Characterizing the neuro-cognitive architecture of non-conscious working memory

Darinka Trübutschek

▶ **To cite this version:**

Darinka Trübutschek. Characterizing the neuro-cognitive architecture of non-conscious working memory. Cognitive Sciences. Sorbonne Université, 2018. English. NNT : 2018SORUS101 . tel-02956592

HAL Id: tel-02956592
https://theses.hal.science/tel-02956592

Submitted on 3 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Pierre et Marie Curie

Ecole doctorale Cerveau, Cognition et Comportement

Inserm-CEA Cognitive Neuroimaging Unit (Unicog, Neurospin)

CHARACTERIZING THE NEURO-COGNITIVE ARCHITECTURE OF NON-CONSCIOUS WORKING MEMORY

Darinka TRÜBUTSCHEK

Dissertation submitted in partial fulfillment for the degree of

Doctor in Philosophy (PhD) in Cognitive Neuroscience

Advised by Prof. Stanislas DEHAENE

and co-advised by Dr. Sébastien MARTI

Presented and publically defended on the 1st of October 2018

In front of a jury and invited members composed of:

Prof. David SOTO	Basque Center on Cognition, Brain, and Language	Reviewer
Prof. Mark STOKES	University of Oxford	Reviewer
Prof. Paolo BARTOLOMEO	Institut du Cerveau et de la Moelle épinière	Examiner
Dr. Lucie CHARLES	University College London	Examiner
Prof. Stanislas DEHAENE	Neurospin	Examiner
Dr. Sébastien MARTI	Neurospin	Examiner
Prof. Claire SERGENT	Paris Descartes	Examiner
Prof. Catherine TALLON-BAUDRY	Ecole Normale Supérieure	Examiner

[This page is intentionally left blank]

An die, die zählen: Mama, Papa, Janka und Henrik

[This page is intentionally left blank]

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Just barely having decided on the basic format for this thesis, I want to focus on its most important aspect – the human, personal part – straightaway. Without the help, support, and love of many, this document would not exist and I would not be the scientist, wife, daughter, sister, and friend that I am today. So please, allow me to thank you all.

My deepest admiration and heartfelt gratitude first and foremost go to my **advisor**: Thank you, Stan, for having given me a chance and having welcomed me into your lab when no one else would; for having introduced me to the world of consciousness, MEG, and decoding; for having challenged me in what, at the time, may have seemed like intense grilling sessions, yet turned out to be the most formative, stimulating, and enriching experiences I could have ever had. You never seized to push, to demand more, to expect the very best. And for this, I will always be grateful, because it turned me into the very best scientist I could be. Mille merci, Stan! I sincerely hope that our collaboration will continue and that we will further pursue our exploration of long-lasting blindsight and activity-silent short-term memory.

At least the same amount of appreciation belongs to you, Seb. Though I could never quite convince the Ecole doctorale to recognize you as my official **co-advisor**, you wholeheartedly took on that role and have become so much more than that: a friend. I will never be able to thank you enough (there is simply not enough space in this manuscript), so here is just a glimpse: Thank you for your time, for having lent me your ear (and hand) whenever I stumbled into your office with yet another question or problem. Thank you for your patience, for not having given up on me despite my surely incessant obsessions with preprocessing, the French administration, and letters of recommendation. And thank you for your kind guidance, for having shared your expertise with me, advised me to pursue the right research topics, and helped me grow. I will always look up to you and look forward to continuing our collaboration in the future.

My many thanks also go to my **collaborators and co-authors** near and far, who helped shape this project and me in countless ways: to Tobias, who nurtured my interest in the human brain and has been a constant source of support throughout all these years; to Josselin, who eased my transition from the States to good old Europe and has granted me an inordinate amount of patience; to Andrés, who introduced me to the study of consciousness and served as my “partner in crime” during the early phase of this thesis; to Jean-Rémi, who tirelessly offered his expertise in machine learning and taught me everything I know about decoding; to Yuanyuan, who modeled our data and generously shared his code; to Misha, whose enthusiasm about our work pushed it to the next level; to Jaco, whose technical and mathematical expertise I could always count on; and to Lionel, with whom I co-authored the probably quickest paper I will ever have published in my career.

All **members of my thesis committee**, thank you very much, for all the attention, insight, and time you have devoted to me and my work. Catherine, our paths crossed early and your work has accompanied me all along this journey. Claire, you are an “eternal” member of my interim committees and have shaped this project with your advice from year 1. David, in a sense, you are the “father” of this thesis, without you, this whole body of work would not exist; I am beyond thrilled to finally have enough time to launch our collaboration. Lucie, your work and thesis served as the best introduction to consciousness science I could have asked for; Mark, your work has been a constant source of inspiration; I cannot wait to start our investigation of activity-silent working memory; Paolo, having been my official advisor from UPMC, I am glad that we had no issues to discuss during my thesis and now get to focus only on the juicy stuff.

It is beyond doubt that this whole endeavor and my stay in France would only have been half as fun had I not made some of the best **friends** anyone could ask for. Valentina, I deeply admire your dedication, kindness, and work ethic – I wish I were at least half as good a German as you are (in your heart). Fabian, we only got to share the same open space for a little while, but not much more was needed to jumpstart our friendship and my lifelong dedication to scikit-learn (and that, mind you, from a Matlab aficionado).

Elo, you quickly turned from Martin's girlfriend into one of my best friends here (though you still owe me a brunch and karaoke night). Martin, what can I say? I have probably had some of the most frustrating, yet also highly entertaining lunch breaks with you and will never forget your incessant flow of ideas. Valérie, perhaps it was fate having been put in the same room as you during our interviews here in Paris, but I think, from the very start, we were just meant to be together. Witek, although it is certainly a rare occasion, your smile always makes me feel better instantly. Allegra, we met too late. We have been riding the bus together for more than a year now and it has been such a blast. I cannot wait to finally get to see one of your apartments (in Rome). Pedro, the only other person as obsessed with food and wine as me is you. I hope we will have many more occasions to celebrate this culture together. Andrès, despite your tendency to disappear, you are one of the kindest and happiest souls I have ever met and, as such, were always able to make me smile. Bianca, your laughter is so contagious, and I love your love of life. Clément, after a somewhat bumpy start, I think we also hit it off with lots of clusters, decoding, and mental rotation. Esther, you introduced me to Chinese culture and one of my favorite bands. Parvaneh, you are incredibly patient and kind. Since you have been gone, something really has been missing in Neurospin. Elisa, thank you a million for showing Henrik and me some of your favorite beaches and for you seemingly endless positive attitude. Dror, I enjoyed every single one of our discussions (about Israel). Thanks for always keeping it real. Liping, I may have missed your talk the other week, but I will never forgot your hospitality in Beijing (and the most appetizing Peking duck I have eaten so far). Lina, your stay may have been short, but these were some of the best and most entertaining weeks at Neurospin for me. Now the German fort is slowly fading. Jennifer, you were one of my first friends here in Paris; thank you, for helping me adjust and learn how to speak French (somewhat) properly.

Many thanks to all of my other colleagues in **Unicog** and **Parietal**, who made every single aspect of my life and work here so much easier: Aaron (whom I bugged with many questions about MEG), Antonio (who was my lifeline many times), Baptiste (who I had the most enriching, philosophical discussions with), Béchir (who gave me the courage to apply for more funding), Benoît (who, together with Vale, was the lab's heart and soul), Christophe (whose didactic skills and deep statistical knowledge saved me more than once), Christos (who is so kind and has the best sweets), Darya (who always smiles at me when we cross in the corridors), Denis (whose enthusiasm about everything is absolutely contagious), Evelyn (whom I shared my very first deep dish pizza with), Fanis (who always greets me with a smile), Fernanda (who helped me set up the most crucial equipment in all of Neurospin: a fan), Florent (the lab's modeling wizard who never got tired to answer my many, many questions), Fosca (who has the most awesome basement in all of Paris and saved me in one of my darkest hours), François (who turns traveling on the RER into an enjoyable experience), Gaël (who answered many, many, many questions about scikit-learn), Ghislaine (who asks the most critical and toughest questions), Giulia (whom I am glad to have around), Isabelle (who tirelessly helped me set up the cluster, again, and again, and again), Laetitia (who was the only other soul to beat in the mornings), Laurence (who always reimbursed me on time and gave me more than expected), Leila (who saved me in the MEG more than once), Lucie (with whom I got to enjoy one of the best conference destinations I have ever been to and who tells the craziest stories), Marie (who gave me a heart attack when, all of the sudden, I heard my own voice reading a series of statements during her defense), Maryline (who makes everything easier), Maxime (who is always there if need be), Milad (who is the most entertaining commute partner ever), Nahuel (who managed to congratulate me for my thesis before I had even submitted it), the nurses here at Neurospin (who tremendously facilitated subject recruitment), Sophie (whom I still owe a breakfast), Tad (who is one of the only people in this environment to appreciate the struggles of an early career scientist), Timo (who offered many times to train me with monkeys and whose help I still hope to take up), Vanna (who, after five years, finally managed to soothe the inherent angst of a German when it comes to administration and paperwork), Virginie (who is someone to live up to), Yvonne (with whom one can have the best conversations about French culture and customs), Zafer (who patiently tried to explain decoding to me many times).

Acknowledgements.

A special shout-out certainly is also due to the **Ecole des Neurosciences de Paris** and the **Fondation Schneider Electric**, both of whom provided me with very generous financial support throughout my entire PhD and time here in Paris.

Zu guter Letzt geht mein Dank natürlich an die, auf die es wirklich ankommt und deren Anerkennung und Achtung mir mehr als alles andere bedeuten: **meine Familie**. **Mama** und **Papa**, danke, dass ihr mich zu der Person erzogen habt, die ich heute bin; dass ihr mich auf meinen Wegen nicht nur immer begleitet, sondern tatkräftig unterstützt (egal in welchen Erdteil sie mich verleiten); dass ihr mir beigebracht habt, ein Kämpfer zu sein, der vielleicht mal vom Pferd fällt, aber wieder aufsteht; dass ihr immer an mich glaubt. Ohne euch hätte ich es nie soweit geschafft, wäre wahrscheinlich schon damals nicht in die USA gegangen und würde euch heute nicht dieses Buch überreichen. Ihr seid meine ganze Kraft. **Janka**, meine Kleine, danke, dass du mich immer wieder an das Wesentliche erinnerst, Mensch zu sein, nicht nur Wissenschaftler; dass du mich aus jeder Misere zu holen vermagst und immer ein Lachen auf meine Lippen zaubern kannst; dass du immer ein offenes Ohr für mich hast, egal wie banal dir meine Probleme und Sorgen auch erscheinen mögen; und dass du mich trotz deiner großen Abneigung immer mal wieder in Paris besucht hast. **Oma**, auch wenn du selbst diese Zeilen nicht mehr wirst lesen können, möchte auch ich dir hier für all die schönen Jahre und Zeiten mit dir danken. Auch du hast mich erzogen und immer das Beste aus mir herausgeholt. Danke noch mal für Alles!

Henrik, dir gebühren selbstredend die allerletzten Zeilen dieser Danksagung. Ohne dich wäre ich heute sicherlich nicht hier. Du hast jede einzelne Phase dieser Zeit meines Lebens hautnah miterlebt, die Ups und Downs, und mich immer wieder motiviert, nicht aufzugeben, nicht das Handtuch einfach hinzuschmeißen. Ich danke dir für deine unermüdliche Güte, Verständnis und Vertrauen. Du bist und bleibst mein Fels in der Brandung, das Licht in jeder noch so dunklen Nacht. Ich kann es kaum erwarten, auf ewig mit dir verbunden zu sein. Danke, für alles, mein Schatz; ich liebe dich. Du weißt: Einzeln sind wir Worte …

ABSTRACT

Our lives hinge on our ability to hold information online for immediate use. For over a century, cognitive neuroscientists have regarded such working memory as closely related to consciousness, with both functions sharing similar features and brain mechanisms. Recent work has challenged this view, demonstrating that non-conscious information may affect behavior for several seconds, and suggesting that there exists a genuine non-conscious working memory system. I here combine behavioral and modeling approaches with time-resolved magnetoencephalography and multivariate pattern analysis to put this proposal to the test. In a first study, I rule out alternative explanations for the long-lasting blindsight effect, showing that it results from a genuinely non-conscious process. Crucially, this non-conscious maintenance is not accompanied by persistent delay-period activity, but instead stores information in "activity-silent" brain states via transient changes in synaptic weights. In a second set of experiments, I systematically evaluate key properties of conscious working memory in the context of long-lasting blindsight. While even multiple items and their temporal order may be stored non-consciously, manipulating stored representations is associated with consciousness and sustained neural activity. Together, these results challenge theories that equate the maintenance of information in working memory with conscious activity sustained throughout the delay period, but also contradict the notion of a genuine non-conscious "working" memory. Instead, I propose the existence of activity-silent short-term memory.

Key words: non-conscious working memory, activity-silent working memory, consciousness, mental rotation, magnetoencephalography (MEG), multivariate pattern analysis

RESUME

Nous avons la capacité de maintenir en mémoire, de manipuler, et de transformer des informations provenant de notre environnement. Depuis plus d'un siècle, les neuroscientifiques considèrent la mémoire de travail comme étroitement liée à la conscience, les deux fonctions partageant des caractéristiques et des mécanismes cérébraux similaires. Des travaux récents ont remis en question ce point de vue en démontrant que des informations non-conscientes peuvent affecter le comportement pendant plusieurs secondes (« vision aveugle »), et suggérant qu'il existe un véritable système de mémoire de travail non-conscient. Nous combinons ici l'étude du comportement, l'imagerie du cerveau à haute résolution temporelle, et la modélisation computationnelle pour tester ces hypothèses. Dans une première étude, nous rejetons plusieurs explications alternatives à la vision aveugle, montrant que celle-ci résulte d'un processus véritablement non-conscient. Nous montrons également que le maintien non-conscient de l'information ne s'accompagne pas nécessairement d'une activité cérébrale soutenue pendant toute la période de maintien, mais pourrait dépendre d'états cérébraux «silencieux» qui sollicitent des changements transitoires dans la connectivité synaptique. Dans une deuxième série d'expériences, nous évaluons systématiquement les propriétés clés de la mémoire de travail consciente dans le contexte de la vision aveugle. Même si plusieurs éléments et leur ordre temporel peuvent être stockés de manière non-consciente, la manipulation des représentations nécessite l'accès conscient et une activité neuronale soutenue. Dans leur ensemble, ces résultats d'une part défient les théories qui assimilent simplement le maintien de l'information en mémoire de travail à une activité consciente soutenue tout au long de la période de maintien. D'autre part, ils contredisent la notion d'une véritable mémoire «de travail» non-consciente. Au lieu de cela, nous proposons l'existence d'une mémoire à court terme « silencieuse ».

Mots clés: mémoire de travail non-consciente, mémoire de travail silencieuse, conscience, rotation mentale, magnétoencéphalographie (MEG), analyses multivariées

LIST OF FIGURES

LIST OF TABLES

PUBLICATIONS OF THE AUTHOR

Part of the work presented in this thesis already has been (or will be) published in internationally recognized, peer-reviewed journals. In addition, several papers and manuscripts, though not included here, have been prepared or published during the course of my PhD and, as such, are worth mentioning at this point.

ARTICLES INCLUDED IN THIS THESIS

Chapter 2 **Trübutschek, D.**, Marti, S., Ojeda, A., King, J.-R., Mi, Y., Tsodyks, M., & Dehaene, S. (2017). A theory of working memory without consciousness or sustained activity. *eLife, 6*:e23871, doi: 10.7554/eLife.23871

Chapter 3 **Trübutschek, D.**, Marti, S., & Dehaene, S. (under review). Temporal order information can be maintained non-consciously.

Chapter 4 **Trübutschek, D.**, Marti, S., Ueberschär, H., & Dehaene, S. (under review). Probing the limits of activity-silent, non-conscious working memory.

OTHER ARTICLES AND MANUSCRIPTS PRODUCED DURING THIS DISSERTATION

Naccache, L., Marti, S., Sitt, J.D., **Trübutschek, D.**, & Berkovitch, L. (2016). Why the P3b is still a plausible correlate of conscious access. A commentary on Silverstein et al., 2015. *Cortex*. Doi: 10.1016/j.cortex.2016.04.003

Coyle, E. F., Fulcher, M., & **Trübutschek, D.** (2016). Sissies, mama's boys, and tomboys: Is children's gender nonconformity more acceptable when nonconforming traits are positive? *Archives of Sexual Behavior*. Doi: 10.1007/s10508-016-0695-5

Trübutschek, D., Henry, C., d'Albis, M.-A., Duclap, D., Hamdani, N., Daban, C., …, & Houenou, J. (under review). Emotional reactivity and limbic networks: A multimodel MRI study in bipolar patients and controls.

Trübutschek, D.*, Kiyonaga, A.*, & Egner, T. (under review). The 'what' and 'how' of working memory: Dissociating neural mechanisms of declarative and procedural functions.

CHAPTER 1 –
GENERAL OVERVIEW OF THE LITERATURE

Without memory no conscious sensation,
without memory no consciousness.
- CHARLES RICHET (1886)

1.1 TRAPPED IN THE MOMENT: THE CURIOUS CASE OF HENRY GUSTAV MOLAISON

Imagine, if only for a fleeting moment, that you were to lose your ability to form new memories. Not just a specific type of memory, all of them. Try to picture how, over the course of your long life, you would slowly cease to recognize your colleagues, your friends, even your closest kin, simply because their physical appearances change as they age. Do not forget the more mundane things either: how you would not be able to keep up with current events, societal changes, technological innovations, or how you would not even be capable to navigate the most banal aspects of your life, such as assuring a healthy, balanced diet. In essence, you would be stuck in the past eternally.

If this description reminds you of the beginning of a horror movie or perhaps a science-fiction drama, be not mistaken. It very well could have been your fate – just as it had been for one of the most famous patients in the history of psychology: Henry Gustav Molaison, or short, H.M. (Figure 1.1A). Plagued by intractable epilepsy, in 1953, Henry agreed to surgery to have the affected parts of his brain removed. This led to the almost complete resection of his bilateral hippocampi, the adjacent parahippocampal gyri, and the left and right amygdalae, all structures buried deep down in the medial temporal lobes of his brain (Scoville and Milner, 1957; Figure 1.1B). While the surgery was successful in treating his epileptic seizures, unfortunately, it also caused profound anterograde as well as a fair amount of retrograde amnesia. Similar to our thought experiment from the beginning of the chapter, Henry had completely lost his capacity to store new (long-term) memories.

Although he was thus severely incapacitated with respect to certain aspects of his life, for instance, never remembering a single one of the scientists, doctors, and nurses that worked with him on a daily basis for several decades, other facets of his cognitive functions and mental life remained remarkably intact. His general intelligence was well above average (Scoville and Milner, 1957), and he was able to learn new motor skills, such as how to draw an outline around an intricate figure when only being able to view his hands and the template through a reflection in a mirror (Pribram and Broadbent, 1970). Most importantly for the

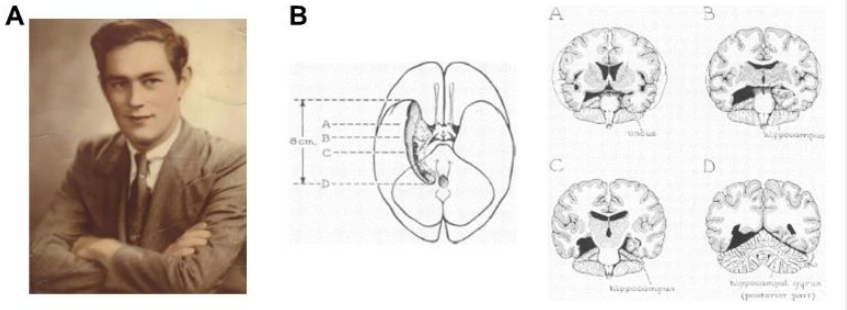


FIGURE 1.1

H.M. – THE MAN WHO WAS TRAPPED IN PERPETUAL PRESENT.

(A) A picture of H.M., taken on the day of his high school graduation, years before a surgery to cure his intractable epilepsy removed a large section of his medial temporal lobe, including bilateral amygdalae, hippocampi, and parahippocampal gyri.
(B) Diagrammatic cross-sections of the human brain, illustrating the extent of the surgical resection of the medial temporal lobe carried out in H.M. (and others): (A) Uncus, (B) Hippocampus, (C) Hippocampus, (D) Parahippocampal gyrus. For display purposes, the resection is shown in one hemisphere only. Adapted from Scoville and Milner (1957).

intents and purposes of this thesis, however, his memory for immediate events (i.e., short-term or working memory) was fully functional. For as long as he paid attention to the task at hand, he had, for example, no trouble remembering three-digit numbers for up to 15 minutes (Pribram and Broadbent, 1970) or remembering sequences of six digits (Squire and Wixted, 2011). Indeed, his conscious experience appeared completely normal, even allowing him to enjoy crossword puzzles in-between his doctors' appointments, and only failing him when requiring integration beyond the present moment, beyond the here and now. To put it in Henry's own words (Pribram and Broadbent, 1970):

> "*You see, at this moment, everything looks clear to me, but what happened just before? That's what worries me. It's like waking from a dream; I just don't remember.*"

Now, for a second, let us imagine the reverse scenario: What would Henry's life have been like, had he not lost his capacity to form (and store) new long-term memories, but rather his ability to remember information for fairly brief periods of time? Would he still have been able to indulge in those crossword puzzles and cherish these moments? Intuitively, most of us would deny this assertion. To develop and maintain a stable, unified sense of the self, it seems, one needs intact long-term memories. But to be fully present in and consciously experience each and every single waking moment of our lives, it appears short-term memories are indispensable.

This notion, that our conscious experience of the world is very closely related to our immediate memory, has indeed been deeply embedded in the thinking of philosophers, psychologists, and neuroscientists for a very long time. The French physiologist Charles Richet, whose famous perspective I already quoted at the outset of this chapter, further stated that

(1) *"for a conscious sensation […] to occur, there must be a present of a certain duration, of a few seconds at least"*, and that

(2) *"[…] to suffer for only a hundredth of a second is not to suffer at all; and for my part I would readily agree to undergo a pain, however acute and intense it might be, provided it should last only a hundredth of a second, and leave after it neither reverberation nor recall"* (Richet, 1884).

Only a couple of years later, in his magnum opus (1890), the father of psychology himself, William James, adopted Richet's stance, thereby setting the stage for most contemporary theories on conscious perception and (working) memory. The body of work I present in this thesis puts these long-held, yet essentially unverified, assumptions to the test and sheds new light on our understanding of the relationship between consciousness and short-term (working) memory. We will now begin with a selective review of the core concepts necessary for a full appreciation of this contribution.

1.2 Working memory – connecting the present to the future

When we think of the term *memory*, we typically envision a single store that holds information about our past, including the knowledge we acquired and the experiences we have had thus far. Yet cases, such as Henry's, do not fit with this kind of conceptualization. If there were really just a single, functionally and neurobiologically unitary memory system, how come Henry could remember information for several minutes, but not for hours or days? How come he could acquire new motor skills even if he had no conscious recollection of ever having been taught this particular movement?

What psychologists and neuroscientists have learned from decades of research is that memory is everything but a single, functionally indistinguishable module of the brain. Instead, we now believe that there may be many separate memory stores or, depending on whom you talk to, states of activation, each with a dedicated function (Figure 1.2). Not all of these are directly relevant for my work, so I will just briefly highlight some of these distinctions before officially introducing the specific type of memory I am interested in here.

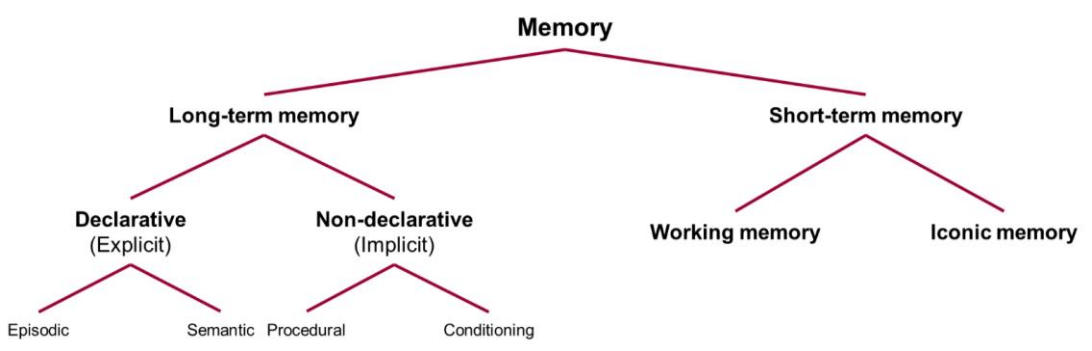


FIGURE 1.2

CLASSICAL TAXONOMY OF HUMAN MEMORY.

According to traditional views, memory may be fractionated into sub-components or stores, each of which is adapted to hold a specific type of content for a particular amount of time. Adapted from Squire and Zola-Morgan (1988).

1.2.1 DELINEATING SHORT-TERM MEMORY FROM LONG-TERM MEMORY

One of the first divisions of memory was proposed by William James. Based entirely on his own introspection, James (1890) distinguished between a *primary* and a *secondary* memory, the former reflecting the current contents of consciousness and lasting for only tens of seconds, and the latter latently storing information about the distant past for an indefinite period of time. The very same division, albeit having been renamed into *short-term* and *long-term* memory, still guides our thinking today.

Part of the evidence rests on the observations of neurological patients. Remember how, following his surgery, Henry was unable to form any new long-term memories, while having no trouble storing information for short periods of time (Pribram and Broadbent, 1970; Scoville and Milner, 1957; Squire, 2009). The converse phenomenon also exists: After damage to his left occipito-parietal cortex, a 28 year-old man, K.S., exhibited a severely impaired short-term memory, not being able to remember more than two digits at a time, while his long-term memory remained virtually intact (Shallice and Warrington, 1970; Warrington and Shallice, 1969). Double dissociations such as these, in which, in one individual, a lesion in brain area A compromises cognitive function X (e.g., long-term memory) and spares cognitive function Y (e.g., short-term memory), while, in another person with a lesion in brain area B, the reverse pattern is observed, are a very powerful method in the toolkit of any psychologist or neuroscientist. They allow us to demonstrate that the two mental processes are dissociable in terms of their function as well as their neural underpinnings, and have thus played an important role in the study of memory.

Clever experiments also supported the distinction between a short- and a long-term memory store. Have a friend read you the following list of numbers at a rate of about one item per second. Then, write them down on a piece of paper in any order you may wish.

64 – 51 – 30 – 80 – 4 – 44 – 81 – 40 – 2 – 57

45 – 94 – 24 – 63 – 78 – 15 – 61 – 62 – 28 – 27

If you had done this experiment multiple times (with different lists of numbers, of course), and had then plotted the average probability of you recalling a specific item as a function of the serial position of the item in the list, you might have obtained a graph such as the one shown in Figure 1.3. You can immediately see that there is a discrepancy in the probability of recall: Whereas both the first and the last few items of the lists had a fairly good chance of being remembered, the intermediate items were far less likely of being recalled (Ebbinghaus, 1885; Murdock, 1962). This serial position effect is typically explained in terms of a division between short-term and long-term memory: Because participants can devote a lot more undivided processing and rehearsal to the first items of the list, these are preferentially encoded into

long-term memory and can thus be preserved until the time of recall (i.e., primacy effect). The last items, by contrast, are still held actively in short-term memory when recall is solicited and are thus fairly accessible to be retrieved. Both neuropsychological and experimental data may therefore be used to delineate short-term from long-term memory.

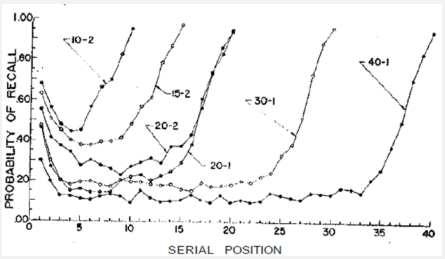


FIGURE 1.3

EMPIRICAL EVIDENCE FOR A DISSOCIATION BETWEEN LONG- AND SHORT-TERM MEMORY.

When asked to freely recall a list of items (in this case, lists of numbers with varying length), participants are more likely to report the items that were presented first (i.e., primacy effect) as well as the items that were presented last (i.e., recency effect) in comparison to items from the middle of the list. This serial position effect has typically been taken as behavioral evidence in favor of separate short- and long-term memory stores. The first few items of the list can be rehearsed without any major interference from other items, thus being relatively successfully encoded into long-term memory. By contrast, the last few items of the list are still held actively in working (or short-term) memory at the time of recall and are therefore also preferentially being retrieved. Adapted from Murdock (1962).

The temporal dimension (i.e., for how long a given piece of information can be stored before being forgotten) is not the only line of division proposed for memory. There exist many more parts to this taxonomy. Long-term memory may, for instance, further be divided into *declarative* and *non-declarative* components, with the main difference between the two traditionally having been attributed to explicit (i.e., conscious) vs. implicit (i.e., non-conscious) recollection of events (Squire and Zola-Morgan, 1988). If I, for example, asked you right now to tell me how you spent 9/11, you would probably be able to fairly accurately recall the events of that particular day, perhaps even bringing to mind a picture of a specific scene. (I, for instance, was sitting on a couch at a friend's house, watching *The Weakest Link*, when the show was interrupted by breaking news.) Similarly, if you had to name the current president of the United States, you would be able to do so. Both of these examples were instances of declarative memory, the first one being *episodic* (i.e., retrieval of specific, personal events) and the second one *semantic* (i.e., retrieval of general knowledge) in nature (Tulving, 1972). Non-declarative memory, in contrast, encompasses, for instance, *procedural memory* (i.e., the type of memory for motor skills that had been preserved in Henry) and *conditioning* effects (i.e., a learned relation between events; Pavlov and Anrep, 2003; Rescorla, 1988). None of these sub-systems of long-term memory are particularly relevant for this thesis, and so I only wanted to touch upon them briefly as an overview. Suffice it to say that, based on decades of research, there now exists an elaborate taxonomy of human long-term memory.

1.2.2 DELINEATING SHORT-TERM MEMORY FROM SENSORY MEMORY

So far, we have primarily discussed the distinction between short- and long-term memory. As their respective names imply, the former allows to store information for short periods of time (in the order of tens of seconds or minutes), whereas the latter can retain memories (almost) indefinitely. Another point of divergence relates to the sheer amount of information that can be stored (Cowan, 2008). Long-term memory has a very large *capacity*: You seem to possess an almost unlimited repertoire of memories, including a hodgepodge of semantic facts and personal, episodic events. Putting an exact number on the capacity of long-term memory is difficult. Previous research has shown that pigeons learned about 800 – 1,200 picture-response associations over the course of 3 to 5 years before their performance on a long-term memory task declined, while baboons had not exhausted their long-term memory capacity after 3,500 – 5,000 items (Fagot and Cook, 2006). Estimates of the lower bounds in humans range between ~10,000 individual items (Standing, 1973) and ~2,500 detailed representations, corresponding to ~228,000 unique codes (Brady et al., 2008). Long-term memory is a truly massive store.

By contrast, the capacity of short-term memory is far more limited. Remember how Henry could recall about six digits when he was focused on the task at hand? If you go back to the graph on the serial position

effect (Figure 1.3), you may also notice that the benefit of the recency effect extends to ~7 – 8 items. Psychologists have formally studied this problematic, asking their participants to retain different stimulus materials (e.g., lists of digits, letters or words; arrays of spatial locations or colors, etc.) presented to different modalities (e.g., auditory, visual) for short periods of time (e.g., Crannell and Parrish, 1957; Luck and Vogel, 1997). The findings of all of these studies converge to a single fact: Our short-term memory is severely limited in capacity, with estimates ranging somewhere from 3/4 to 7 items (or chunks of information; Cowan, 2001; Halford et al., 2007; Figure 1.4). This was probably most famously summarized by Harvard psychologist George A. Miller (1956), when he proposed "the magical number seven, plus or minus two" as the limit for the span of immediate (i.e., short-term) memory.

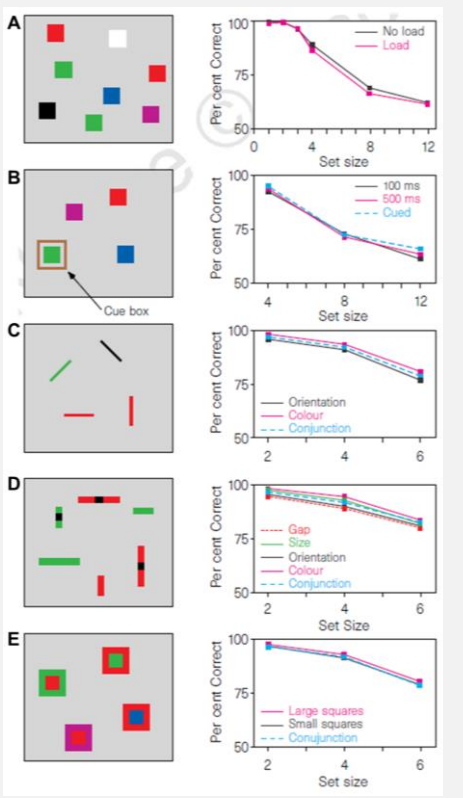


FIGURE 1.4

EMPIRICAL EVIDENCE FOR A CAPACITY LIMIT IN SHORT-TERM MEMORY.

In their seminal study, Luck and Vogel (1997) presented their participants with a series of different stimulus arrays (A – E, left) and asked them to retain this information over a short delay. Strikingly, irrespective of the actual material used, subjects' working memory capacity was limited to about 4 items (A – E, right). Note that this capacity did not vary as a function of the complexity of the items, such that 4 individual colors could be maintained equally well as the conjunction of 4 colors in a given spatial position.

Is our memory inherently biased towards the long-term retention of information? The long-term store I have described so far just seems so much more powerful than its short-term counterpart, imposing almost no limits on the quantity of or duration for which memories may be maintained. Is that really all that there is to the fractionation of memory or could there be an equivalent to this massively unlimited long-term store on the short-term memory side as well? Early work by yet another Harvard psychologist, George Sperling, suggested that this may indeed be so (1960). Consider a slightly different setup of the experiment you did previously: Instead of having the digits read to you one by one, imagine them appearing all at once on a computer screen, grouped in five rows of four numbers each and disappearing after 50 ms of exposure. What would have happened, if you had to report all of them immediately, that is, give an immediate *full report*? – Just like Sperling observed in his subjects, you would have shown the typical effects of a capacity-limited short-term memory, on average not being able to report more than ~4 individual numbers. Now envision yet another scenario: Directly after the offset of the stimulus array, a sound will be played, telling you which specific row of numbers to report. That is, you will be asked for a *partial report*. How many of these four numbers would you be able to recall correctly? This question may seem a bit odd at first sight: If the capacity of short-term memory is limited to about four items and you paid careful attention to all of the numbers in the original display, then, on average you should be able to report at most one of the numbers from the cued row. Strikingly, this is not at all, what Sperling found. Based on the data from such partial reports, he estimated that his participants could correctly remember at least twice as much information as under the standard full report condition, yet this advantage disappeared entirely when the delay between the offset of the stimulus array and the presentation of the cue approached 1 s (Sperling, 1960). Follow-up experiments demonstrated that the presentation of an intervening stimulus between the original stimulus array and cue (i.e., a mask) also disrupted the benefits of these partial reports (Averbach and Coriell, 1961; but see Smithson and Mollon, 2006 for a contrasting opinion - this will become important later on).

As such, Sperling interpreted his findings as evidence for a quickly decaying, yet high-resolution "after-

image" of the original stimulus (Sperling, 1960), but this initial conceptualization was soon recast as yet another type of memory: *sensory* or *iconic memory* (Neisser, 1967). The basic idea here is simple: There are two independent memory buffers for the short-term storage of information. The first, iconic memory, functions as a high-capacity system, retaining information in a very high-fidelity and rather literal format, but persisting only for a couple hundred milliseconds. The second, short-term memory store, by contrast, has a much more limited capacity of only a handful of chunks of information, but, even in the absence of any rehearsal, could maintain a given memory trace for several seconds. Knowing which of the rows was going to be probed allowed you to sample from this high-capacity iconic representation in the partial-report task, thereby helping you avoid having to rely solely on your limited short-term memory store (Dick, 1974). More recent evidence from behavioral experiments (Sligte, 2010; Sligte et al., 2008) and brain imaging studies (Sligte et al., 2009) confirms the existence of such an iconic memory store separate from traditional short-term memory and even proposes yet another memory system, *fragile visual short-term memory*, intermediate between iconic and short-term memory in terms of both capacity and durability of the memory trace. We will revisit this idea at a later time, so, for now, just keep in mind the two main buffers discussed.

1.2.3 Cognitive models of working memory

In summary, at the crudest level, memory may be divided into three components: (1) a massive long-term store, (2) a capacity-limited short-term buffer, and (3) a quickly decaying, yet high-capacity iconic memory trace. But which out of all these systems is the one most relevant for the work I conducted? If you consider the initial question as well as the studies discussed so far, it probably comes as no surprise that, for the rest of this thesis, I will primarily focus on short-term memory or, to be more exact, on a very specific type of short-term memory, called *working memory*.

In essence, everything I have said so far about short-term memory also applies to working memory. In fact, many researchers use these two terms interchangeably (Aben et al., 2012), and I have, up until now and primarily for historical and didactic reasons, chosen to stick to the traditional term of short-term rather than the more recent conceptualization of working memory. From here on out, however, I will reserve the term working memory to refer specifically to a memory system that encompasses short-term memory in addition to other processing mechanisms that allow us to make use of the information held in short-term memory. In other words, I will use working memory in-line with its original connotation of a memory system "for the execution of our Plans" (Miller et al., 1960, p. 65). This definition allows me to highlight one important distinction with respect to the other memory systems we have talked about: the inherently prospective nature of working memory (Fuster, 2015, p. 144). Long-term memory, short-term memory (in its restricted sense) and iconic memory are all, more or less, accurate, permanent, and static snapshots of your past, helping you to integrate your previous experiences (or states) with your current ones. Working memory, if you will, allows you to connect your present to your future. It permits you to keep track of events in a sequence (e.g., when cooking your favorite dinner), hold and manipulate information in your mind (e.g., when mentally solving a difficult arithmetic question), and plan your future decisions (e.g., when mapping out your daily commute to work). In a sense, it serves as the sketchpad of your consciousness (but we will discuss this in much more detail later). Let us first consider some of the cognitive models psychologists have developed to account for the experimental findings we have already encountered (i.e., temporary memory trace and severe capacity limits).

1.2.3.1 Systems-based models of working memory

One of the ultimate goals of experimental psychology and cognitive neuroscience is to explain and predict (human) behavior and mental processes. An important step along the way is the development of theoretical models about specific cognitive functions, such as working memory. They help us summarize

and extract common patterns from the existing body of empirical evidence, and guide our future research by identifying key unknowns and making specific, testable predictions. Most importantly, they are not set in stone, but only reflect the current state of the evidence, thus evolving continuously.

Over the last 70 years, many different cognitive models of working memory have been proposed. One of the earliest was developed by Donald E. Broadbent (1957, 1958) and did not actually deal with working memory per se. Rather, it was primarily meant to explain another type of mental process we call *attention*. Imagine yourself at your best friend's wedding: The music plays audibly in the background, the room is full of other guests and buzzing with chatter, and you are deeply engaged in a conversation. How is it possible that, despite the constant source of incoming noise, you can fully focus on the conversation at hand, seemingly tuning out all other stimulation? Broadbent's *filter model of attention* helps to explain this well-known *cocktail party effect* (Cherry, 1953): He recognized that, while, at all times, our brain receives input from its sensory organs in parallel, not all of this information can and will be processed further. As such, he proposed that, at first, incoming stimuli would be held in a temporary short-term store (similar to what, today, we might term working memory), before being filtered and selectively passed on to a capacity-limited perceptual

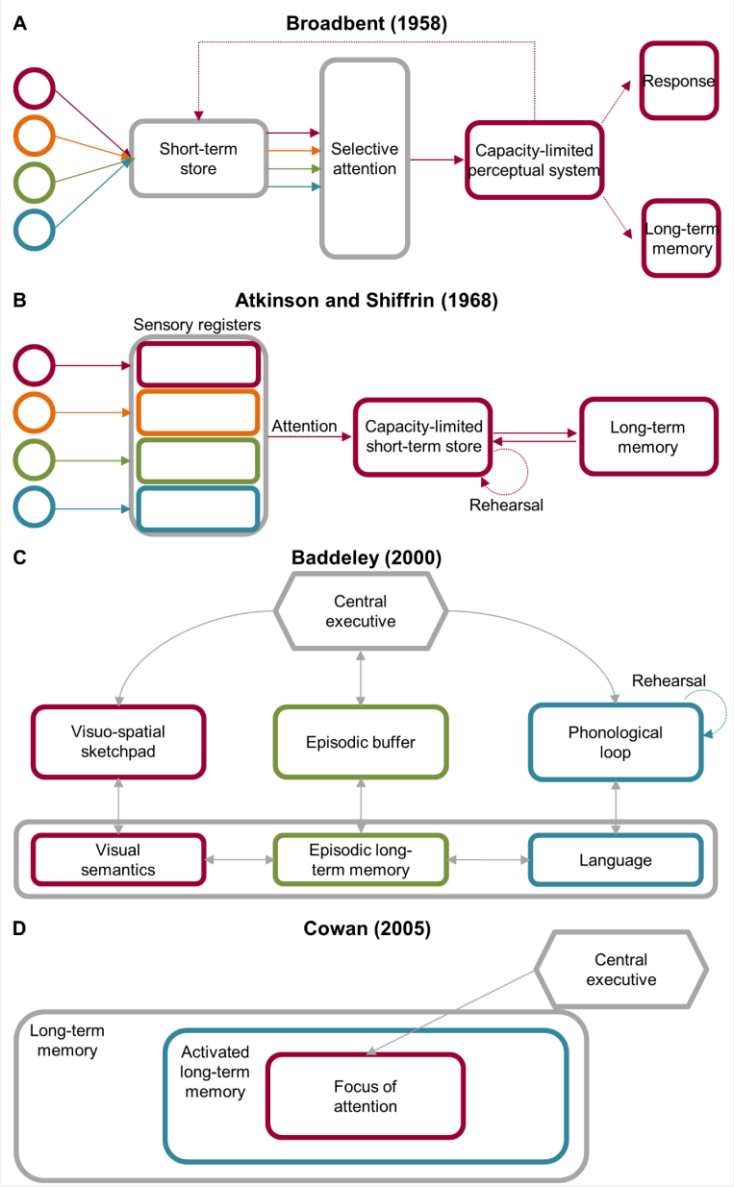


FIGURE 1.5

COGNITIVE MODELS OF WORKING MEMORY.

(A) Schematic view of Broadbent's selective filter model of attention. Incoming sensory information (colored circles on the left) competes for access to a central, capacity-limited perceptual system. The short-term store initially ensures the pre-selection temporary maintenance of all incoming information and, once relevant information has been selected, may then hold only the currently relevant item for as long as no new information is received. As such, it constitutes one of the earliest instantiations of a working memory. Adapted from Broadbent (1958).
(B) Diagrammatic view of Atkinson and Shiffrin's multi-store model of memory. Sensory information is first thought to be held in a high-capacity, modality-specific sensory memory buffer (i.e., iconic memory), before being selected for further processing. Material currently in the capacity-limited short-term store is under the direct and willful control of the subject, allowing him or her, for instance, to refresh the memory trace through rehearsal, use it, or transfer it to the permanent long-term store. Adapted from Atkinson and Shiffrin (1968).
(C) Schematic representation of Baddeley's multi-component model of memory. Within working memory, a "central executive" coordinates, selects, and triggers storage of information in three dedicated, specialized short-term buffers. Adapted from Baddeley (2000).
(D) Diagrammatic representation of Cowan's embedded-process model as an exemplar of state-based models. Attentional processes may select relevant latent representations from long-term memory, thereby activating them and making them accessible for further processing. Adapted from Cowan (2005).

buffer, from which they could either (1) be reverted back to the short-term store, (2) be used to formulate a response or (3) be stored in long-term memory (Figure 1.5A).

This initial model was incomplete in many respects, for instance, leaving out important mechanisms for allowing processing of previously unattended information (e.g., permitting you to shift your attention away from your current conversation with your friend when someone else calls your name) or detailing exactly how information might be shuffled between the different buffers. It was therefore soon elaborated on and mathematically formalized by Richard Atkinson and Richard Shiffrin (1968, 1971) and featured the tri-partite division of memory we have been talking about for quite a while (Figure 1.5B). There was the high-capacity, but quickly decaying iconic memory buffer (i.e., sensory registers), the capacity-limited working memory (i.e., short-term store), and the massive long-term memory store. What is important about this *multi-store model*, however, is not only this fractionation of memory, but also the emphasis that is put on the processes governing the flow of information from one system to the next. Atkinson and Shiffrin (1968, 1971) thought that subjects could willfully control (and operate on) the material currently held in working memory. They could, for example, rehearse it, visualize it, or use it to guide their future decisions. In perhaps the truest sense of the term, they considered their working memory a "store in which decisions are made, problems are solved, and information flow is directed" (Atkinson and Shiffrin, 1971, p. 5).

You may have noticed the fit between this model and the behavioral data from the experiments we talked about before. The limited capacity of working memory as measured by, for instance, digit span, as well as primacy and recency effects are all accounted for by Atkinson and Shiffrin's (1968, 1971) proposition. However, what about the neurological observations? Henry's case appears fairly consistent with the model: His iconic and working memory were all intact, allowing him to maintain and use information for brief periods of time. On the other hand, he was incapable of storing any new memories in his long-term memory, so, most likely, his brain lesion interrupted the transfer from working into long-term memory. But what about patients like K.S., the man with a selective deficit of short-term memory, yet an intact system for long-term memories (Shallice and Warrington, 1970; Warrington and Shallice, 1969)? As it turns out, Atkinson and Shiffrin's (1968, 1971) model is fully incompatible with this type of deficit. Given that information first has to pass through short-term memory to be encoded into long-term memory, any lesion causing a disruption of short-term/working memory should also automatically lead to a gross impairment of long-term memory. To overcome inconsistencies and problems such as these, Alan Baddeley and Graham Hitch proposed their highly influential *multi-component model* of working memory (1974; Figure 1.5C). According to the original conceptualization, working memory comprises a capacity-limited "central, executive" in addition to two "slave systems", one for visuo-spatial information (i.e., visuo-spatial sketchpad) and the other for auditory items (i.e., phonological loop; Baddeley, 1992a, 1983, 1993; Baddeley and Hitch, 1974). Baddeley (2000, 2003, 2012) then later also added the episodic buffer to specifically account for multi-modal, episodic representations. Put simply, the central executive was viewed as a system that could actively regulate the distribution of limited attentional resources and coordinate, select, and trigger short-term maintenance of information in the three dedicated slave buffers. It thus corresponds to the *working* part of working memory and, in stark contrast to the previous models, highlights the separation of the storage from other processing components in working memory.

1.2.3.2 State-based models of working memory

Despite their obvious differences, all of the cognitive models of working memory presented above also share a few commonalities. Most importantly, similarly to what we discussed in the beginning of this chapter, they all seem to presuppose the existence of at least two independent stores for the long-term and short-term retention of information. Another school of thought that, apart from a brief mention, we have not yet discussed so far instead focuses on differences in the *current state of activation* of a

representation in the brain to delineate short-term from long-term memory (Cowan, 1997; Jonides et al., 2008; McElree, 2001; Oberauer, 2002; Figure 1.5D).

In essence, this entire class of models holds that there exists only a single, long-term memory store, in which permanent memories are coded as structural changes in the physical connections between neurons (i.e., *neurons that fire together, wire together*; Hebb, 1949). These long-term representations are typically dormant (i.e., latent), but may, if the current task or situation demands it so, be brought into the focus of attention, thus being activated for short periods of time, during which they then not only populate your mind, but also influence your current thoughts and actions. As such, these attentionally-dependent, activated portions of long-term memory correspond conceptually to the short-term storage of information in working memory (although a more fitting term might perhaps be "working attention").

Exactly how many such states may exist and what their specific capacity limitations might be varies between different instantiations of this type of model. Some distinguish only between latent long-term memory and a very narrow focus of attention (McElree, 2001, 2006), while others differentiate between up to four distinct states, long-term memory, activated long-term memory, a state of direct access (i.e., broad focus of attention), and a narrow focus of attention (Oberauer, 2001, 2002, 2005). The specifics here are not directly relevant for the work I conducted, so I just wanted to mention them in passing. Two aspects that I would like to draw your attention to, in contrast, are the following: First, state-based models are typically considered to be a fairly recent theoretical development. However, even Atkinson and Shiffrin (1971) already stated that their

> *"account of STS [short-term memory] and LTS [long-term memory] does not require that the two stores necessarily be in different parts of the brain, or involve different physiological structures. It is possible, for example, to view STS simply as a temporary activation of some portion of LTS."*

The idea of different states of activation thus appears to have been around for much longer than usually assumed. Second, it is important to note that the clear advantage of this group of models as opposed to more traditional conceptualizations of working memory lies in the inherent flexibility they offer: Information may, for instance, co-exist in several forms at once, and different states of activation might correspond to differences in the amount of processing required (i.e., maintenance vs. manipulation). As such, state-based models represent an especially parsimonious and attractive cognitive theory of working memory.

1.2.4 Searching for the neural correlates of the working memory engram

Up until now, we have explored key properties of working memory (i.e., short-lived, capacity-limited system) and discussed some of the most influential cognitive models attempting to explain said features. What we have not yet touched upon, however, is the issue of how exactly our brain accomplishes this extraordinary feat. How does it allow us to keep in mind information for immediate use, to flexibly store, update, and retrieve it? Over the past half century, many a researcher has attempted to tackle these fundamental questions, and a few notable theories have largely dominated the field of (cognitive) neuroscience. In what follows, I would like to highlight two core questions that generations of psychologists and neuroscientists have grappled with and that continue to spark heated debate even today: (1) Where exactly in the brain is the seat of working memory? and (2) Which specific neural mechanism(s) support(s) the short-term storage and manipulation of information? We will begin with a consideration of the first issue.

1.2.4.1 Dedicated or distributed neural system for working memory?

Most of the initial cognitive models of working memory emphasized the existence of two distinct stores for the long-term and short-term retention of information. These views were consistent with early

neurobiological evidence suggesting that the dorsolateral prefrontal cortex might serve as a dedicated neural system for working memory. Even before the turn of the 20th century, scientists studying the effects of the experimental resection of or accidental injury to the prefrontal cortex in monkeys had already noted the devastating effects these lesions produced on the behavior of the animals, including the loss of "attentive and intelligent observation" (Ferrier, 1876) and of the "coordination and fusion of the incoming and outgoing products of the several sensory and motor areas of the cortex" (Bianchi, 1895). In a series of seminal studies, Jacobsen (1935) then demonstrated that lesions to the lateral prefrontal cortex in primates led to a specific impairment of the capacity "for immediate or for recent memory."

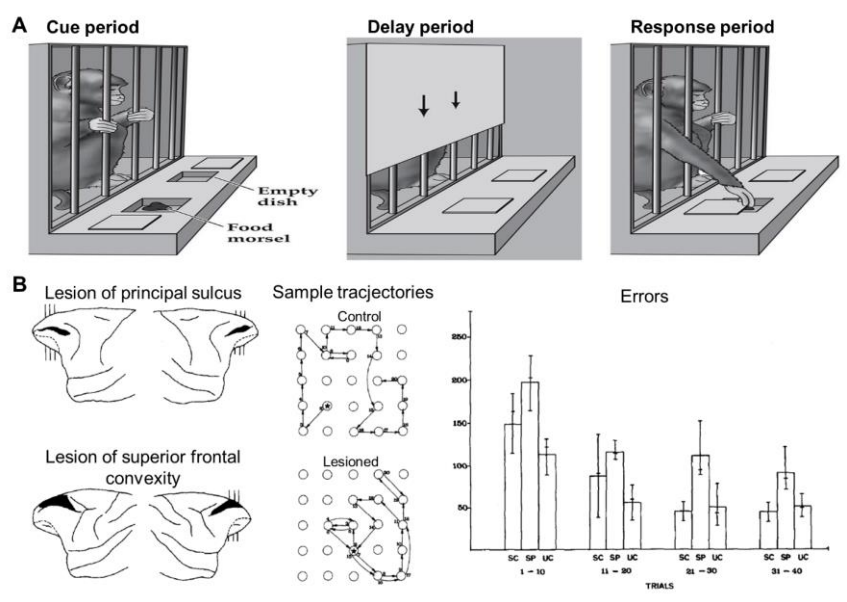


FIGURE 1.6

A CAUSAL ROLE OF THE PREFRONTAL CORTEX IN WORKING MEMORY.

(A) Illustration of the delayed-response paradigm. The delayed-response paradigm as well as its variants (e.g., memory-guided saccade task) are one of the most widely used groups of tasks to study working memory in animals, but also humans. In this particular version of the task, a food reward is first placed in a randomly selected well that is clearly visible to the monkey (cue period). A screen is then lowered for a delay period of several seconds, during which the monkey no longer receives any direct visual input about the placement of the reward and therefore has to hold this location in working memory. Once the screen is raised, the monkey has one shot to choose the correct location to receive the reward. If he succeeds, he must have held a representation of the location in his working memory.
(B) Empirical evidence for the role of the prefrontal cortex in working memory. (Left) Here, monkeys were first lesioned either in the principal sulcus (SP; upper brain) or the superior frontal convexity (SC; lower brain), or were not lesioned at all (UC). They then had to perform a search task, in which they had to retrieve as many food rewards as possible from 25 locations, being only allowed to visit each location once. (Middle) Sample trajectories for two representative monkeys. The starting location is indexed by a star. (Right) Number of errors as a function of lesion status and trial. It is immediately evident that lesioned monkeys committed far more errors than the non-lesioned controls. Crucially, the same lesioned monkeys were perfectly capable of performing the task, when the working memory requirement was removed, suggesting that prefrontal cortex plays a specific role in working memory processes. Adapted from Passingham (1985).

How was he able to reach that conclusion? Let us first take a close look at the type of task he used (Figure 1.6A). A monkey is presented with a food source, randomly placed in one of two wells before his eyes. In order to receive this reward, he will have to keep this location in his mind, even when he can no longer see it. In other words, he will have to store this location in his working memory to correctly perform the task. Jacobsen (1935) noted that lesioned animals were no longer capable of performing this kind of spatial *delayed-response task*, while they showed no impairments in similar tasks without a working memory requirement. Further lesion work in primates (Butters and Pandya, 1969; Goldman and Rosvold, 1970; Levy and Goldman-Rakic, 1999; Malmo, 1942; Passingham, 1985; Petrides, 1995) and humans (Manes et al., 2002; Owen et al., 1990) confirmed the disruptive effects of lesions in the dorsolateral prefrontal cortex on performance in these types of working memory tasks, leading to the idea that this particular region of the brain served as a dedicated storage buffer for representations held in working memory (Figure 1.6B).

It quickly turned out that this initial conclusion may have been a bit too rushed. Just because damage to a given brain area (in our case, the dorsolateral prefrontal cortex) leads to decrements of a particular cognitive function (here, short-term maintenance of information in working memory) this need not

imply that the brain area under consideration is the seat of the cognitive function per se. It could, for instance, also be the case that the lesion prohibited two (or more) brain regions from communicating with each other, thereby interfering with the task at hand. Alternatively, it may also have fuddled with a related mental process, that is necessary (but not sufficient) for the execution of the actual cognitive function. For example, you may imagine how, for working memory specifically, different brain areas might be in charge of the different components, such as the actual storage of information and the control and transformation processes we talked about before (e.g., central executive, attention, rehearsal, manipulation, etc.).

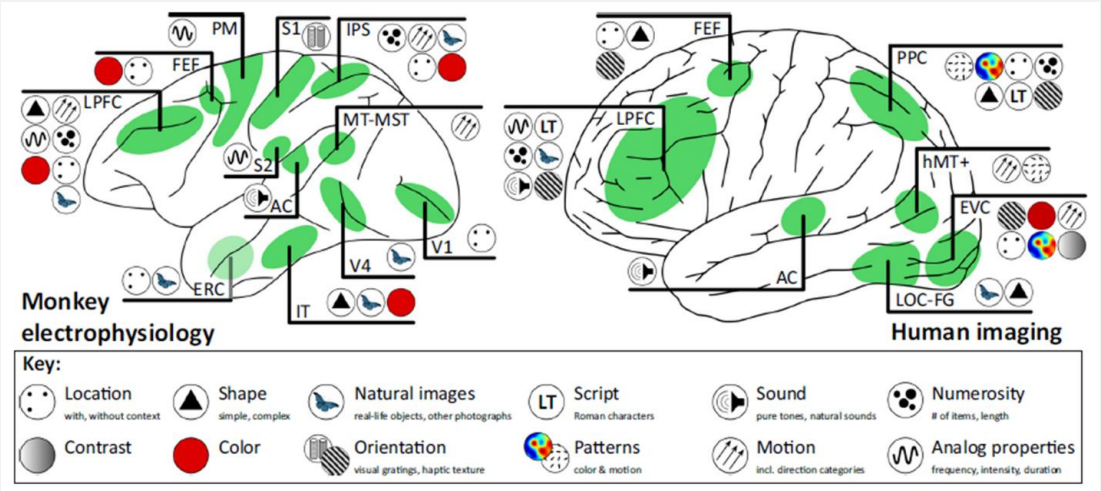


FIGURE 1.7

A DISTRIBUTED NETWORK OF BRAIN AREAS SUPPORTS WORKING MEMORY.

Illustration of major cortical areas exhibiting stimulus-selective activity during working memory delay periods in the macaque (left) and human brain (right). While early findings pointed towards a dedicated system for working memory in the dorsolateral prefrontal cortex (LPFC), subsequent studies reported memory-related activity in a multitude of brain areas, leading to the idea that working memory functions in the brain are subserved by a highly distributed network of brain regions. AC = auditory cortex; ERC = enthorinal cortex; EVC = early visual cortex; FEF = frontal eye fields; FG = fusiform gyrus; hMT+ = human analog to MT/MST; IPS = intraparietal sulcus; IT = inferior temporal cortex; LOC = lateral occipital complex; LPFC = lateral prefrontal cortex; PM = premotor cortex; PPC = posterior parietal cortex. Adapted from Christophel et al. (2017).

This indeed turned out to be the case. On one hand, evidence from further lesion studies suggested that the dorsolateral prefrontal cortex may not be implicated in the storage of working memory representations per se (D'Esposito and Postle, 1999; D'Esposito et al., 2006), but may rather subserve the monitoring and manipulation of the current contents of working memory (Barbey et al., 2013; Petrides and Milner, 1982; Tsuchida and Fellows, 2009). Brenda Milner (1982), for instance, observed that, following lesions to their dorsolateral prefrontal cortex, patients exhibited performance decrements on a short-term recognition task only when the stimulus material was reused across trials as opposed to when a non-repetitive set of memoranda was utilized. The lesions thus did not produce a general deficit in short-term retention (as they should if information held in working memory were stored in this area of the brain), but seemed to be more specifically involved in resolving distraction from previously relevant memories. On the other hand, over the course of the years, many more brain regions were found to be implicated in working memory (Figure 1.7). Directly recording single- or multi-unit neural activity as well as local field potentials in the monkey brain, for instance, revealed additional contributions of other frontal areas, such as the frontal eye fields (Buschman et al., 2011) and the pre-motor cortex (Lemus et al., 2009; Vergara et al., 2016), as well as of more posterior areas, including inferior temporal cortex (Fuster and Jervey, 1981; Miller et al., 1993) and primary visual cortex (van Kerkoerle et al., 2017). Similarly, neuroimaging studies with human participants consistently reported memory-related activity in posterior parietal cortex (Jerde et al., 2012; Sprague et al., 2014, 2016) and early sensory areas (Emrich et al., 2013; Ester et al., 2009; Harrison and Tong, 2009) in addition to prefrontal cortex (Ester et al., 2015; Lee et al., 2013).

The existence of such a distributed network of brain areas activated by different working memory tasks casts considerable doubt on the purportedly specific role of the dorsolateral prefrontal cortex for working memory. If, after all, the dorsolateral prefrontal cortex really serves as a dedicated system for working memory in the brain, then what exactly are all these other regions doing? The development of highly sophisticated, multivariate statistical methods, allowing the identification of item/category-specific patterns of neural activity, may have provided a first approximate answer. Just as we speculated before, there appears to exist a division of labor between different cortical regions when it comes to working memory. On one hand, posterior sensory brain areas appear to be primarily involved in the storage of the actual precise contents of working memory, with different types of memoranda recruiting the corresponding brain regions. For example, if you were asked to remember low-level visual features, such as orientation, color, or motion, your primary visual areas would temporarily maintain your memories (Christophel et al., 2012; Harrison and Tong, 2009; Riggall and Postle, 2012; Serences et al., 2009). If, by contrast, the stimulus material were auditory in nature, the corresponding representations would be stored in your primary auditory cortex (Huang et al., 2016; Kumar et al., 2016). On the other hand, there is now mounting evidence that the role of dorsolateral prefrontal areas during working memory primarily reflects "top-down" attentional control and management processes, serving to bias stimulus-specific activity in sensory regions in the service of goal-directed behavior (Curtis and D'Esposito, 2003; Sreenivasan et al., 2014; Warden and Miller, 2010). Quentin and colleagues (2018), for instance, very recently observed a dissociation between ventrolateral prefrontal cortex coding (and storing) representations of task rules, and a distributed network of occipito-parietal brain areas specifically maintaining the to-be-remembered contents. Consistent with the state-based cognitive models of working memory we discussed before, a highly distributed network with different functional specializations thus seems to underlie working memory in the brain.

1.2.4.2 Stable, persistent neural activity or dynamic, activity-silent processes as a candidate mechanism for working memory?

Let us now switch gears a little bit and turn to the consideration of the second question I raised at the outset of this section. What we have seen so far is that the orchestrated action of a highly distributed network of brain areas appears to underlie the storage and manipulation of information currently held in working memory. We still do not know, however, which specific neural mechanism(s) might support these working memory functions. In order to address this problematic, we will have to shift our focus away from simply trying to localize the seat of working memory in the brain, and instead look more closely at the actual dynamics of neural activity during typical working memory tasks.

One of the most fundamental concepts in the neurobiological study of working memory is the idea that information is maintained by persistent neural activity. That is, even in the absence of external, sensory stimulation, neurons selective for the to-be-remembered information continue to generate activity until that information is no longer needed. According to this view, if, for example, your partner had sent you to the grocery store without a shopping list to pick up some carrots, nuts, and cinnamon, those neurons (or those populations of neurons) in your brain coding for these items (or, perhaps, for the more abstract notion of carrot cake), would have to remain active until you retrieved each ingredient from its shelf.

Some of the earliest support for this notion came from electrophysiological experiments in awake, behaving animals. In their seminal work, Fuster and Alexander (1971) recorded extracellular activity from single units in the prefrontal cortex in five rhesus monkeys performing the type of spatial delayed-response task we discussed before, and observed that some of the neurons increased their firing rate specifically during the delay period (i.e., that stage of the task, during which the to-be-remembered stimulus was no longer physically present, yet monkeys had to maintain an active representation of it in order to complete the task successfully). You can see in Figure 1.8A how, over five consecutive trials, the neuron depicted here systematically seems to bridge the gap between the presentation of the memorandum and the execution of the response. Similar electrophysiological findings were also reported with different

experimental paradigms, such as the delayed alternation task (requiring monkeys to alternate lever presses; Kubota and Niki, 1971) or oculomotor versions of the delayed-response task (requiring monkeys to make a saccade to a spatial location held in working memory; Funahashi et al., 1989).

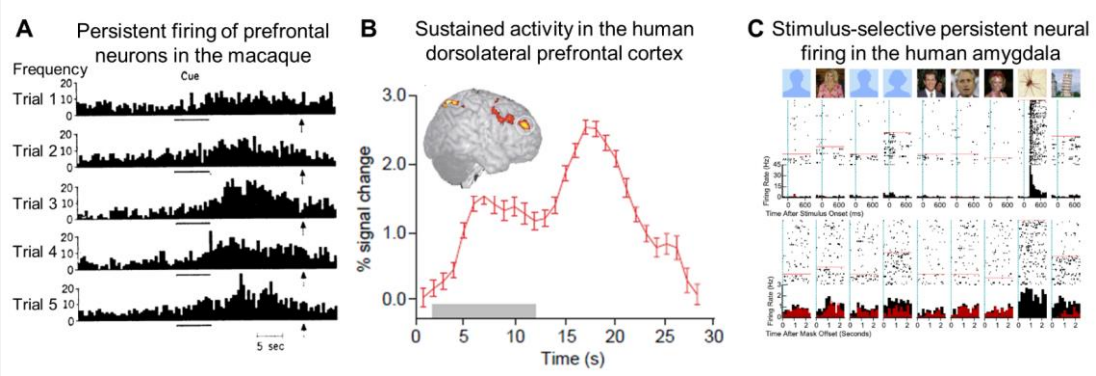


FIGURE 1.8

EMPIRICAL EVIDENCE FOR SUSTAINED NEURAL ACTIVITY AS THE CORRELATE OF THE WORKING MEMORY ENGRAM.

(A) Single unit recordings in five consecutive trials from a representative neuron in the prefrontal cortex of a macaque during the execution of a spatial delayed-response task. The horizontal line corresponds to the presentation of the memory item, the arrow marks the beginning of the response period. It is clearly evident that, starting from the cue period, this cell increased its firing rate until the termination of the response. Adapted from Fuster and Alexander (1971).
(B) Time-course of BOLD signal in the human dorsolateral prefrontal cortex during the performance of an oculomotor delayed-response task. Gray bar denotes delay period. Again, sustained activity is present during the maintenance period. Adapted from Curtis and D'Esposito (2003).
(C) Single neuron activity of one representative neuron from the human amygdala demonstrating stimulus-selective persistent neural firing. Adapted from Kornblith et al. (2017).

How did these early studies interpret this sustained neural activity? It might surprise you that these authors (in particular the ones of the first two papers) actually did not equate the persistent firing with the storage of representations in working memory (as is typically assumed to be the case today). They instead suggested that it reflected the animals' sustained attention towards the internally stored representations of the to-be-remembered information, thus corroborating the contemporary views on prefrontal cortex function we talked about before. Subsequent investigations, however, were able to provide increasingly convincing evidence in favor of the role of persistent neural activity as a genuine neural correlate of the working memory engram (Goldman-Rakic, 1995): They, for instance, ruled out possible alternative explanations for the sustained neural firing observed in the original studies (e.g., preparation of upcoming motor response) and, crucially, demonstrated that it was in particular those neurons that were selective for the current content of working memory that exhibited this kind of persistent delay-period activity (Funahashi et al., 1989; Miller et al., 1996). Since then, sustained working-memory related activity has been observed in many brain imaging studies in humans (Courtney et al., 1997, 1998a, 1998b; Jansma et al., 2000; Sakai et al., 2002; Figure 1.8B) and stimulus-selective sustained neural firing has very recently even been reported in single neurons of the human medial temporal lobe (Kamiński et al., 2017; Kornblith et al., 2017; Figure 1.8C) as well as populations of neurons in the prefrontal cortex (Haller et al., 2018). In their totality, these findings thus support an extraordinarily appealing view of the nature and neural substrates of working memory: In order for us to keep a stable thought in mind, neurons coding for the respective information will have to remain active until that information is no longer needed. What could possibly be more beautiful than that?

After all, the story might not be quite that simple. If sustained neural activity truly were the only mechanism supporting the short-term storage of information in working memory, then, if ever it were to be prematurely terminated or abolished, the memory should be lost. However, this turns out not to be the case. Watanabe and Funahashi (2014), for example, recorded single neuron activity from the prefrontal cortex of two macaques during the performance of a dual-task requiring the temporary memorization of a spatial location in addition to the focusing of attention on a spatial position. Though the attention-

demanding, secondary task did not interfere with the monkeys' ability to accurately remember the relevant spatial location, it did interrupt the content-specific delay-period activity associated with the memorized position. The neurons' spatial selectivity for the memory cue dropped dramatically during the dual-task portion of the task, only to re-emerge anew following the successful completion of the secondary task. Even more striking examples than this temporary disruption of the chain of neural firing also exist. Directly requiring monkeys to simultaneously remember two (or more) spatial locations while again recording spiking activity and local field potentials (LFP) in prefrontal cortex, Lundqvist and colleagues (2016) found no evidence for persistent neural activity whatsoever. Instead, at the single-trial level, memory contents appeared to be maintained via intermittent, discrete bursts of gamma oscillations, accompanied by a decrease in beta-burst probability (Figure 1.9A). A follow-up study with even more complex stimulus materials (i.e., sequences of spatial patterns) recently confirmed these initial findings

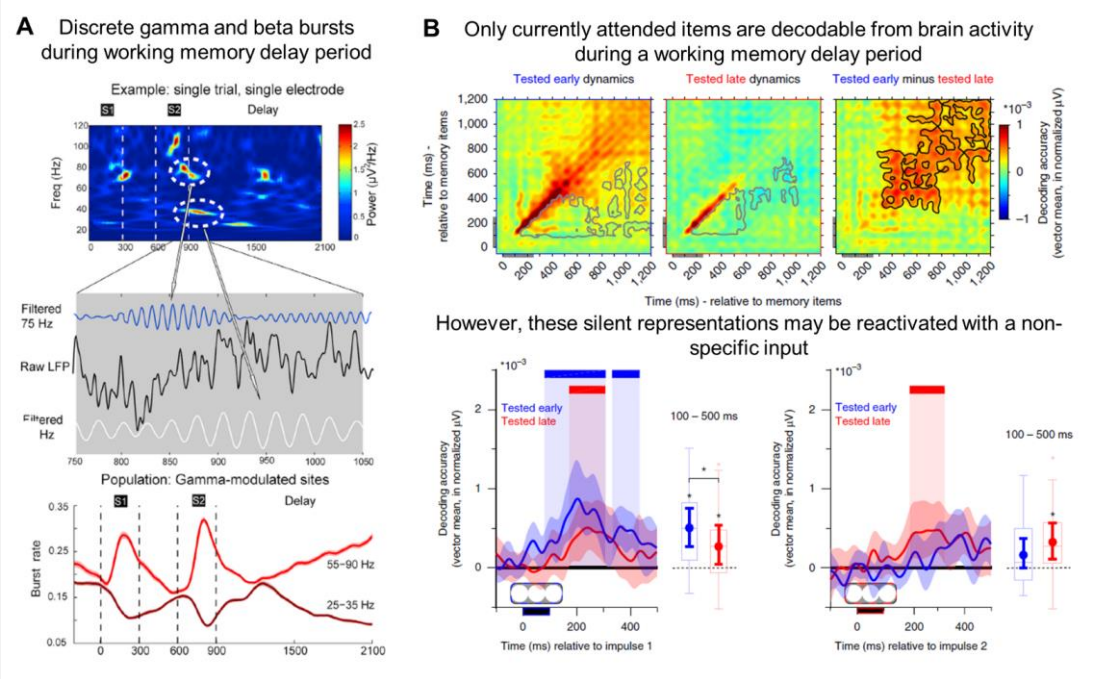


FIGURE 1.9

EMPIRICAL EVIDENCE FOR DYNAMIC, ACTIVITY-SILENT BRAIN STATES DURING WORKING MEMORY MAINTENANCE.

(A) Monkeys were trained to retain two or more spatial locations in working memory. (Top) During the working memory delay period, there was no evidence for sustained activity at the single-trial level. Instead, there were only brief bursts in gamma and beta frequency bands associated with encoding and reactivation of the information. (Middle) Zoom-in on the raw local-field potentials displayed in the upper graph. (Bottom) Gamma-burst probability increases towards the end of the delay period, while, simultaneously, beta-burst probability decreases. Together, these results suggest that information in working memory may be maintained in the absence of sustained neural activity. Adapted from Lundqvist et al. (2016).
(B) Human participants performed a working memory task, for which they had to retain and report two items. Crucially, one of the items had to be recalled a couple of seconds before the other, such that, for the first half of the delay, only the first item, and for the second half of the delay, only the second item was task-relevant. (Top) During the first half of the delay, only the first item can be decoded from brain activity. The second seems to have vanished. (Bottom) A non-specific signal, however, reactivates decodability of all generally task-relevant items during the first (left) and second (half) of the delay. As such, it appears as if currently attended items are being stored in neural firing, while currently unattended ones might rely on "activity-silent" mechanisms. Adapted from Wolff et al. (2017).

(Lundqvist et al., 2018), and complementary observations have also been made in humans using both functional magnetic resonance imaging (fMRI; Lewis-Peacock et al., 2012; Rose et al., 2016; Sprague et al., 2016) and time-resolved electroencephalography (EEG; LaRocque et al., 2013; Wolff et al., 2015, 2017; Figure 1.9B). While currently attended representations appear to be stored in patterns of neural firing, unattended, yet still task-relevant, items do not seem to require accompanying neural activity. At least in some instances, then, storage of information in working memory clearly occurs in the absence of sustained neural firing.

The partial or complete absence of persistent delay-period activity is not the only challenge models of

sustained neural activity in working memory have to grapple with. Another problem is related to the stability of the mental representations themselves. If our memories were really stored in persistent neural activity, then one might imagine that this firing should be fairly consistent and stable over the course of such delays. However, especially in the prefrontal cortex, this is also not the case: Even over the course of a single trial, neurons may adapt their coding preferences (Buonomano and Maass, 2009; Cavanagh et al., 2017; Spaak et al., 2017). For instance, neurons in the prefrontal cortex of the macaque have been found to dynamically change their tuning profiles in order to accommodate changes in the behavioral context, first only coding for the physical properties of the stimuli and then transitioning to a representation of decision-related features (Stokes et al., 2013). Even more recently, Parthasarathy and colleagues (2017) showed that the representations stored in the lateral prefrontal cortex of two monkeys during a delayed saccade task were surprisingly flexible. Initially, representations reflected only the current contents of working memory. However, after the presentation of a distractor stimulus, this memory code was morphed with the distracting location, albeit without losing any task-relevant information. Neural activity is therefore much more dynamic than previously thought.

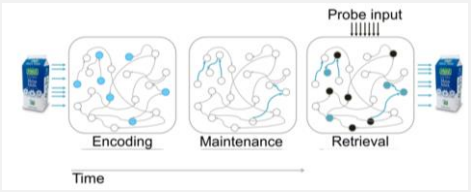


FIGURE 1.10

THE ACTIVITY-SILENT DYNAMIC CODING FRAMEWORK FOR WORKING MEMORY.

Information in working memory may be stored in dynamic, hidden states. Consider a population of neurons encoding a specific item (left column). Short-term synaptic plasticity links the active neurons coding for the respective memorandum (thick arrows), allowing the cell assembly to go dormant during the delay (middle column). When a non-specific recall signal reactivates the neural population, the response will be patterned according to the previous input, thereby allowing the information to be retrieved (right column). Adapted from Stokes (2015).

If neurons thus do not always fire persistently during working memory delay periods, but rather exhibit dynamic patterns of activity along with periods of inactivity, how exactly can we even keep a coherent, stable thought in mind? An exciting, new development is the idea that storage in working memory may be supported by temporarily restructuring the functional connectivity of neural networks through short-term synaptic plasticity (Barak and Tsodyks, 2007; Fiebig and Lansner, 2017; Mongillo et al., 2008; Stokes, 2015; Sugase-Miyamoto et al., 2008). Sounds complicated? The basic idea is fairly simple: Consider a population of neurons that codes for a specific concept, such as a bottle of milk (Figure 1.10). On top of the existing structural connections (light arrows), short-term synaptic plasticity effectively links the active neurons (thick arrows), allowing the cell assembly to go dormant during the delay. When a non-specific recall signal reactivates the neural population the response will be patterned according to the previous input. Information may thus be stored in working memory without any accompanying neural activity in so-called *activity-silent brain states*. Crucially, this mechanism offers a much more *dynamic* view of working memory than do classical models based on persistent activity: An otherwise flexible neural network has been modified so as to bias the processing of subsequent input in accordance with the information currently held in working memory. Memories are therefore no longer just still-frame snapshots of the past, but rather our brain's response potential to *future* input. Currently, we still know very little about how exactly such dynamic, activity-silent mechanisms might cooperate with more persistent patterns of neural activity to subserve the storage and manipulation of information in working memory. A first few attempts have been made to reconcile the dynamic, activity-silent framework with more traditional models, for instance, drawing the distinction between (1) the maintenance of attended vs. unattended information (Rose et al., 2016; Sprague et al., 2016; Wolff et al., 2017), (2) the storage of task rules vs. stimulus information (Quentin et al., 2018), or (3) the maintenance vs. the manipulation of representations (Masse et al., 2018), but an overarching theoretical perspective is still lacking. The future of working memory research is therefore certainly ripe with major discoveries.

1.3 Consciousness – our subjective experience of the here and now

So far, we have focused our discussion almost entirely on working memory, its properties, cognitive nature, and neurobiological bases. However, when you go back to our story about Henry, you may recall that it was not only about his memory. It was also about his ability to – in spite of his devastating handicap – seize the moment, to fully experience and enjoy the here and now. In a sense, it was the tale of a man whose life was restricted to a permanent present tense, to an eternal succession of fleeting, subjective experiences.

For many of us, layman and scientist alike, the mere existence of these inherently subjective, conscious experiences remains one of the great mysteries of our universe (Adolphs, 2015). How is it possible that a physical organ, such as the brain, can generate any subjective experience at all? There are, after all, many highly complex machines and organisms that do not appear to have any conscious sensations. Take your smartphone as an example. No matter the exact model, chances are that you have at least a couple of apps running that measure the current state of wellbeing of your phone: There is the quintessential application tracking your smartphone's battery life, temperature, and remaining storage capacity – to name but a few. Despite this abundance of information, however, I bet you anything your smartphone has never expressed any concern or insight about its present state. Why is it that we tend to feel tired, when our battery runs out at the end of the day, or feel at the very least extremely uncomfortable, when our body temperature rises, yet that your phone does not seem to experience any of these? My goal for this section of the thesis is to introduce you to the scientific study of consciousness, provide a brief overview over some of the most prominent cognitive and neurobiological models of conscious perception, and highlight the non-conscious processing capacities of our brain.

1.3.1 A scientific approach to the study of consciousness

1.3.1.1 A brief history of consciousness science

For readers who are not that familiar with the topic of my research, it might have been surprising to hear a scientist talk about consciousness. Should this not rather belong to the domain of philosophy? For a good part of history, this was actually the case: As the human body and mind generally tended to be viewed as distinct and separable entities, with the former being physical and the latter metaphysical in nature, considerations about subjectivity and/or consciousness were relegated to the more philosophically inclined. The French philosopher, mathematician, and physiologist René Descartes (1637) is typically credited for this kind of *dualistic* worldview, but the general idea has been around for much longer than that. Almost 2 millennia earlier, Plato (360 BC) had already conceived of human beings as immortal souls trapped in mortal bodies. Aristotle (ca. 350 BC) and Thomas Aquinas (ca. 1250) both also defended somewhat weaker forms of dualism, before in 1637, René famously attributed the pineal gland as the connecting organ between the body and the soul. If you now think that these somewhat mystical concepts are all remnants of our distant past, you are mistaken. Contemporary conceptualizations of an immortal soul form part and parcel of many religious beliefs and schools of thought (e.g., Christianity, Hinduism, etc.), and are still actively proposed by many acclaimed philosophers (and other scientists; Chalmers, 1997; Popper and Eccles, 1984).

What, then, caused the shift in our thinking, turning the question of subjectivity and consciousness into a valid subject of scientific inquiry? Perhaps one of the most defining moments in the history of psychology (and consciousness research) was the realization that "the mental world can be grounded in the physical world by the concepts of information, computation, and feedback" (Pinker, 2002). With the dawn of the *cognitive revolution* in the 1950's, not only had the workings of our inner, mental life all of the sudden been admitted as a veritable topic for scientific investigation, but they had also become amenable to the scientific method itself. Just like a computer, the mind was likened to an information-processing device, receiving an input, using and/or modifying this information, and producing an output. To infer the

functional architecture of the human mind as a whole and dissect individual mental processes, one thus only had to manipulate the system's inputs and observe the effects of such variations on its outputs. A second driving factor may have been the fractionation of consciousness itself. Just ~20 years ago, the philosopher David Chalmers (1995) proposed a distinction between the *easy* and the *hard problem* of consciousness. The former, he argued, "only" requires an explanation of the properties of consciousness (e.g., integration of information, reportability, deliberate control of behavior, etc.) and is, as such, immediately susceptible to the standard methods of psychology and neuroscience because it can be reduced to computational or neural mechanisms. By contrast, the hard problem would necessitate an account of our phenomenology itself, essentially addressing the issue of how (and why) there is any subjective experience at all. Chalmers considered this to likely be intractable. While thus perhaps a bit of a pessimistic opinion, Chalmers' perspective drew attention to the fact that consciousness may be broken down into smaller constituents, thereby paving the way for contemporary consciousness science.

1.3.1.2 What is consciousness?

If I am not actually going to be talking about consciousness in laymen's terms, then what exactly am I referring to? Let us take a closer look at Figure 1.11. You immediately notice that consciousness appears to vary along at least two dimensions (Dehaene and Changeux, 2011; Laureys, 2005): The first concerns your current *state of arousal* (or vigilance) and denotes your general level of wakefulness. It is thought to cover the entire spectrum between a fully comatose and a fully vigilant state, and is, obviously, of great clinical importance. The second dimension, on the other hand, reflects the current *contents of your consciousness* (or awareness), the kinds of materials your conscious mind currently has *access* to. At first sight, this latter element may seem a bit odd. Do we not usually have conscious access to every aspect of our waking life? Though, intuitively, this indeed appears to be the case, this notion of unlimited conscious access is actually flawed. After all, how would you be able to know about something if you did not have prior conscious access to it already? Listen up for a moment: Were you aware of the ticking of the clock, the buzzing of the cars on the street, or the humming of your refrigerator before I drew your attention to it? Did you feel your back press against the chair, couch, pillow, or whatever other mode of support you may have chosen while reading this dissertation? Probably not – until just now. The point that I would like to make here is simple: At any given moment, our brain receives and processes much more information than anyone of us is currently consciously aware of. We will talk about the depth of such non-conscious processing in much more detail later on. For now, just remember that there exist different degrees of awareness, and that it is this very idea of *access consciousness* that I will be focusing on for the remainder of this work.

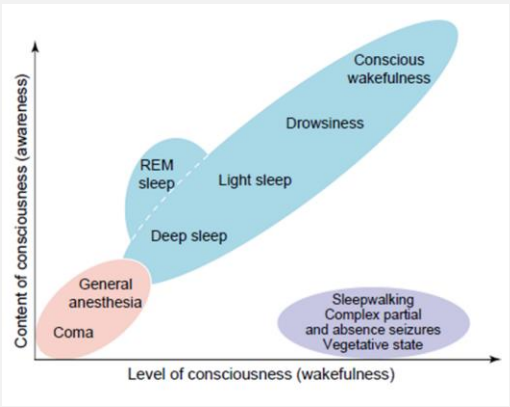


FIGURE 1.11

THE MULTI-DIMENSIONAL NATURE OF CONSCIOUSNESS.

At its simplest, consciousness may be conceptualized to vary along two dimensions: awareness (i.e., subjective report) and arousal (i.e., wakefulness). Adapted from Laureys (2005).

1.3.1.3 How do we manipulate consciousness?

Even though we now have a working definition of consciousness, this still leaves the thorny issue of how exactly we are going to manipulate it experimentally. The majority of research on consciousness tends to focus on conscious visual perception. A first reason for this relative overemphasis on vision as compared to any of our other senses is certainly related to the field's general bias. Out of all our senses, vision has

been the one most heavily studied and, as such, we already know a fair amount about where (and how) the brain processes visual inputs (e.g., Felleman and Van Essen, 1991).

Even more importantly, however, visual stimuli lend themselves fairly easily to create ideal experimental situations. There exist many visual illusions and experimental techniques that allow researchers like myself to carefully manipulate conscious perception without even having to change the physical input the eyes (and brain) receive (Kim and Blake, 2005). Take a look at the image displayed in Figure 1.12A. Did you notice how your perception was constantly fluctuating between two different percepts? Sometimes you were seeing a man playing a saxophone, sometimes the face of a young woman. This kind of *bistable perception* results from the brain's propensity to construct subjective reality from noisy and inherently ambiguous sensory input (Helmholtz, 1866), leading to situations such as these, in which two equally likely interpretations compete against each other for conscious access.

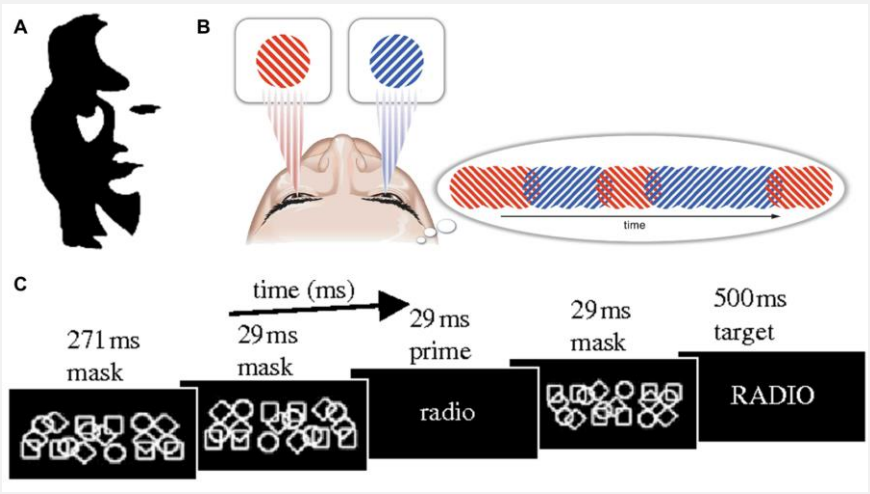


FIGURE 1.12

OVERVIEW OVER SOME EXPERIMENTAL TECHNIQUES TO MANIPULATE CONSCIOUS PERCEPTION.

(A) An example of a visual illusion inducing bistable perception. Here, you either perceive the saxophone player or the face of a woman, but not both at once.
(B) Schematic illustration of binocular rivalry. Two different images are presented to your two eyes. Instead of perceiving a fused image, you will, once again, experience a continuous fluctuation of the two different percepts.
(C) An example of a masking paradigm. The target stimulus (i.e., "radio") is surrounded (in space and time) by another highly salient visual stimulus. As such, the target word will be perceived on some trials, but not on others. Note that, for each of the techniques displayed here, the physical input to the eyes is kept constant, while conscious perception will alternate. As such, one is able to create a minimal contrast between conscious and non-conscious conditions, with the main difference being attributable to changes in visual awareness. Adapted from Kouider and Dehaene (2007).

Crucially, this phenomenon can be induced experimentally in the laboratory, using a technique known as *binocular rivalry* (Logothetis et al., 1996; Figure 1.12B). Here, two different images will be presented to the two eyes. In contrast to what one might expect, one will not perceive a single, fused image, but an individual's perception will instead alternate between the two possibilities, just as was the case for the visual illusion. If one of the images shown consisted of a series of continuously flashing shapes or Mondrian patterns, however, the input to the other eye would remain invisible for a much longer time period than in the standard version of this paradigm (i.e., *continuous flash suppression*; Tsuchiya and Koch, 2005). Tasks, such as these, have played an important role for research on consciousness, permitting, for instance, to evaluate which factors might improve (or degrade) access to consciousness (Jiang et al., 2007; Tsuchiya et al., 2009; Yang et al., 2007) or to interrogate the neural correlates of non-conscious processing (Yang et al., 2014). They also can be easily adapted in other primates (e.g., Leopold and Logothetis, 1996) and, as such, open the door for comparative studies on conscious perception.

A second class of techniques possesses the capability of rendering a stimulus fully invisible. *Visual masking* essentially degrades the quality of the input signal to such an extent that it will severely reduce or even eliminate subjective awareness. Here, additional, highly salient stimuli are shown in close temporal (i.e., *forward* or *backward masking*) and spatial contiguity (i.e., *metacontrast* or *pattern masking*) to the actual target stimulus (Breitmeyer and Öğmen, 2006; Enns and Di Lollo, 2000). In the example displayed in

Figure 1.12C, for instance, the lower-case word "radio" has been surrounded by two mask stimuli, each composed of a set of irregular shapes. Had it stood on its own, even if only being flashed on a computer screen for 30 ms, you would have had no trouble seeing it whatsoever. However, due to the mask, chances are you would have missed it entirely.

An equally effective, yet methodologically different technique to interfere with conscious perception consists in directly manipulating attention. In experiments relying on the *attentional blink*, two target stimuli are embedded in a sequence of otherwise irrelevant distractors, flashed one by one on a screen. Participants, whose task consists in reporting both targets, will typically miss the second if the delay between the two was sufficiently short (i.e., < 50 ms) – most likely because their attention was still directed towards the processing of the first target by the time the second arrived (Raymond et al., 1992). Perhaps even more spectacular is the case of *inattentional* or *change blindness* (Simons and Ambinder, 2005; Simons and Chabris, 1999). If you are not familiar with this terminology, watch this short clip on youtube before reading any further. As you have just experienced yourself, sometimes people fail to notice even fairly large changes to visual scenes and might miss entire objects altogether (such as a gorilla walking amidst a crowd of ball players). Lack of attention certainly plays a rather prominent role for such phenomenal lapses of conscious perception, but prior expectations most likely contribute as well.

Though thus rather diverse in terms of experimental setup and underlying cognitive and neural mechanisms, all of these methodologies permit to manipulate conscious perception in a highly controlled manner. The physical properties of the input stimuli may, for instance, be kept constant for the duration of the experiment, thereby minimizing potential sources of noise and variance. Moreover, by carefully choosing the experimental parameters (e.g., contrast of the mask, temporal delay between the two targets for the attentional blink, etc.), participants' perception may be titrated to a very specific threshold, with a certain percentage of detected and undetected target stimuli. These latter points turn out to be of tremendous importance. One of the most frequently used approaches in the scientific study of consciousness consists in creating a *minimal contrast* between conscious and non-conscious conditions, with, in ideal circumstances, the only difference between the two being the subjects' conscious perception (Baars, 1994). Both the behavioral and brain responses to conscious and non-conscious stimuli may then easily be compared with each other, thereby permitting researchers like myself to characterize conscious and non-conscious processing and to identify the neural correlates of conscious perception (Koch, 2004).

1.3.1.4 How do we measure consciousness?

If you have been following me until now, you may have noticed that an important element still seems to be missing from our ideal study on consciousness. So far, we have defined our terminology and discussed a variety of experimental approaches to manipulate conscious perception. But how are we to know whether our subjects perceived something consciously or not? How do we know whether they saw the masked target that was flashed on the screen, or whether they detected the second stimulus during the attentional blink? To differentiate between seen and unseen (or between heard and unheard, etc.) trials, we will have to have a direct measure of our participants' subjective experience.

There exist two general approaches to address this issue: Either we trust our subjects' own introspection and directly ask for a *subjective report*, or we rely on *objective measures*, such as chance performance. Perhaps the most intuitive option, given the nature of the topic, consists in openly asking our participants about their perceptual experience. At the end of each trial, we may, for instance, demand subjects to report their visibility of the target on a binary (i.e., seen vs. unseen), discrete (i.e., clearly seen, weakly seen, glimpse, not seen) or quasi-continuous scale (e.g., from 0/unseen to 100/seen). Intriguingly, it turns out that, even if given the opportunity to report extremely fine variations in visibility, most individuals still opt for a binary visibility judgement, largely ignoring intermediate ratings (Sergent and Dehaene, 2004). Popular measures of subjective visibility therefore still tend to focus on largely restricted scales, offering between two and four distinct categories (e.g., *Perceptual Awareness Scale*; Ramsøy and

Overgaard, 2004). Second-order commentaries, such as confidence ratings, in which subjects indicate their confidence in their decision (Cheesman and Merikle, 1986; Dienes et al., 1995; Lau and Passingham, 2006), or post-decision wagers, in which participants place a monetary bet on the accuracy of their decision (Persaud et al., 2007), may substitute these purely introspective, first-order visibility ratings in order to further increase the subjects' motivation to respond truthfully (Schurger and Sher, 2008).

The advantages of these types of measures are readily apparent: They constitute the most direct way of assessing our topic of interest (i.e., a subjective experience) and permit analyses at the single-trial level. On the flipside, however, they may also be prone to response bias or might potentially contaminate the associated behavioral and neural responses. Let us focus on each of these issues in turn. First, imagine a situation in which several participants have to perform the same difficult task: A heavily masked, small digit is flashed on the screen for 17 ms, and subjects first have to determine whether the target digit is inferior or superior to *5* and then have to rate their visibility on a 4-point scale. Do you think that all of our subjects will apply the same criteria to decide whether a given stimulus was seen? Probably not. Would you, for instance, judge a digit as seen if you had detected an ambiguous shape on the screen without being able to unequivocally determine its identity? Some participants might, others might not. While some subjects might adopt a more conservative strategy, in which they report even partially seen stimuli as unseen, others might be more liberal in their visibility ratings. Either way, these kinds of response biases introduce a certain level of noise (i.e., error) into our measurements, because both the group of seen and the group of unseen trials might contain a non-negligible portion of trials of the other category. Objective measures, grounded in signal detection theory (Green and Swets, 2000), may partially remedy this dilemma. Here, non-conscious processing is essentially equated with chance-performance on some direct task of stimulus detection or classification. For instance, if, in the previous example, our participants performed at chance in the classification task (i.e., accuracy does not differ significantly from 50%), this would be considered non-conscious processing. Though thus immune to the type of response bias we talked about before, these measures come with their own bag of problems: Their computation requires multiple trials and necessitates a change to the physical properties of the input stimulus when conscious and non-conscious conditions are to be compared.

A second concern over the use of subjective measures that has recently been voiced pertains to the nature of information that may be isolated with such paradigms (Tsuchiya et al., 2015). Remember how most of the research on consciousness relies on a contrast between conscious and non-conscious conditions to isolate just that aspect of our behavior and brain activity that subserves conscious processing (Baars, 1994)? As it turns out, it may not be quite that straightforward. Simply contrasting two conditions with distinct reports (e.g., seen vs. unseen) might potentially confound the neural correlates of consciousness with other, closely related, cognitive functions, such as attention, working memory, or expectation (Aru et al., 2012; Kok et al., 2012; Melloni et al., 2011). A future alternative might be to switch to *no-report paradigms*, for instance inferring subjects' perceptual state from physiological measurements (e.g., eye movements, heart rate, etc.; Tsuchiya et al., 2015), but even these measures will first have to be validated with subjective reports. Any approach chosen will therefore come with its own set of drawbacks and, in the end, it is up to the researcher to decide on the most appropriate course of action.

1.3.2 From cognitive to neurobiological models of conscious access

Now, we are in a position similar to where we were when we began our detailed exploration of working memory: We have talked about what consciousness is, how to manipulate and measure it. But how exactly do psychologists and neuroscientists conceptualize conscious access in cognitive terms? And what is the brain's role in all of this? In this section, I would like to introduce you to some of the most prominent cognitive and neurobiological theories of consciousness to date, setting the stage for empirical findings on (non-)conscious processing.

1.3.2.1 Cognitive models of attention and working memory as precursors for theories of consciousness

Do you still recall how our discussion of cognitive models of working memory began with Broadbent's (1957, 1958) filter model of attention? In a sense, this very same model may also serve as one of the precursor models of conscious access (de Gardelle and Kouider, 2009). Initially processed in parallel, incoming sensory information is thought to compete for access to a capacity-limited central perceptual system (Figure 1.5A). Though not described as such, processing within this central system may be taken as synonymous with conscious processing: At the top of the cognitive hierarchy, it only received the most relevant and pre-processed information, and, itself, was chiefly responsible for higher-order computations (e.g., motor planning, decisions, etc.) and coordination of activity in lower-level systems.

This notion of a centralized, conscious homunculus permeated almost all of the models we have talked about so far. Atkinson and Shiffrin (1968, 1971) specifically endowed their capacity-limited short-term store with a host of control processes (e.g., rehearsal, visualization, decision-making; Figure 1.5B), and Baddeley and Hitch (1974) considered all of the peripheral, modality-specific buffers to be slaves to the actions and decisions of the central executive (Figure 1.5C). A further distinction between automatic and conscious processing was then introduced by Donald Norman and Tim Shallice (1986): In their action-selection model, a "supervisory-attentional system" was in charge of selecting appropriate actions based on the evidence received from independent, sensory processors, thereby again being capable of implementing flexible behavior in a goal-directed fashion.

1.3.2.2 Brief overview over contemporary cognitive models of consciousness

While none of the work we discussed in the previous section directly dealt with conscious perception (or access) per se, it still laid the foundation for most of the contemporary cognitive theories on consciousness. Just like these models distinguished between central and peripheral systems, conscious and non-conscious processing is typically divided along the same lines, with the former being associated with the flexible control of behavior and the latter with automatic hard-wired (re)actions. However, though all of these models provide an excellent description of a general cognitive architecture, they also fall short in their characterization of this conscious, capacity-limited central processor. Sure, its capabilities and responsibilities are fairly clear, but who (or perhaps, rather, what) exactly is this internal observer? Have we learned anything about how this inner homunculus gives rise to subjective, conscious experience? Or have we just avoided the entire discussion by nominating another conscious entity to govern our behavior?

Contemporary cognitive accounts of consciousness typically deal with this kind of issue by introducing distributed cognitive architectures, obviating the need for the existence of such homunculi. Jackendoff (1987) and Prinz's (2000, 2010) *intermediate level theory* of consciousness, for instance, stipulates that conscious perception arises through the attentional amplification of intermediate level representations in the brain. Consider your own current, visual experience. What exactly are the contents of your visual awareness? An adequate description might be that you perceive objects as a whole from a specific vantage point: I, for my part, am mostly aware of the front of my laptop, but also the TV screen in the background. By contrast, none of us ever seem to be aware of the individual elements (or pixels; e.g., edges, oriented lines, etc.) these objects are made up of, nor of their context-independent abstractions (e.g., view-invariant representations, etc.). Yet all of these representations exist somewhere along the visual hierarchy in the brain. Primary visual cortex (i.e., V1), for example, encodes information in retinotopically arranged cells that respond selectively to wavelength, movement, and edges at various orientations (Felleman and Van Essen, 1991; Hubel and Wiesel, 1968), while inferotemporal cortex is largely indifferent to the size, orientation, and position of objects in the visual field and codes for more abstract, view-invariant representations (Booth and Rolls, 1998; Logothetis et al., 1995). Based primarily on neurological evidence, both Jackendoff (1987) and Prinz (2000, 2010) argue that it is in particular representations intermediate between these two extremes that drive our conscious visual experience.

Another and perhaps the most well-known and influential out of all cognitive theories of consciousness is Bernard Baar's (1988) *global workspace theory*. Here, conscious perception (and consciousness more generally) is thought to result from the global broadcasting of information to a large audience of non-conscious processors. Just like his predecessors, Baars (1988) envisions an army of specialized, non-conscious processors compete in parallel for access to a capacity-limited central stage (i.e., global workspace). Local coalitions between these independent modules may be formed in order to increase their voice in this competition. If successful, the winning processor (or coalition) will dominate the workspace in a winner-take-all fashion, allowing it to recruit further processors, transmit its information globally, and make it available to conscious experience. The losing processors, by contrast, will continue to work in isolation, thereby remaining inaccessible to consciousness.

1.3.2.3 Searching for the neural correlates of subjective, conscious experience

As you have already seen in the case of working memory, any good theory should not only be grounded in cognitive, but also in neurobiological terms. However, we have not yet talked at all about how a physical organ, such as the brain, may give rise to subjective, conscious experience.

Recall some of the major dividing lines surrounding the neural substrate(s) of working memory. There used to be a debate about whether this particular cognitive function was subserved by a dedicated neural system or a widely distributed network of brain areas. A somewhat similar division may also be drawn when it comes to consciousness. Though the consensus largely points towards a distributed view of consciousness in the brain, isolated "localist" accounts have also emerged. Given the general emphasis of the field on vision, most of these models tend to focus on some part of the brain recruited during visual processing. A complete review of all of these theories is clearly well beyond the scope of this introductory chapter here, but I would nevertheless like to briefly discuss some of these areas (and models).

One of the first structures in the brain to process visual input is the thalamus or, to be more precise, the *lateral geniculate nucleus* of the thalamus. Strategically located between the retina and the cortex, it serves as a relay station, sending retinal information to the primary visual cortex (Koch, 2004). Most importantly for the intents and purposes of our current discussion, the thalamus itself is tightly and reciprocally connected to almost all other major cortical areas (Jones, 2002), thereby being extremely well positioned to integrate a wide variety of cortical computations. Moreover, it appears to be implicated in the regulation of arousal, being a frequent locus of brain damage in vegetative-state patients (Adams et al., 2000; Kinney et al., 1994) and of the actions of general anesthetics (Alkire and Miller, 2005; Alkire et al., 2000), and its activity is modulated during binocular rivalry (Haynes et al., 2005). Based on these or similar findings, many researchers have proposed the thalamus as the seat of consciousness in the brain. Ward (2011), for instance, argued that, while the cortex is necessary to analyze, compute, and process the contents of consciousness, the thalamus itself generates the associated phenomenological experience through synchronized neural activity. According to this view, then, the thalamus is the only structure in the brain responsible for our subjective, conscious experience.

Once having reached the thalamus, visual information will be gated to the occipital cortex at the very back of the brain. In the context of the intermediate level theory of consciousness (Jackendoff, 1994; Prinz, 2000, 2010), we have already talked about the functional specialization of visual cortex. Neurons towards the lower end of the visual hierarchy appear to respond preferentially to individual features. Indeed, different cortical patches are systematically recruited for the analysis of distinct features from the visual environment: Whereas most of the orientation-selective cells are found in area V2 and V3, area V4 seems to host the majority of color-selective and area V5 the bulk of motion-selective neurons (Zeki, 1978). To make matters even worse, these distinct subgroups of cells also seem to become involved in the generative process of vision at slightly different times. Different attributes of a visual scene are not perceived simultaneously, with color leading orientation (i.e., shape) by ~40 ms and motion direction by ~80 ms (Zeki, 2015). Why then is it still the case that we perceive our surroundings as a unified whole as opposed to a

conglomerate of disjoint features? Semir Zeki's (2003) solution to this problem is quite unusual: According to his view, each of the aforementioned processing sites also functions as a perceptual site, meaning that heightened activity within that particular cortical area leads to conscious experience of the corresponding feature. These individual "micro-consciousnesses" are then bound together post-consciously, that is, only after each of the attributes has already been experienced consciously, to form a "macro-conscious," unified percept. How exactly this binding process is supposed to be accomplished is, unfortunately, not further specified. What is clear, however, is that, here, consciousness is thought to arise from the activity in a number of isolated, localized modules, corresponding to sensory brain areas.

A very last brain region that I would like to touch upon during this brief overview is the prefrontal cortex. Certainly not an area specialized for visual processing, the prefrontal cortex nevertheless receives (and sends) anatomical projections from (and to) all other sensory brain areas and sits atop the information processing hierarchy in the brain (Goldman-Rakic, 1995). Similar to the thalamus, it therefore appears well placed to integrate information from diverse, disparate processors. Moreover, empirical evidence has consistently linked activity in prefrontal cortex with conscious awareness (Dehaene and Changeux, 2011; Kleinschmidt et al., 1998; Lumer, 1998; Lumer and Rees, 1999; Rees et al., 2002). In stark contrast to the theories we have talked about so far, Hakwan Lau (2007; Lau and Passingham, 2006; Lau and Rosenthal, 2011) therefore considers the prefrontal cortex and, in particular, its dorsolateral aspect, to be the seat of subjective, conscious experience. According to his account, sensory (i.e., first-order) representations are not sufficient by themselves to generate conscious awareness. Subliminal stimuli, for instance, though presented below the threshold for conscious awareness, still influence behavior (Hannula et al., 2005; Kouider and Dehaene, 2007) and therefore must have been coded in the brain. Higher-order representations, by contrast, permit a reflexive, second-order view of the corresponding first-order mental states (e.g., the quality of or efficacy with which the first-order, sensory signal was processed) and, as such, give rise to subjective, conscious experience. The dorsolateral prefrontal cortex, as mentioned before, is in an ideal position to accomplish this task and thus thought to be the main locus of consciousness in the brain.

Let us now switch gears a little bit. At the outset of this section, I drew a clear distinction between such fairly uncommon, localist models of consciousness and the more widely accepted accounts. Remember how contemporary theories of working memory consider this cognitive function to arise from the orchestrated action of a widely distributed network of brain areas? Very similar proposals have also been made in the name of consciousness. Highlighting either synchronous oscillatory activity at a specific frequency (Crick and Koch, 2003; Engel and Singer, 2001; Lutz et al., 2002; Melloni et al., 2007; Tallon-Baudry, 2009; Tononi and Koch, 2008) or re-entrant/recurrent feedback processes (Dehaene and Changeux, 2011; Dehaene et al., 2014; Koch et al., 2016; Lamme, 2006; Lamme and Roelfsema, 2000; Tononi and Edelman, 1998; Tononi and Koch, 2015; Tononi et al., 1994, 1998), this class of models as a whole views *communication between disparate populations of neurons* as a central ingredient of conscious processing. Let us take a moment to take a closer look at the three most influential proposals.

How would you describe your conscious, visual experience? If you think like Giulio Tononi and his collaborators – first Gerald Edelman and later Christof Koch – you might say that any conscious experience is coherent (or *integrated*) as well as specific (or *differentiated*; Tononi and Edelman, 1998; Tononi and Koch, 2015). Let us consider each of these descriptors in turn. Take a standard piece of white paper and write down the word "armchair" in uppercase, black letters. Then, grab another piece and spell out, in as much detail as possible, your current visual experience. What did you end up writing? Most likely, at some point, you said something along the lines of: "I see the word armchair." But why did you not experience this compound noun separately, for instance stating that you saw the word "arm" on the left as well as the word "chair" on the right? Tononi and his colleagues would argue that this is because each subjective conscious experience is irreducible, that is, it is *integrated* as a whole. Moreover, this very experience of yours was also highly unique and specific: It happened exactly the way it did, though you can probably think of a myriad variations. Perhaps you completed our little experiment while sipping your morning

coffee in your kitchen, but it very well could also have happened in another location or at another time of the day. In Tononi's language, your experience was *differentiated*, having been composed of a specific set of specific phenomenal aspects.

Importantly, Tononi and collaborators (Tononi and Edelman, 1998; Tononi and Koch, 2015) suggest that there exists a neural substrate of such integrated and differentiated subjective experiences. According to the *dynamic core hypothesis*, re-entrant interactions primarily, though not exclusively so, between posterior thalamic and anterior cortical brain areas lead to the formation of a temporary, short-lived functional cluster (i.e., dynamic core) that is both integrated, gathering a large amount of information across the entire thalamocortical system, and differentiated, having been selected out of a large repertoire of possibilities (Figure 1.13). The amount of *integrated information* that a given system carries (i.e., θ) may be quantified via a precise mathematical formula (Tononi, 2004), with higher values being indicative of "more consciousness." *Integrated information theory* thus actually conceptualizes consciousness as a graded phenomenon, implying that even simple systems possess a certain degree of consciousness, and, as such, has been criticized on the account of promoting panpsychism (Dehaene et al., 2014).

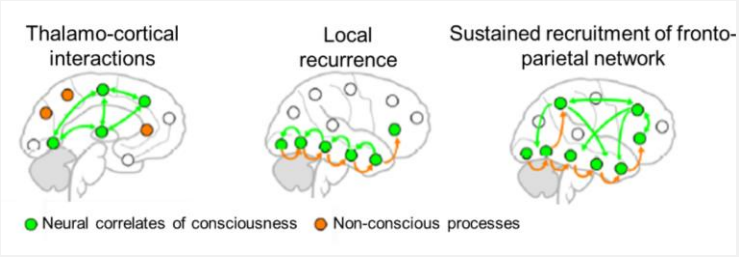


FIGURE 1.13

OVERVIEW OVER NEUROBIOLOGICAL ACCOUNTS OF CONSCIOUSNESS.

There exist many different neurobiological models of consciousness. I here show a select overview of three of the most prominent theories, each of which features the maintenance of information via different neural mechanisms. (Left) Integrated information theory by Tonini and Koch. (Middle) Recurrent feedback loops by Lamme. (Right) Global neuronal workspace by Dehaene and Changeux. Adapted from King (2014).

There is, however, at least some empirical support for this proposal. Changes to the state of consciousness, for instance, tend to be associated with the brain's capacity for information integration. Using high-density EEG recordings in awake, yet differentially conscious patients, King and colleagues (2013) reported increases in long-distance information-sharing capability (as measured by weighted symbolic mutual information) as a function of conscious state: Vegetative-state, minimally conscious, and fully conscious patients could all be distinguished based on this measure alone. Perhaps even more strikingly, Massimini and collaborators (2005) demonstrated, in the same subjects, how the brain's response to a focal, perturbational transcranial-magnetic stimulation (TMS) impulse varied during different states of consciousness. When the participant was awake, the cortex responded with a complex pattern of widespread, and spatially and temporally differentiated activations. By contrast, once the subject had transitioned to a state of deep, non-REM sleep, the cortical response was far more fragmented, remaining locally organized but globally disintegrated. Similar observations have also been made with fMRI in anesthetized monkeys (Barttfeld et al., 2015) and with intracranial recordings in anesthetized patients with epilepsy (Lewis et al., 2012). While thus perhaps not ideally suited as an explanatory mechanism of consciousness in the brain, integrated information surely appears to serve as a viable signature of conscious processing.

A somewhat different stance on the neural basis of consciousness has been taken by Victor Lamme (Lamme, 2006, 2010; Lamme and Roelfsema, 2000; Lamme et al., 2000; Figure 1.13). When visual information first reaches the primary visual cortex, it is rapidly transmitted to all other areas of the visual cortical hierarchy through a cascade of feedforward connections. Though this *fast feedforward sweep* may spread up all the way to motor and prefrontal cortex and may already extract all sorts of behaviorally relevant information (e.g., orientation, shape, color, etc.), by itself, it is not considered sufficient to induce subjective, conscious experience. Masked stimuli, for instance, even if unseen, still entail selective feedforward activation in both visual and non-visual areas (Lamme et al., 2002) and similar feedforward

activity may also be observed in anesthetized, non-conscious animals (Lamme et al., 1998). What, then, is the proposed origin of conscious perception and awareness? Lamme argues that *recurrent processing* via horizontal connections and top-down signaling lies at the heart of consciousness, as it is in particular this kind of neural activity and architecture that permits the sharing, maintenance, and integration of information into a unified, coherent percept.

Perhaps some of the strongest empirical evidence in favor of the role of such recurrent processing for consciousness in humans comes from TMS studies (Lamme et al., 2002; Silvanto et al., 2005). Pascual-Leone and Walsh (2001), for example, applied TMS pulses to visual cortex in order to probe the involvement of back projections from extrastriate area V5 to V1 in motion awareness. When only targeting V5, subjects typically report the perception of a moving flash of light (i.e., phosphene), whereas they perceive stationary phosphenes for hierarchically lower sites of stimulation. A critical test of Lamme's theory therefore consists in disrupting activity in V1 at the time feedback from area V5 arrives: If these feedback connections are at all relevant for subjective awareness, interfering with processing in V1 should also preclude the perception of attributes encoded by area V5. This is exactly what Pascual-Leone and Walsh (2001) observed. When pairing a first TMS impulse over V5 with a second one over V1 ~5 to 45 ms later, participants perceived either no phosphenes at all or only stationary ones, suggesting that feedback from V5 to V1 is central for the conscious perception of motion. Even more recently, Boehler and colleagues (2008) confirmed that awareness of a masked stimulus indeed correlates with fast modulations of recurrent activity in V1 and, importantly, ruled out that these recurrent modulations reflected top-down attentional signals. Recurrent processes therefore do appear to play a role in conscious perception.

Perhaps the most well-known out of all neurobiological theories of consciousness is the neural instantiation of Baar's (1988, 1994) *global workspace* model, developed by Stanislas Dehaene, Lionel Naccache, and Jean-Pierre Changeux (Dehaene and Changeux, 2011; Dehaene and Naccache, 2001; Dehaene et al., 1998a, 2006, 2014). Here, just like in Baar's original theory (1988, 1994), conscious access is thought to arise from the flexible and global broadcasting of information, previously confined to an independent, peripheral module and then made available for further processing throughout the entire cortex. This *global neuronal workspace* is envisioned to be formed by a network of interconnected high-level cortical regions, comprising the dorsolateral prefrontal cortex as well as inferior parietal, mid-temporal, and cingulate cortices. But what might be so special about these areas? According to Dehaene and colleagues, a cytoarchitectonic feature might turn out to be the distinguishing factor. Excitatory, cortical pyramidal neurons with long-range axons happen to be particularly dense in all of these areas, thereby rendering them particularly suitable to first amplify and maintain a given neural representation, and then to transmit it to as many other processors as needed (Figure 1.13). As such, this model predicts that conscious perception should be accompanied by a large, sudden amplification of brain activity (i.e., *non-linear ignition*) in an extended network of fronto-parietal brain areas.

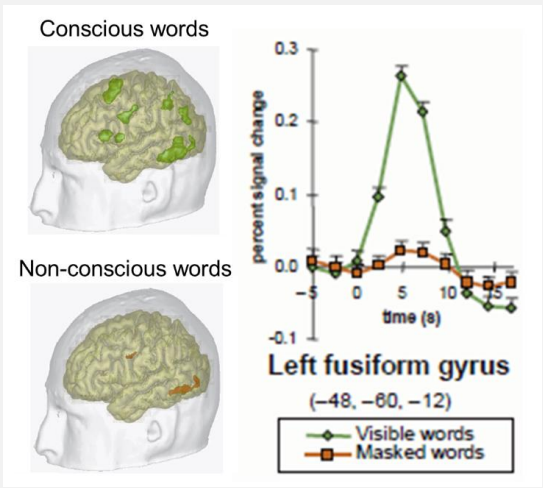


FIGURE 1.14

EMPIRICAL SUPPORT FOR THE GLOBAL NEURONAL WORKSPACE MODEL.

While non-consciouly perceived words still induce activity in visual brain areas typically implicated in reading, only conscious words recruit a highly distributed network of fronto-parietal brain regions. Adapted from Dehaene et al. (2001).

Empirical support for this theory is fairly substantial. On one hand, only consciously perceived stimuli tend to recruit the fronto-parietal network in a sustained manner (e.g., Beck et al., 2001; Grill-Spector et

al., 2000; Sadaghiani et al., 2009). For instance, while heavily masked, unseen words still elicited activity in brain areas typically implicated in reading, only activity following unmasked, seen words was found to spread to prefrontal and parietal cortices (Dehaene et al., 2001; Figure 1.14). Similarly, when monkeys were trained to report the presence (or absence) of a weak, visual stimulus, reported stimuli were associated with strongly sustained activity in prefrontal cortex. Frontal activity elicited by unreported stimuli, by contrast, failed to result in this type of prolonged activation, being much weaker to begin with and decaying rapidly (van Vugt et al., 2018). On the other hand, in direct agreement with the notion of a non-linear ignition accompanying conscious perception, late and sustained brain responses are generally reserved for conscious trials (Del Cul et al., 2007; Salti et al., 2015). Intracranial recordings during a masking paradigm, for example, revealed that, while still entraining early (i.e., < 300 ms) event-related potentials and feedforward gamma band activity, subliminal stimuli, in contrast to their conscious counterpart, no longer evoked late (i.e., > 300 ms) and long-lasting frontal potentials (Gaillard et al., 2009). A late ignition of brain activity may thus serve as a marker of conscious access.

1.3.3 The depth of non-conscious processing

So far, we have primarily talked about conscious processing: We have learned how to manipulate and measure consciousness, and have discussed some of the proposed cognitive and neurobiological bases of conscious perception. However, there is also the flipside of the coin, non-conscious processing, that we have not yet touched upon directly. You are probably familiar with the legend about subliminally presented advertisements by *Coca Cola*. In 1957, social psychologist and market researcher James McDonald Vicary claimed that he had systematically exposed ~45,000 cineastes to subliminal images of merchandise. Supposedly, for a number of movies, he had inserted a single frame, depicting either the phrase “Hungry? Eat Popcorn” or “Drink Coca-Cola.” Because these images had been flashed on the screen below the threshold for conscious perception, none of the moviegoers had ever seen these messages. Yet, to everyone’s astonishment, Vicary reported a stark increase in concession sales: According to him, sales went up by 57.5% for popcorn and by 18.1% for Coca Cola. Unfortunately, it turned out later that none of these events had ever happened, and that Vicary had invented the entire experiment. The myth of subliminal advertising nevertheless remains alive today and continues to spark ethical, legal, and scientific discussions.

Let us imagine for a second that Vicary really had conducted this study. Would he actually have observed any effects on people’s behavior? While perhaps not quite as drastic as he had made it out to be, we now know that there indeed goes on much more behind the veil of consciousness than we are typically aware of. For example, meet patient G.Y. When G.Y. was 8 years old, he incurred a head trauma due to a traffic accident, leading to the (almost) complete ablation of his left primary visual cortex (i.e., V1). As a result, he was henceforth completely blind in that part of the visual field corresponding to the site of his lesion. However, despite this complete absence of awareness for a fairly large portion of space, G.Y. retained remarkable abilities to detect, localize, and discriminate between objects shown only in this blind field. He could, for instance, accurately trace the path of a moving target along a variety of straight and curved trajectories (Weiskrantz, 1996) or identify the color of a stimulus (Brent et al., 1994), all the while denying any conscious visual experience of the target. Crucially, this paradoxical phenomenon of *blindsight* is not only restricted to human subjects, but may also be observed in other primate species (Cowey and Stoerig, 1995; Humphrey, 1974; Mohler and Wurtz, 1977; Moore et al., 1995).

At first sight, these findings sound counter-intuitive. How is it possible that one can perform above chance on an objective task, when one does not experience any of the accompanying subjective sensations? Let us briefly revisit the anatomy of the visual system: Lightwaves hit the retinae, this information is transmitted to the thalamus and then ascends to the cerebral cortex. Usually, the majority of visual signals is first sent to primary visual cortex and then up the visual hierarchy for further processing. In blindsight patients, this is obviously no longer possible. Any visual information (regardless of complexity)

will therefore have to bypass the lesioned area. There indeed exist direct projections from the lateral geniculate nucleus of the thalamus to extrastriate cortex that might serve as a potential neural substrate of the residual capabilities of blindsight patients, but even today the jury is still out on this particular issue (Leopold, 2012). What is the case, however, is that blindsight in and of itself is a striking demonstration of non-conscious processing. Though no longer being able to entrain subjective, conscious experience, visual brain signals must nevertheless have been strong enough to bias the overt behavior of blindsight patients. Non-conscious processing may therefore indeed affect our everyday actions.

But what is the extent of such non-conscious cognition? Is it limited to such fairly simple, almost instinctive behavioral responses or does it also extend to other domains of our lives? Decades of research in experimental psychology and cognitive neuroscience provide clear evidence for the latter hypothesis: A remarkably large amount of processing may occur in the complete absence of awareness (Dehaene and Naccache, 2001). Some of the earliest support for such far-reaching consequences of non-conscious processing came from masked *priming experiments*. Here, a masked stimulus (i.e., the prime) is shown shortly before a visible target. If the delay between the prime and the target is short enough (i.e., < ~100 ms), the masked stimulus may still influence the processing of the subsequent target stimulus. In their seminal work, Greenwald and colleagues (1996), for instance, demonstrated that categorization of a target word as either pleasant or unpleasant was facilitated (as measured by reaction time and error rates) when the preceding prime was of the same semantic category (as opposed to the opposite semantic category) as the target. Slightly later work by a different group of researchers replicated these semantic priming effects with numerical stimuli and, crucially, also showed that non-conscious primes had a measurable effect on brain dynamics and activity (Dehaene et al., 1998b). On incongruent trials, covert motor activation was initially observed on the incorrect side, suggesting that participants may have non-consciously applied the task instructions indexed by the prime. A host of other studies also confirmed the existence of non-conscious semantic processing (e.g., Gaillard et al., 2006; Nakamura et al., 2018; Weibel et al., 2013; Yeh et al., 2012). In their totality, these findings thus suggest that even fairly high-level, semantic representations may be activated and accessed non-consciously.

Since then, there have been many reports stretching the capacities of non-conscious cognition even further, and, crucially, also demonstrating that virtually any brain area may be recruited by non-conscious processes. Non-conscious motivational and learning processes may, for instance, activate typical reward-related brain areas (Pessiglione et al., 2007, 2008), and even prefrontal networks involved in control and inhibitory functions may be engaged non-consciously (e.g., van Gaal et al., 2010; Lau and Passingham, 2007; Figure 1.15). Similarly, cognitive control (Reuss et al., 2011) and error detection (Charles et al., 2013, 2017) have been observed to occur outside the realms of conscious awareness. Very recently, Sklar and colleagues (2012) even claimed that multistep, effortful arithmetic equations could be solved non-consciously, by showing that reaction times to targets congruent with the equation's solution were reduced as compared to incongruent ones. Though these findings were recently replicated by another group (Karpinski et al., 2018), they have also engendered considerable criticism both on statistical and theoretical accounts (Moors and Hesselmann, 2018; Shanks, 2017) and are, as such, in my opinion, at best inconclusive. What is certainly clear is, however, that non-conscious processing does not stop at simple visual awareness, but may include a wide variety of higher-level cognitive functions and brain areas.

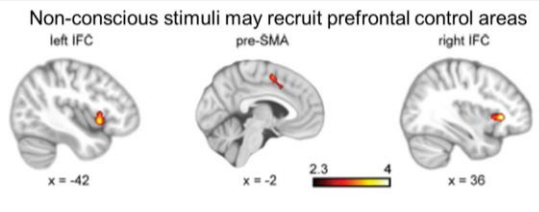


FIGURE 1.15

NON-CONSCIOUS STIMULI MAY RECRUIT EVEN HIGHER-LEVEL BRAIN AREAS IN PREFRONTAL CORTEX.

Participants completed a masked version of a go/no-go paradigm, in which a rare stimulus cues the inhibition of a pre-potent motor response. Although subjects did not consciously perceive the instructive cue, it still led to a slowing-down of their motor response and brief activation of prefrontal control areas. Adapted from van Gaal et al. (2010).

1.4 Putting it all together: Out of sight, out of mind?

Up until now, we have largely treated working memory and conscious perception as separate phenomena. However, while reading through the previous sections, you may already have noticed certain commonalities between these two functions: They appear to be grounded in the same cognitive theories, and share similar characteristics and brain mechanisms. My goal for this part of the thesis is to make this link even more apparent and explicit, critically review the theoretical and empirical literature directly relevant for the question at hand, and set the stage for the presentation of my own experimental contributions. We will begin this last section by revisiting some of the cognitive models we talked about before.

1.4.1 Traditional theoretical perspectives support intimate relationship between conscious perception and working memory

Definitions of working memory emphasize the short-term storage and manipulation of information for prospective use and goal-directed behavior. Intuitively, it seems quite obvious that such a system should be closely tied to your current conscious experience. At any given moment of your waking hours, you are unavoidably taxing your working memory: When watching a movie or reading a book, you have to keep in mind the past characters and events in order to be able to follow the plot. Similarly, when taking the subway or a train, you have to have your ultimate destination handy to not miss a stop or take the wrong connection. In a sense, your working memory seems to serve as the *sketchpad of your consciousness*. Whatever sensation, impression, thought you are currently holding onto, they also automatically appear to be part and parcel of your current subjective experience. Perhaps there really is some truth to the phrase "out of sight, out of mind" (or should I rather say "out of mind, out of sight")?

Though not always made explicit, there indeed has been a long-standing tradition in psychology to equate working memory with conscious experience. We have already seen a snippet of how Charles Richet (1884) and William James (1890) construed the relationship between consciousness and memory in general. For William James, in particular, the contents of working (i.e., primary) memory and of conscious experience were inextricably linked and, in fact, indistinguishable from each other. Talking about the act of retrieval from long-term (i.e., secondary) memory, he states:

> *"But an object of primary memory is not thus brought back; it never was lost; its date was never cut off in consciousness from that of the immediately present moment. In fact it comes to us as belonging to the rearward portion of the present space of time, and not to the genuine past."*

Since then, not much has changed. Most of the influential cognitive models of working memory we have discussed before implicitly or explicitly assumed a tight link between consciousness and working memory. Take Broadbent's (1957, 1958) filter model of attention (Figure 1.16A). Here, selective attention gates access to a central capacity-limited "perceptual system," that, in and of itself, appears to be a conscious processor. As we have already discussed before, it not only sits atop the proposed processing hierarchy, thereby receiving only a very limited amount of the most highly relevant information, but also coordinates the actions of the remaining buffers and stores, deciding whether to keep a representation alive for further processing (i.e., sending it back to the short-term store) or to transmit it to downstream systems, such as long-term memory or motor output. No covert or overt action appears possible without the intervention of this perceptual system, making it more than a likely candidate for a genuinely conscious module. By contrast, and by virtue of the sheer amount of information that may be stored, both the short-term as well as the long-term store may certainly operate outside the realms of conscious perception.

The intimate link between consciousness and working memory becomes even more apparent in subsequent models. Atkinson and Shiffrin (1968, 1971) explicitly equated the contents and operations of their capacity-limited short-term store with conscious awareness, stating that "the thoughts and

information of which we are currently aware can be considered to be part of the current contents of STS [working memory]" (Figure 1.16B). Similarly, several components of Baddeley and Hitch's (Baddeley, 2000, 1992b, 2003; Baddeley and Hitch, 1974) multicomponent model of working memory integrate features typically associated with conscious processing (Figure 1.16C). In this regard, the episodic buffer certainly constitutes the clearest cut case. It was specifically conceived to integrate information from long-term memory as well as the independent, peripheral slave systems into a unified, multi-modal representation, and, importantly, to be directly accessible to conscious awareness. In a sense, it is meant to serve as the interface between memory and consciousness. The role of some of the other components, in contrast, may not be as readily evident. Both the visuo-spatial sketchpad and the phonological loop contain elements that require access to consciousness, while others do not. For instance, the phonological loop has been closely associated with verbal rehearsal and the visuo-spatial sketchpad with visual imagery (Baddeley, 2003), both operations for which at the very least the outputs need to be conscious. However, unattended speech, even if presented in a foreign language, still has been shown to interfere with the retention of visually presented items (e.g., Salamé and Baddeley, 1986), suggesting that information entering the peripheral stores may do so automatically and be held there for several seconds. A similar set of considerations also appears to apply to the central executive. Though closely related to Norman and Shallice's (1986) supervisory attentional system, and, as such, responsible for the initiation of voluntary, conscious cognitive control processes (e.g., item rehearsal, attentional shifting, inhibition, etc.), some of its actions (e.g., sifting through the contents of long-term memory) may not necessarily be accessible to

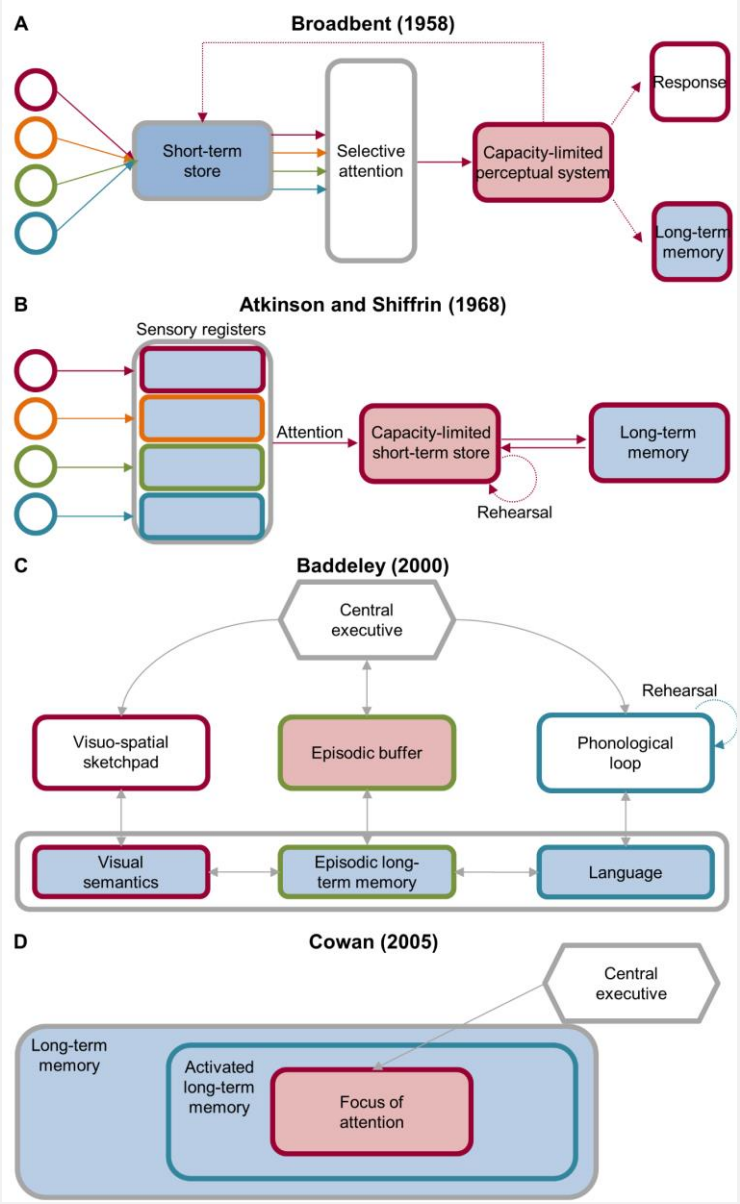


FIGURE 1.16

THE ROLE OF CONSCIOUS AWARENESS FOR INFLUENTIAL MODELS OF WORKING MEMORY.

(A) Broadbent's filter model of attention.
(B) Atkinson and Shiffrin's multi-store model of working memory.
(C) Baddeley and Hitch's multi-component model of working memory.
(D) Cowan's embedded process model. Same conventions as in Figure 1.5 apply throughout. Components thought to depend on (or reflect) conscious processing are highlighted in red, components presumably not necessarily requiring access to consciousness are highlighted in blue.

conscious awareness. All in all, then, it seems as if, in Baddeley's model, consciousness primarily operates through working memory.

While these systems-based models thus tend to consider all or, at the very least, the large majority of the contents of working memory to be synonymous with the currently experienced contents of consciousness, the state-based models favor a slightly different perspective. Here, only a subset of the information currently activated in working memory will receive sufficient attentional amplification to gain access to awareness. Take Cowan's embedded process model as an example (1997; Figure 1.16D). Both the contents of long-term memory as well as its activated portion are clearly non-conscious. Only items within the current focus of attention constitute the contents of consciousness. At any given time, the contents of working memory may thus be in two different representational states. Either they are conscious (due to sufficient amounts of attention), or they are non-conscious (though may easily be brought into awareness). Access to consciousness therefore serves as a central aspect of these models.

We have just seen that, whether implicitly or explicitly, consciousness features as an integral part in most contemporary models of working memory. But what about the converse direction? What do cognitive theories of consciousness have to say about working memory? Here, too, working memory is thought to play a prominent role. In Prinz's (2010) development of Jackendoff's (1994) intermediate-level theory of consciousness, for instance, intermediate level representations are thought to gain access to consciousness only when they become available for encoding in working memory as a result of attentional amplification. The contents of working memory are thus once more being equated with the contents of conscious awareness. Similarly, in Baar's global workspace model, the global workspace, or theater stage to stick with Baar's own metaphor, corresponds to working memory (Baars, 1988, 1997; Baars and Franklin, 2003), with only its focal contents being broadcast globally and thus made available to consciousness. In a sense, the global workspace provides a working memory space, from which attention selects the most relevant representations for further, conscious processing. Sounds a bit like the state-based models of working memory we just talked about, does it not? Just like in these models, the contents of consciousness are supposed to be determined by the current contents of working memory. All in all, then, both cognitive theories of working memory as well as of conscious access are built on and involve similar features and concepts. The terminology may be a bit different, but it is beyond doubt that consciousness and working memory are intricately interwoven in all of these models.

1.4.2 Conscious perception and working memory share common characteristics and brain mechanisms

Psychological theory largely tends to corroborate our own intuitions. Conscious perception and working memory are considered to be tightly linked, with the active contents of working memory overlapping substantially with those of your current, conscious experience, and some sort of conscious entity (i.e., homunculus) featuring in many of the earlier models of working memory. But what about actual data? Should the results of experiments lead you to conclude that subjective, conscious experience and working memory are indeed closely related with each other? Let us start by revisiting some of the properties of working memory we discussed earlier and see how they might also apply to conscious processes.

1.4.2.1 Access to working memory and consciousness guarantees longevity, stability, and robustness of representations

Perhaps the most defining feature of working memory is to *maintain information for short-periods of time*. This may sound obvious, but is an important characteristic to consider. Once information has entered working memory, even in the complete absence of rehearsal or refreshing of the corresponding representation, the memory trace will at the very least persist for several seconds (Muter, 1980). Indeed,

the underlying causes of "forgetting" are still actively debated in the community (Jonides et al., 2008). While some authors argue vigorously in favor of passage of time constituting the major factor in determining the integrity of working memory representations (i.e., *decay theory*; Barrouillet and Camos, 2012; Barrouillet et al., 2004, 2011), others hold that it is solely interference from other items that leads to the disintegration of working memory representations (i.e., *interference theory*; Lewandowsky and Oberauer, 2009; Lewandowsky et al., 2009). An in-depth discussion of this problematic is certainly far beyond the scope of my work here and is also not the point that I would like to make. Rather, I just wanted to draw your attention to the fact that working memory lends a fair amount of *durability* and *stability* to its contents. Without it, as is, for example, the case for iconic memory, representations would very quickly decay and fade away (Dosher et al., 2005).

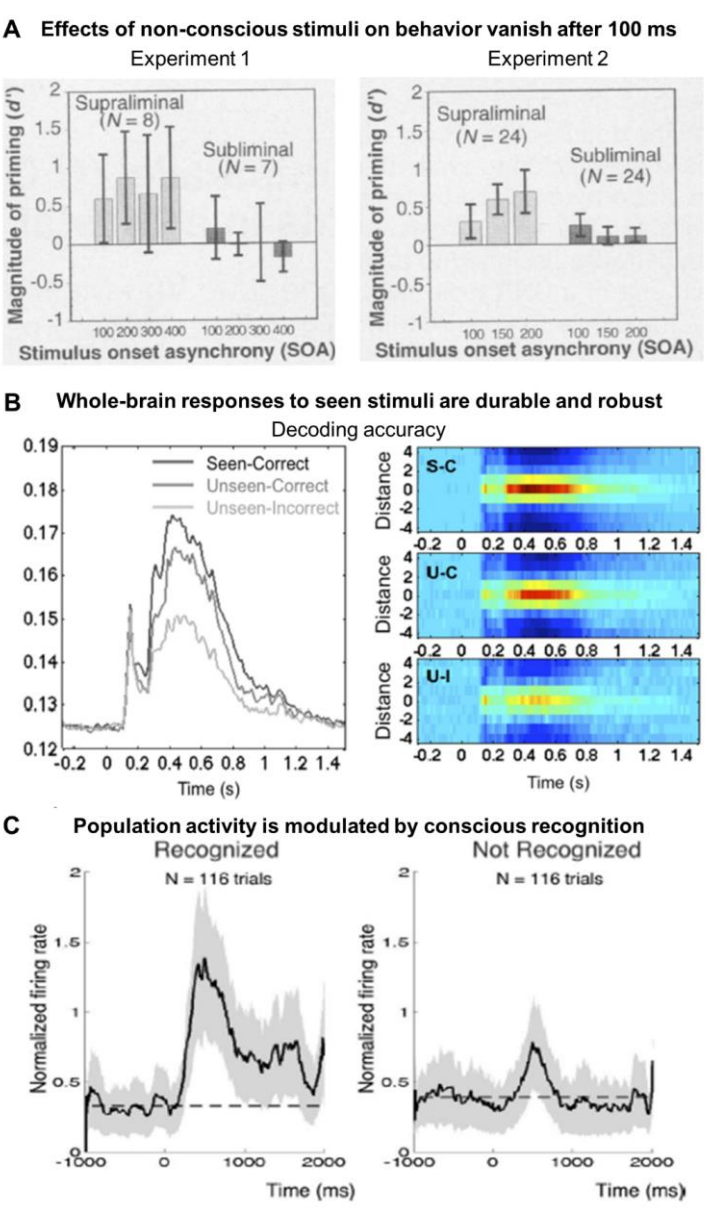


FIGURE 1.17

ACCESS TO CONSCIOUSNESS PERMITS THE MAINTENANCE OF INFORMATION.

(A) Magnitude of priming effects for conscious and non-conscious primes in two experiments as a function of prime-target SOA. While conscious (supraliminal) primes consistently affected target processing across all four delays studied (100, 200, 300, and 400 ms), the effects of non-conscious (subliminal) primes vanished after about 100 ms. Adapted from Greenwald et al. (1996).
(B) Time course of decoding performance for spatial location as a function of visibility and accuracy. Up until ~270 ms, decoding accuracy, providing a rough estimate of the amount of information present in the brain, is indistinguishable between all three conditions. Crucially, hereafter, it is stronger and more durable for conscious trials than even for blindsight trials (on which subjects responded correctly in the absence of any subjective experience). Chance performance corresponds to 12.5%. Adapted from Salti et al. (2015).
(C) Normalized activity of a population of neurons in human entorhinal cortex as a function of subjective recognition. Even for objectively identical inputs, spiking activity is clearly higher and more stable for consciously recognized trials. Adapted from Quiroga et al. (2008).

A similar beneficial effect is also associated with conscious processing. Remember those masked priming studies we talked about? Though, as we have discussed before, the general finding is that non-conscious primes do influence processing of subsequent target stimuli, these effects are typically much more short-lived than the ones observed following a conscious prime (Dupoux et al., 2008). For instance, going back to Greenwald and colleagues' (1996) seminal work on non-conscious semantic processing, these authors showed that, for their non-conscious primes to have any effect whatsoever, the delay between the prime

and the target (i.e., prime-target *stimulus onset asynchrony* [SOA]) could not exceed ~100 ms. By contrast, conscious primes were consistently strong, perhaps even increasing in strength, across all of the four SOAs studied (i.e., 100, 200, 300, and 400 ms; Figure 1.17A). Moreover, on conscious, but not on non-conscious trials, the prime-target relationship on the previous trial modulated the size of the effect on the current trial. Conscious representations thus influenced subjects' behavior in a much more stable, durable fashion than did non-conscious representations, consistent with the idea that, similar to working memory, access to consciousness permits maintenance of information.

Similar differences also exist at the brain level. We have talked quite a bit about how content-specific, persistent neural firing is still considered to be the hallmark of working memory (though this may be shifting a bit). When it comes to the distinction between conscious and non-conscious processing, late, robust, and sustained brain responses are also typically reserved for the former (Dehaene et al., 2017; Koch et al., 2016). The P3b, a late (i.e., > ~300 ms) event-related potential (ERP) observed over central sensors, for instance, appears to index conscious, as opposed to non-conscious perception (Dehaene and Changeux, 2011; Del Cul et al., 2007; Lamy et al., 2009; Naccache et al., 2016; Polich, 2007), while most of the earlier components are present for both seen and unseen stimuli. Even more strikingly, once a target stimulus has crossed the threshold for subjective, conscious perception, content-specific neural activity is also more robust and durable than when the very same stimulus has not been detected – even if the participant's response is strictly identical. Take a recent study by Salti and colleagues (2015) as an example. Here, subjects were first asked to report 1 out of 8 possible masked spatial locations, and then to rate their visibility of the target on a scale from 1 to 4. Behaviorally, even when participants had not perceived the actual stimulus, they were still able to identify the correct target location much better than predicted by chance alone. This is the blindsight effect we talked about previously. For the intents and purposes of our current discussion, this also means that there were two groups of correct trials: a subset, on which the subjects had seen the target, and a subset, on which they had not seen the target. What did the authors observe in terms of brain responses for these two conditions? Crucially, they found that, although initially encoded identically, after ~270 ms, brain activity started to diverge. Information associated with seen trials was selectively amplified and persisted for longer than its non-conscious counterpart (Figure 1.17B).

Direct neural recordings from the monkey or human brain corroborate these findings. Quiroga and colleagues (2008), for instance, compared single neuron responses from the human medial temporal lobe as a function of whether their subjects had consciously recognized a target photograph or not (Figure 1.17C). Here, too, neurons fired much more vigorously and longer when the items had crossed the threshold for conscious recognition than when they had not. Similarly, as we have already discussed before, van Vugt and colleagues (2018) demonstrated just very recently that, at all stages of the visual hierarchy in the monkey brain (up to and including prefrontal cortex), weak, unreported (i.e., unseen) stimuli elicited consistently weaker and more short-lived responses than their conscious counterparts. They then compared their data to predictions of a computational model of brain activity, in which a non-linear ignition process in higher cortical areas led to conscious perception. The model nicely fit their empirical data, suggesting, once more, that access to conscious awareness is an all-or-none phenomenon, which results in the maintenance and global broadcasting of selective information.

Let me finish this section on just two more related thoughts. Though this has been my main focus so far, conscious perception may not only amplify and prolong neural activity, it may also render it more *reproducible* and *stable*. The main evidence here comes from two recent magnetoencephalography (MEG) and one fMRI study (Baria et al., 2017; Schurger et al., 2010, 2015). On one hand, all three agree that brain activity tends to be less variable (i.e., more reproducible) across trials. On the other hand, their conclusions regarding the stability of brain activity associated with conscious processing are inconclusive. While Schurger and colleagues (2015) presented evidence in favor of more stable cortical activity during a conscious event, Baria and collaborators (2017) argued for robust, albeit transient dynamics. How can both of these accounts be true? One possible explanation, put forward by one of the authors of the second

study (He, 2018), may be related to differential dynamics as a function of frequency, with lower frequencies exhibiting fast-evolving activity trajectories for seen, but not unseen stimuli, and higher frequencies displaying the opposite pattern. It may, however, also signal that the maintenance of a stable representation need not necessarily be accompanied by stable neural patterns. Remember our discussion about the neural substrates of working memory? In this context, we have already seen that neural activity tends to be much more dynamic than previously thought, with neurons changing their coding preferences over the course of a single trial (Parthasarathy et al., 2017; Stokes et al., 2013). Quick successions of meta-stable patterns of brain activity also turn out to be a fairly common feature in many recent decoding studies on conscious perception and working memory (King and Dehaene, 2014; Marti and Dehaene, 2017; Spaak et al., 2017; Wolff et al., 2017). Having been taken to reflect sequential processing, such transient dynamics are, in fact, compatible with theoretical models of conscious access, such as the global neuronal workspace, which presume a cascade of processes once a piece of information has become conscious. Perhaps even more importantly, however, the information content of representations may persist, even when underlying neural patterns are changing dynamically (Myers et al., 2015). As such, the contents of your consciousness (or working memory, for that matter) can remain stable in the absence of stable brain activity.

In sum, then, the essence of all of these findings reviewed above boils down to this: Both conscious perception and working memory appear to permit the stable, short-term maintenance of information. Unseen stimuli do affect behavioral and brain responses, but, in stark contrast to their conscious counterpart, these effects are weak and variable, and vanish rapidly. At least in terms of the main characteristic of working memory, there is thus a large overlap with consciousness.

1.4.2.2 Both working memory and consciousness are capacity-limited, central systems

Both conscious perception and working memory are oftentimes described in terms of severely capacity-limited, central systems. In the case of working memory, we have already considered this particular property when delineating it from both iconic and long-term memory. Whereas we can presumably store an unlimited amount of information in our long-term memory, and, apparently also possess a high-capacity, but quickly decaying iconic memory (Neisser, 1967; Sperling, 1960), time after time, it has been shown that, depending on the person and situation, we can only store about 4 to 7 high-fidelity representations in working memory (e.g., Bays and Husain, 2008; Cowan, 2001; Luck and Vogel, 1997; Vogel and Machizawa, 2004). Many, though certainly not all, researchers in consciousness science also argue in favor of such a bottleneck when it comes to conscious perception (Baars, 1988; Cohen and Dennett, 2011; Dehaene and Changeux, 2011; Dehaene and Naccache, 2001; Dehaene et al., 2017; Kouider et al., 2010; Lau and Rosenthal, 2011; Marois and Ivanoff, 2005, but see Block, 2011; Cohen et al., 2016; Lamme, 2010 for a different perspective). Take the global neuronal workspace as an example: As visual information traverses the processing hierarchy, it first passes through an early, non-conscious stage with unlimited capacity, before being admitted (via attentional selection) to the capacity-limited, serial workspace.

Empirical support for this position comes in a variety of different flavors. For instance, do you recall our discussion on bistable perception and binocular rivalry? Here, your brain either receives an ambiguous input or two different images via your two eyes. Yet, although both representations are clearly coded simultaneously and continue to be processed non-consciously (Leopold and Logothetis, 1996; Logothetis et al., 1996; Panagiotaropoulos et al., 2012), they do not populate your mind at the same time. At any given moment, you are only aware of a single percept.

Change or inattentional blindness constitutes another excellent demonstration of the limited nature of our conscious perception. In their seminal work, Rensink and collaborators (1997) employed the now well-known “flicker paradigm,” in which the presentation of photographs was interspersed with the

display of blank screens (thereby creating the impression of a flickering image). Crucially, at some point during the sequence, the original picture was replaced with a modified version, in which objects had either been (1) removed entirely, (2) colored differently, or (3) displaced spatially. As you have already experienced yourself when you watched the inattentional blindness experiment on youtube, with this particular setup, subjects were remarkably bad at identifying even blatant changes to objects in the scene, sometimes requiring up to 50 s before noticing the discrepancy. Despite our propensity to consider our own visual experiences as fairly rich and vivid, they might thus, in fact, be much more limited and sparse.

A last example that I would like to mention in this context is the attentional blink. Consider a recent experiment by Marti and colleagues (2015; Figure 1.18). Here, subjects had to perform two tasks simultaneously: They first had to discriminate the pitch of a sound (target 1) and then the identity of a letter (target 2), embedded in a random series of letters. The delay between target 1 and target 2 varied across trials between 100 (lag 1) and 900 (lag 9) ms. Behaviorally, these authors reported typical effects of the psychological refractory period and the attentional blink: The closer the two targets appeared in time, the slower participants responded to the second target (i.e., psychological refractory period) and the fewer instances of target 2 did they actually detect (i.e., attentional blink). Intriguingly, at these shortest lags, the brain initially appeared to non-consciously process information for the two targets simultaneously and in parallel, but then, after ~500 ms, switched to a serial mode, shortening processing of the first target, while also delaying dealing with the second. Marti and collaborators took these differential dynamics to suggest that, during the execution of the first task, conscious perception of the second target is temporarily put on hold, only able to succeed if processing of the first task does not take longer than the decaying representation of the second target. As such, this study highlights the differences in capacity limits between conscious and non-conscious processes. While the latter may perform multiple computations in parallel, the former appear to be able

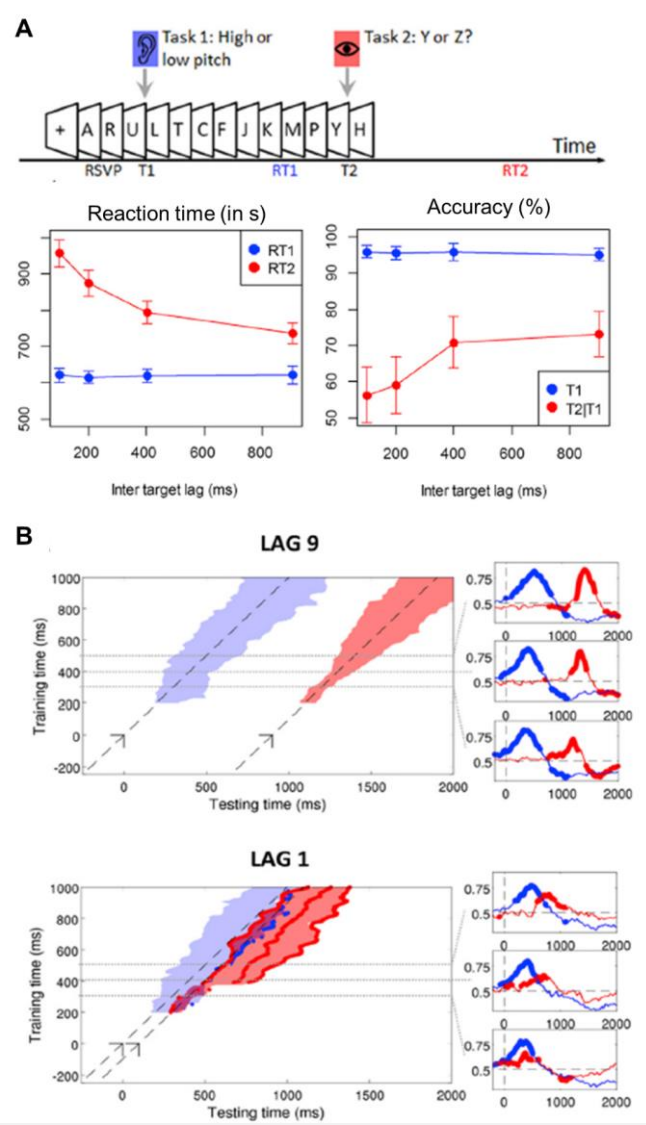


FIGURE 1.18

CONSCIOUS PERCEPTION IS SERIAL AND CAPACITY-LIMITED.

(A) (Top) Participants performed a dual-task, in which they first had to identify the pitch of a sound, and then discriminate between two target letters. Crucially, the temporal delay between the two targets was varied across trials. (Bottom) Subjects display typical effects of the psychological refractory period and attentional blink: For shorter lags, they respond more slowly and less accurately, suggesting that they did not always detect the second target.
(B) Although both targets could initially be processed in parallel, after ~500 ms, processing of the second target was put on hold, implying that, if it did cross the threshold for conscious processing, it did not do so before processing of the first target had been terminated. Adapted from Marti et al. (2015).

to only handle a handful of information at once, and, as such, may have to revert to serial operations. In this respect, then too, conscious perception closely resembles working memory.

1.4.2.3 Contents of working memory and consciousness may be manipulated and reported

If you have been following me until now, you may have noticed that I have left out a central aspect of the definition of working memory in our current consideration. Do you still recall how I introduced you to this particular type of memory? That I emphasized its *working* component? In addition to the simple storage of information, most of the influential theories of working memory also highlight its role in goal-directed behavior and preparing representations for prospective use (Atkinson and Shiffrin, 1968, 1971; Baddeley, 1992a; Baddeley and Hitch, 1974; Cowan, 1997; Luck and Vogel, 2013; Miller, 1956; Oberauer, 2005; Wager and Smith, 2003). We have already discussed this many times, but I think it is important to mention it again: Working memory allows you to connect your present to your future, briefly storing information, but, crucially, also transforming it in a meaningful way, so that you can apply it dynamically and flexibly. Given this definition, it is no surprise that working memory is indispensable for any kind of complex operation. Whether you forgot your shopping list (as my fiancé just did) and then need to call home and ultimately remember the items on the list by heart, or are trying to solve an arithmetic equation in your head (e.g., 3267 + 845 + 67 + 23), you are automatically relying on your working memory to do the trick.

The question we now have to ask, in the context of our current discussion, is how conscious and non-conscious cognition map onto this. We have already seen that non-conscious signals tend to have widespread consequences on behavior and elicit activity in many different brain areas (Dehaene and Naccache, 2001). Perceptual, motor, semantic and higher-level control processes may all proceed in the complete absence of subjective awareness (Boy et al., 2010; Charles et al., 2013, 2017; Greenwald et al., 1996; Merikle and Reingold, 1990; Nakamura et al., 2018) and may sometimes even trigger neural responses in the prefrontal cortex (van Gaal et al., 2010; van Vugt et al., 2018). But are there any limits to the depth of non-conscious processing, or may any operation, no matter how complex, also occur non-consciously?

Conscious perception in particular and consciousness more generally appear to be necessary for *abstract*, *symbolic*, *sequential*, *and rule-following computations* (Dehaene and Naccache, 2001). For instance, Sackur and Dehaene (2009) demonstrated that piping, or chaining, mental operations appears to require prior access to consciousness. Participants first had to add (or subtract) the number 2 to (from) a target digit, and then compare the result with 5. In the critical condition of this experiment, the target digit was rendered subliminal by means of backward masking. While subjects were able to perform each of the elementary operations in the absence of subjective awareness (that is, they could add or subtract 2 from the target, or they could directly compare the target digit with 5), they did not exceed chance when being required to execute both operations in sequence. Another striking example of the boundaries of non-conscious cognition comes from the inclusion/exclusion procedure (Debner and Jacoby, 1994; Merikle et al., 1995). Here, a masked word is quickly flashed on a computer screen and immediately followed by a presentation of its stem. The target word *spice*, for instance, might be followed by the stem *spi__*. Participants are required to complete this stem with any word other than the original target stimulus. That is, they are being asked, to exclude a certain representation from their repertoire of responses. While, under conscious conditions, people are generally fairly good at complying with these task instructions, they have much more difficulty when they did not perceive the target, oftentimes completing the stem with just that target word. Conscious awareness thus also seems to be essential in inhibiting a prepotent automatic response and, at the same time, deploying a novel, unusual strategy. Note that, in contrast to these findings, there has also been a recent report claiming to have found evidence for complex, multistep operations outside the realms of conscious awareness (Sklar et al., 2012). However, we have already seen how diverse replication attempts have led to, at best, inconclusive evidence for such non-conscious

processing capabilities (Moors and Hesselmann, 2018; Shanks, 2017). As such, I am still inclined to align myself with the bulk of the data and favor a more conservative stance. Though non-conscious processing is certainly deep and widespread, it also appears to be insufficient for exactly those kinds of computations and mental operations working memory is most known for. There thus appears to be yet another link between conscious perception and working memory.

Allow me to finish this section by drawing your attention to one last related, yet often overlooked, consequence of access to consciousness or working memory: *reportability*. Almost by definition, as soon as a piece of information enters working memory or crosses the threshold for conscious perception, it becomes available for verbal as well as behavioral report. If I were to ask you right now to describe your current subjective experience or the contents of your working memory, you would easily be able to do so. We have already seen that, even empirically, report is often considered to be the standard operational index of both working memory and consciousness. Subjects may, for instance, be asked to report back a list of items (e.g., Murdock, 1962), or to rate their visibility of a target stimulus on an ordinal scale (Salti et al., 2015; Sergent and Dehaene, 2004). As banal as this fact may seem at first sight, it actually might also tell us something about the nature of the underlying representations. Both the contents of consciousness and working memory appear to exist in a readily accessible, easy-to-use format that is amenable to covert or overt report. This is not the case for non-conscious representations and, as such, once more highlights a similarity between working memory and consciousness.

1.4.2.4 Similar brain mechanisms appear to subtend working memory and conscious perception

Before moving on, let us quickly summarize what we have covered so far. Both from a theoretical and an empirical point of view, conscious perception and working memory appear to be intimately related, sharing a number of defining characteristics and properties. What we have only touched upon in passing is their relationship at the neural level: Do they recruit similar brain areas and rely on comparable mechanisms, or will this finally constitute the point of clear divergence between these two cognitive functions?

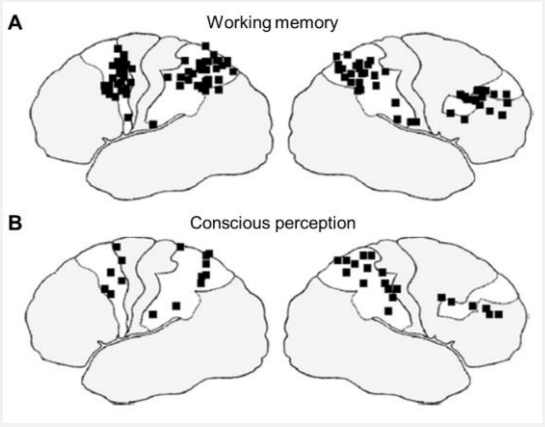


FIGURE 1.19

CONSCIOUS PERCEPTION AND WORKING MEMORY RECRUIT SIMILAR BRAIN AREAS.

(A) Peaks of fMRI activations from a selected number of different studies examining working memory.
(B) Same as in **(A)** but for conscious perception. Adapted from Naghavi and Nyberg (2005).

When we talked about the neural substrates of working memory, I attempted to emphasize their distributed nature (Figure 1.7). Posterior sensory regions appear to be primarily responsible for the storage of high-fidelity representations, while more anterior brain areas and, in particular, the dorsolateral prefrontal cortex, seem to figure most prominently as a "top-down" control and management system. A very similar network of brain regions also seems to play an important role for subjective, conscious experiences (Rees, 2007). For instance, as soon as a visual stimulus crosses the threshold for conscious perception, activity in higher-level, but also early, *visual cortex* is greatly amplified as compared to when the very same stimulus fails to reach awareness (Haynes et al., 2005; Polonsky et al., 2000; Ress and Heeger, 2003; Williams et al., 2008). In addition, as we have already seen before, the recruitment of a *distributed network of bilateral and prefrontal cortices* is also frequently implicated in conscious access. Neuroimaging experiments have consistently revealed a strong correlation between activity in these fronto-parietal areas and subjective visibility, with these brain areas only being activated for consciously perceived stimuli (Dehaene et al., 1998b, 2001; Kleinschmidt et al., 1998; Lau and Passingham, 2006; Lumer, 1998). Crucially,

evidence from lesion and brain stimulation studies supports a causal role of such fronto-parietal involvement. Damage to these areas may lead to visuospatial neglect, a clinical condition, in which patients seize to perceive (and respond to) any type of stimulation in the part of their visual field contralateral to their lesion (Bartolomeo et al., 2007; Corbetta et al., 2005). Similarly, TMS stimulation of the dorsolateral prefrontal cortex has been shown to induce a reduction of subjective visibility in the absence of any detrimental effects on objective sensorimotor performance (Rounis et al., 2010; Ruby et al., 2017), although these findings have recently been challenged (Bor et al., 2017). Most importantly for the intents and purposes of our current discussion, there is a clear overlap in these fronto-parietal regions recruited for working memory and conscious perception (Naghavi and Nyberg, 2005; Figure 1.19). While this observation, by itself, does not necessarily imply that this network performs these two cognitive functions, it very strongly suggests that, at the very least, similar types of computations may be involved in conscious perception and working memory. It is, for instance, conceivable that, posterior sensory cortices play a central role in representing the actual contents of working memory and conscious perception, while these anterior fronto-parietal areas select, integrate, and transform this information.

In addition to this similarity in brain networks recruited, other neural markers also seem to be shared between consciousness and working memory. *Elevated, sustained neural activity* is an obvious contender. We have already discussed at length how persistent neural firing is still considered to be the prime candidate for the neural correlate of the working memory engram (e.g., Courtney et al., 1997, 1998b; Funahashi et al., 1989; Fuster and Alexander, 1971; Haller et al., 2018; Kornblith et al., 2017), and how late and sustained brain responses are also typically associated with conscious processing (Dehaene et al., 2017; Koch et al., 2016). Indeed, irrespective of the theoretical stance, maintenance of information features as a key ingredient in most neurobiological models of conscious perception, be it in the form of synchronous thalamocortical activity (Tononi and Koch, 2008), cortical recurrence (Lamme and Roelfsema, 2000), or the sustained recruitment of a fronto-parietal network in a global neuronal workspace (Dehaene and Changeux, 2011).

What might, perhaps, be a little less apparent is the role of *long-range connectivity* and *neural synchrony*. Given that a widely distributed network of brain areas appears to underlie both conscious perception and working memory, these regions have to be able to communicate with each other in order to exchange information. Modulations of long-distance functional connectivity have indeed been reported for working memory. Functional connectivity between prefrontal cortex and posterior sensory brain areas may, for instance, increase during the delay period of a working memory task (Gazzaley et al., 2004; Kuo et al., 2018). Crucially, such coupling appears to be directly relevant for behavior. Galeano Weber and colleagues (2017) measured functional connectivity between occipital and parietal brain areas during the encoding of information in working memory and observed increased precision in behavioral performance measures when occipito-parietal connectivity was higher. Though the neural basis of such increases are still fairly poorly understood, it might reflect synchronization of neural activity (Liebe et al., 2012). Similar observations have also been made in the domain of visual awareness. Long-distance synchrony in beta as well as gamma frequency bands is consistently increased during conscious perception (Gaillard et al., 2009; Gross et al., 2004). Consider a seminal study by Buschman and Miller (2007). Monkeys were trained to search for (and detect) a visual target under two different conditions: Either the physical features of the target were sufficiently different from those of the distractors to make it pop-out quickly (bottom-up attention condition), or it could only be identified by means of an effortful search guided by a template representation held in working memory (top-down attention condition). Once the monkeys attended to (and thus became consciously aware of) the target stimulus, fronto-parietal synchrony was enhanced, either in lower beta-band frequencies during top-down, or in higher gamma-band frequencies during bottom-up attention. Long-distance connectivity and neural synchrony thus appear to be equally important for both working memory as well as conscious perception.

1.4.3 Putting prevailing views to the test: May there be non-conscious working memory as well?

So, here we have it, then. There exist myriad reasons to consider working memory and conscious perception to be intricately linked: Cognitive theories of both functions build on each other and are tightly interwoven, and both share central features and neural mechanisms, including their capacity-limited nature, their role in the maintenance of information and brain states, and their similarly distributed neural bases. It indeed appears as if the contents and operations of conscious awareness might equally well be described in terms of the contents and operations of working memory.

Yet, arguably, up until now, we have largely considered tangential, peripheral evidence. While I have pointed out glaring similarities that have emerged in the two fields of research, we have yet to talk about studies that actually look at these two phenomena directly and in a joined manner. Otherwise, how can we be sure that these commonalities are specific to the relationship between consciousness and working memory? Perhaps, had I reviewed other pairs of cognitive functions, such as working memory and attention, or attention and visual awareness, I might also have observed a comparable overlap.

Given the pervasiveness of the assumption of an intimate coupling specifically between conscious perception and working memory, one might expect there to be a plethora of empirical evidence that directly speaks to this question. However, this is not the case at all. This specific problematic actually falls into a newly minted area of research that has only just begun to attract attention. What, then, is the result of this renewed interest? Do the findings support the prevailing perspective of a close relationship? Or do they challenge it? Let us first take a look at the existing evidence.

1.4.3.1 Visual working memory may operate outside the realms of conscious awareness

A first question that we might pose when evaluating the relationship between conscious perception and working memory is whether the actual operations of working memory require access to consciousness. That is, are we always aware of engaging our working memory, or may it sometimes also store, manipulate, and transform information implicitly, without our conscious knowledge and intention? Empirical evidence for this particular issue remains sparse and still few and far between.

In one of the earliest attempts, Hassin and collaborators (2009) presented their subjects with a series of sequentially displayed disks and asked them to judge a simple, perceptual feature (i.e., whether they were black or white). Crucially, the authors manipulated the order of disk presentation, such that, on a subset of trials, it followed a predictable spatial pattern. If participants were able to extract the implicit spatial structure of the sequence, they should be able to anticipate the upcoming location, and thus decrease their time needed to perform the judgment task. While there was thus no overt working memory requirement, task performance was facilitated if subjects managed to store the locations of the disk in the order of their appearance and transform this information into an abstract, symbolic representation of the spatial pattern. This is exactly what Hassin and colleagues (2009) observed. Although participants systematically failed to demonstrate any subjective awareness of the existence of such a spatial structure (as assessed by, for instance, a post-experimental questionnaire, an announced free recall test of the very last spatial pattern, etc.) they nevertheless appeared to have used this information to guide their behavior and improve their performance. The authors thus take this reduction in reaction time as evidence to argue that working memory may operate both unintentionally and outside the realms of conscious awareness. However, note that the size of the observed effects was fairly small, at most amounting to an ~40 ms decrease in reaction time, and, that, crucially, once subjects were explicitly made aware of the potential existence of spatial patterns, any facilitatory effect of spatial structure was lost. Whether this particular task thus only activated the working memory system we have described so far or perhaps also recruited more implicit sequence learning based on subcortical structures is not unequivocally clear.

A slightly more indirect approach to assess the possibility of a non-conscious operation of working memory was recently taken on by Bona and colleagues (Bona and Silvanto, 2014; Bona et al., 2013). If the computations performed in working memory depended on conscious awareness, then, as a pre-requisite, the contents of working memory should accurately and directly be accessible to conscious awareness. In other words, we should have direct access to the actual working memory trace. Bona and collaborators (2013) suggest that this may not necessarily be the case. In this study, participants had to complete two tasks with respect to an item (i.e., orientation of visible grating) held in working memory: They first had to compare its orientation to the one of a probe stimulus (objective task), and then rate their subjective vividness of this working memory representation as it had been at the end of the delay period (subjective task). Importantly, on half the trials, a masked distractor grating could be presented throughout the delay period (Figure 1.20A). As you may already have seen in Figure 1.20B and C, the distractor turned out to differently affect subjects' conscious experience of their working memory contents and their actual objective performance. While, irrespective of subjective visibility of the distractor stimulus, delayed discrimination was impaired only for the most disruptive distractors, subjective vividness decreased for all non-conscious distractors, irrespective of their orientation. As such, the actual contents of working memory and whatever participants experienced as such could be dissociated, leading a subset of these authors to propose a *conscious copy* model of working memory introspection (Jacobs and Silvanto, 2015). According to this theoretical stance, the contents of working memory are, by default, non-conscious, requiring a new representation, with different functional properties, to be created specifically for the conscious domain. Another possibility might, however, simply be that, in the above experiment, a third variable, such as subjects' current attentional state, might have led to fluctuations in working memory performance and detection of the distractor stimulus. At the moment, there thus exists little direct evidence to permit an unequivocal evaluation of the topic under consideration.

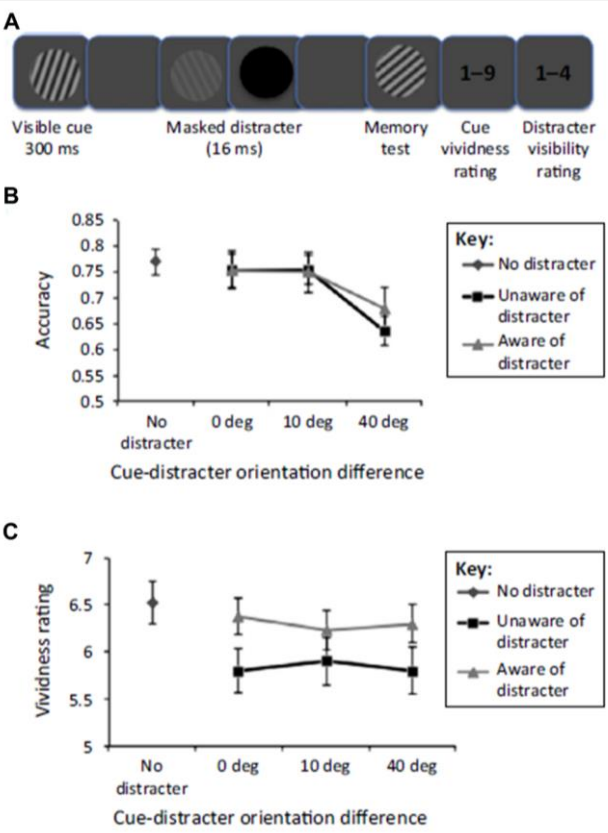


FIGURE 1.20

EMPIRICAL EVIDENCE FOR A DISSOCIATION BETWEEN CONSCIOUS PERCEPTION AND THE OPERATION OF WORKING MEMORY.

(A) Subjects first compared the orientation of a probe stimulus to the one of an item held in memory, and then rated the vividness of this working memory representation as it had been at the end of the delay. Crucially, a masked distractor stimulus could appear during the maintenance period. **(B)** The distractor impaired objective performance irrespective of visibility, but only when it had been sufficiently different from the memory stimulus. **(C)** Intriguingly, a different pattern of effects was observed for subjective judgments, suggesting that working memory contents may be dissociated from subjective experience. Adapted from Soto and Silvanto (2014).

1.4.3.2 Visual working memory may operate on non-conscious input

In addition to determining the capacity of working memory to perform its computations, operations, and transformations non-consciously, we should also consider whether the actual information working memory operates on needs to be conscious. Initial findings from a handful of experiments suggest that

this does not necessarily have to be the case. Take the seminal work by Soto and colleagues (2011) as an example. Clearly breaking with the tradition in the domain of consciousness research, these authors chose to combine a masking paradigm with a delayed-response task, requiring participants to, after a delay of up to 5 s, first compare the orientation of a masked memory cue with the orientation of a probe stimulus and then rate their subjective visibility of the memorandum on a scale from 1 to 4 (Figure 1.21A). In stark contrast to everything we have talked about so far, these authors reported that, in all four of their experiments, subjects performed the objective discrimination judgement better than would have been predicted by chance, even when they reported no awareness for the target stimulus whatsoever (Figure 1.21B). This blindsight effect was weak, but persisted in the face of a visible distractor stimulus, leading the authors to conclude that non-conscious information may be stored in non-conscious working memory (as opposed to non-conscious iconic memory). Results from a different group, relying on continuous flash suppression to render their stimuli invisible, corroborated these early findings, and supposedly, also demonstrated that a conjunction of two non-conscious features could be retained (Bergström and Eriksson, 2015). However, given that the features chosen comprised a spatial location and object identity, it seems likely that, even in this case, automatic feature-integration processes already reduced the remembered stimulus to a single, bound representation (Bapat et al., 2017).

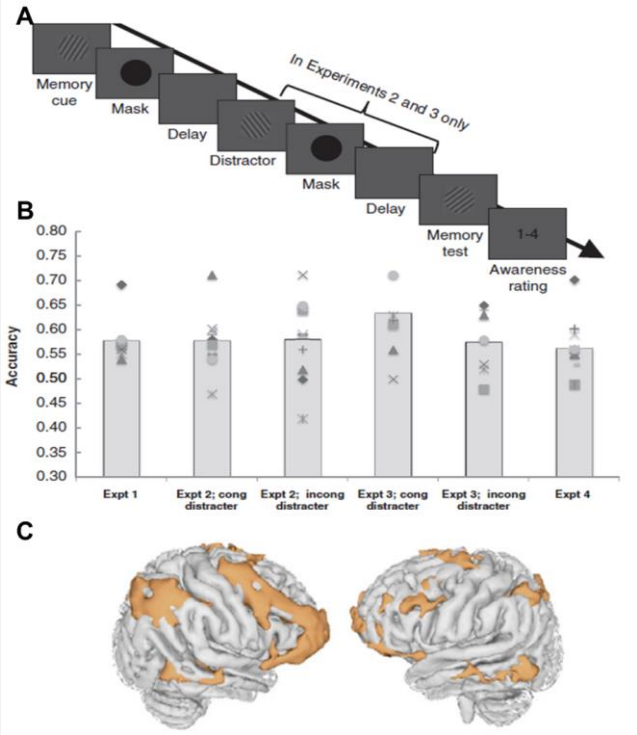


FIGURE 1.21

CURRENT EMPIRICAL EVIDENCE FOR NON-CONSCIOUS WORKING MEMORY.

(A) Delayed-masking paradigm, in which subjects had to retain the orientation of a masked Gabor patch for up to 5 s, compare it to a probe stimulus, and then rate their visibility of the target.
(B) Across a series of four experiments, participants performed consistently above chance, even when they had not seen the target stimulus. Adapted from Soto et al. (2011).
(C) On non-conscious trials, activity in a fronto-parietal network correlated with performance. Adapted from Dutta et al. (2014).

Irrespective of these particularities, both studies thus demonstrated that a non-conscious stimulus may influence behavior for much longer than previously thought: up to 15 s. But is this really enough to abandon the notion of a tight link between subjective, conscious experience and working memory, and invoke the existence of a non-conscious working memory system? Clearly, it cannot be. If you think about it, there may be many reasons, other than non-conscious working memory, that might explain the long-lasting blindsight effect. Perhaps subjects accidentally miscategorized a small subset of *seen* trials as *unseen*, or guessed the response to the objective task right after the presentation of the subliminal targets and then stored the resulting guess in conscious working memory? Similarly, even if genuinely non-conscious, perhaps the information never made it into working memory, but was held in another memory system, such as iconic, fragile, or long-term memory? Or perhaps there exists yet another type of memory?

The list of alternatives is seemingly endless, and we have barely just begun to leave a dent in the mountain of possibilities. On one hand, there is now some suggestive evidence that non-conscious stimuli may only be maintained if needed for prospective use. Pan and colleagues (2014), for example, examined the time that it would take for a subliminal face presented to one eye to break though suppression from

Mondrian noise patterns presented to the other eye. Critically, the authors reported that suppression of the subliminal face was shortened, when participants simultaneously had to maintain a matching conscious or non-conscious face in working memory as opposed to when they only had to attend to (but not remember) this initial cue. These results appear in line with our conceptualization of working memory as a memory system in the service of goal-directed behavior and also highlight that even non-conscious contents of working memory may modulate the gating of subsequent information into conscious awareness.

On the other hand, there have also been some attempts at gauging the neural underpinnings of such storage of non-conscious information, in an effort to demonstrate that this would require similar neural substrates as conscious working memory. Dutta and collaborators (2014) recorded fMRI while subjects performed the same masked spatial-delayed response task as displayed in Figure 1.21A. They observed that BOLD signal change in a fronto-parietal network correlated with working memory performance even on the unseen trials, and that subsequent transcranial direct current stimulation (tDCS) of prefrontal cortex modulated delayed discrimination performance (Figure 1.21C). However, it is unclear whether the observed changes in brain activity are actually causally related to working memory or not: The short duration of the chosen delay period (i.e., 1.5 s) in combination with the sluggishness of the BOLD signal renders it impossible to separate the differential contributions of the encoding, maintenance, and retrieval period to the signal changes observed. Moreover, it is even conceivable that there already existed pre-stimulus differences in activity in these prefrontal areas, potentially reflecting fluctuations in conscious top-down attentional control (and hence in signal amplification) received by unseen correct and unseen incorrect trials. As there was, unfortunately, no control task without a working memory requirement, similar concerns also affect the stimulation paradigm. While subsequent efforts by a second group, who employed an attentional blink paradigm and prolonged the duration of the delay period to 15 s, may have mitigated some of these initial critiques, they, too failed to provide a convincing picture of the neural substrates of non-conscious working memory (Bergström and Eriksson, 2014). In comparison to a target-absent control condition, maintenance of a non-conscious target appeared to recruit right mid-lateral prefrontal cortex specifically during the delay period. However, contrasting the very same control condition with conscious targets revealed no significant BOLD signal changes during the delay period whatsoever, thus rendering the observed activity on the unseen trials an unlikely candidate of non-conscious working memory. Perhaps, here too, subjects exerted more top-down control in an effort to complete the task for non-conscious targets. All in all, then, there thus appears to be fairly convincing, behavioral evidence that non-conscious stimuli may exert a long-lasting influence on behavior. However, the exact nature of this long-lasting blindsight effect is still unknown.

1.5 Outstanding questions

So, it seems as if we are at a bit of an impasse right now. Decades of theoretical reflections and research on conscious perception and working memory point to an intimate and intricate relationship between the two. Both cognitive functions are typically conceptualized in terms of a capacity-limited, central system with a role in abstract, complex behaviors. Amplification and maintenance of information features prominently in both conscious perception and working memory, and a similarly distributed network, centered on fronto-parietal and posterior sensory cortices, appears to subtend both phenomena. Yet, as we have just seen, very recent behavioral (and, to a lesser extent, neuroimaging) evidence challenges these prevailing assumptions and dominant views. Non-conscious stimuli appear to influence behavior for much longer periods of time than previously assumed, and may potentially even recruit some of the same prefrontal brain areas as consciously stored representations do. Moreover, it might even be the case that some of the computations and operations performed by working memory may occur implicitly, outside the realms of our subjective, conscious experience. What are we to make of this? Should we abandon ship and accept the notion of a non-conscious working memory system? Or should we disregard the recent findings, perhaps attributing them to a process other than working memory?

My goal for this thesis was to shed novel insights into some of these questions. Adopting a combination of behavioral, electrophysiological, time-resolved decoding, and modeling techniques, I first attempted to rule out some of the most fundamental objections to the observed long-lasting blindsight effect and then to systematically characterize its neuro-cognitive architecture. The logic behind this approach is simple. In order to determine whether this long-lasting blindsight constitutes genuine non-conscious working memory, we first need to evaluate this effect in terms of alternative explanations (e.g., accidental miscategorization, conscious maintenance of a guess; Chapter 2), and then in light of the characteristics and features of conscious working memory. Here, I chose to focus on two particular properties: its role in storing multiple items and their related temporal order (Chapter 3), and its ability to integrate, manipulate, and transform information in the service of goal-directed behavior (Chapter 4). As is so often the case in science, this endeavor may have led me down a quite unexpected path. I hope that you will enjoy this journey as much as I did!

CHAPTER 2 –
A THEORY OF WORKING MEMORY WITHOUT CONSCIOUSNESS OR SUSTAINED ACTIVITY

Nothing in life is to be feared, it is only to be understood.
Now is the time to understand more, so that we may fear less.
- MARIE SKLODOWSKA-CURIE

2.1 ABSTRACT

Working memory and conscious perception are thought to share similar brain mechanisms, yet recent reports of non-conscious working memory challenge this view. Combining visual masking with magnetoencephalography, we investigate the reality of non-conscious working memory and dissect its neural mechanisms. In a spatial delayed-response task, participants reported the location of a subjectively unseen target above chance-level after several seconds. Conscious perception and conscious working memory were characterized by similar signatures: a sustained desynchronization in the alpha/beta band over frontal cortex, and a decodable representation of target location in posterior sensors. During non-conscious working memory, such activity vanished. Our findings contradict models that identify working memory with sustained neural firing, but are compatible with recent proposals of 'activity-silent' working memory. We present a theoretical framework and simulations showing how slowly decaying synaptic changes allow cell assemblies to go dormant during the delay, yet be retrieved above chance-level after several seconds.

2.2 INTRODUCTION

Prominent theories of working memory require information to be consciously maintained (Baars and Franklin, 2003; Baddeley, 2003; Oberauer, 2002). Conversely, influential models of visual awareness hold information maintenance as a key property of conscious perception, highlighting synchronous thalamocortical activity (Tononi and Koch, 2008), cortical recurrence (Lamme and Roelfsema, 2000), or the sustained recruitment of parietal and dorsolateral prefrontal regions (i.e., the same areas as in working memory; Naghavi and Nyberg, 2005) in a global neuronal workspace (Dehaene and Changeux, 2011; Dehaene and Naccache, 2001). Experimentally, non-conscious priming only lasts a few hundred milliseconds (Dupoux et al., 2008; Greenwald et al., 1996) and unseen stimuli typically fail to induce late and sustained cerebral responses (Dehaene et al., 2014). Conscious perception, in contrast, exerts a durable influence on behavior, accompanied by sustained neural activity (King et al., 2014; Salti et al., 2015; Schurger et al., 2015). The hypothesis of an intimate coupling between conscious perception and working memory is thus grounded in theory and supported by numerous empirical findings.

Recent behavioral and neuroimaging evidence, however, has questioned this prevailing view by suggesting that working memory may also operate non-consciously. Unseen stimuli may influence behavior for several seconds (Bergström and Eriksson, 2015; Soto and Silvanto, 2014). Soto and colleagues (Soto et al., 2011), for instance, showed that participants recalled the orientation of a subjectively unseen Gabor cue above chance-level after a 5s-delay. Functional magnetic resonance imaging suggests that prefrontal activity may underlie such non-conscious working memory (Bergström and Eriksson, 2014; Dutta et al., 2014).

The verdict for non-conscious working memory is far from definitive, however. Delayed performance with subjectively unseen stimuli was barely above chance (Soto et al., 2011) and could have arisen from a

small percentage of errors in visibility reports, with subjects miscategorizing a seen target as unseen (miscategorization hypothesis). If this were the case, then the blindsight trials, on which subjects correctly identified the target while denying any subjective awareness of the stimulus, should display similar, if not identical, neural signatures and contents as the seen trials. Alternatively, participants could also have ventured a guess about the target as soon as it appeared and consciously maintained this early guess (conscious maintenance hypothesis). Many priming studies have shown that fast guessing results in above-chance objective performance with subjectively unseen stimuli (Merikle et al., 2001). The observed blindsight effect would then reflect a normal form of conscious working memory (Stein et al., 2016). This alternative hypothesis is hard to eliminate on purely behavioral grounds; it can only be rejected by tracking the dynamics of working memory activity, for instance using brain imaging, and determining whether this activity occurs immediately after the target even on unseen trials.

Here, we set out to address these issues, focusing on four main objectives: First, we probed the replicability of the long-lasting blindsight effect reported by Soto et al. (2011) as well as its robustness with respect to interference from distraction and a conscious working memory load in order to delineate it from other forms of prolonged iconic or sensory memory. Second, we interrogated the link between conscious perception and conscious working memory, evaluating whether the maintenance period in working memory could be likened to a prolongation of a conscious episode. Third, we tested the reality of non-conscious working memory by systematically examining the neural correlates of the blindsight effect and using them to assess the above two alternative hypotheses (the miscategorization and conscious maintenance hypothesis). Lastly, we propose a neuronal theory to offer a mechanistic account of conscious and non-conscious working memory.

2.3 Results

We combined magnetoencephalography (MEG) with a spatial masking paradigm to assess working memory performance under varying levels of subjective visibility (Figure 2.1A and Methods). On 80% of the trials, a target square was flashed in 1 of 20 locations and then masked. Subjects were asked to localize the target after a variable delay (2.5–4.0 s) and to rate its visibility on a scale from *1* (not seen) to *4* (clearly seen). On the remaining 20% of trials, the target was omitted, allowing us to contrast brain activity between target-present and -absent trials. A visible

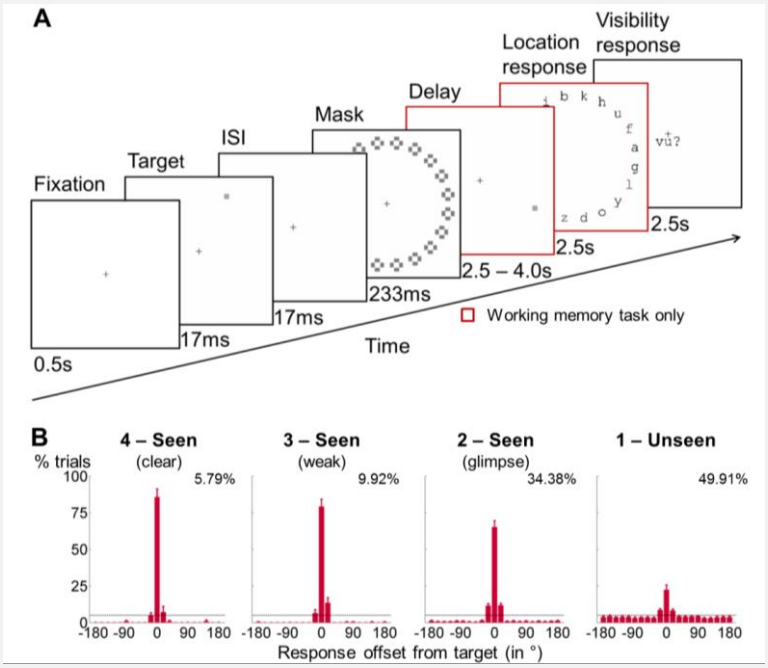


Figure 2.1

General experimental design and behavioral performance in the working memory task.

(A) Experimental design. A subsequently masked target square was flashed in 1 out of 20 positions. Subjects were asked to report this location after a delay of up to 4 s and to rate the visibility of the target on a 4-point scale. A visible distractor square with features otherwise identical to the target was shown on 50% of the trials during the retention period (at 1.75 s). In a perception-only control condition, the maintenance phase and location response were omitted, and subjects assessed the visibility of the target immediately after the mask.
(B) Spatial distributions of forced-choice localization performance in the working memory task (experiment 1; 0 = correct target location; positive = clockwise offset). Error bars indicate standard error of the mean (SEM) across subjects. The horizontal, dotted line illustrates chance-level at 5%. Percentages show proportion of target-present trials from a given visibility category. Due to low number of trials in individual visibility ratings *2*, *3*, and *4*, all *seen* categories were collapsed for analyses.

distractor square was presented 1.5 s into the delay period on half the trials, challenging participants' resistance to distraction and enabling us to evaluate the robustness of the blindsight effect behaviorally. In addition to this working memory task, subjects also completed a perception-only control condition without the delay and target-localization periods (perception task), so that we could isolate brain activity specific to conscious perception (without a working memory requirement) and investigate its link with working memory.

2.3.1 Behavioral maintenance and shielding against distraction

We first examined objective performance in the working memory task as a function of target visibility. Overall, subjects reported the exact target location with high accuracy on seen trials (collapsed across visibility ratings > *1*: $M_{correct}$ = 69.1%, $SD_{correct}$ = 17.4%; chance = 5%; $t(16) = 15.2$, $p < .001$, 95% CI = [55.2%, 73.1%]; Cohen's d = 3.7). As subjective visibility of the target increased from glimpsed (visibility = *2*) to clearly seen (visibility = *4*), there was a corresponding monotonic increase in accuracy (*ps* < .05 for all pair-wise comparisons; Figure 2.1B). Crucially, performance remained above chance even on unseen trials (rating = *1*: $M_{correct}$ = 22.4%, $SD_{correct}$ = 13.8%; $t(16) = 5.2$, $p < .001$, 95% CI = [10.3%, 24.4%]; Cohen's d = 1.3). This blindsight remained substantial after a 4s-delay ($M_{correct}$ = 21.1%, $SD_{correct}$ = 14.7%; $t(16)$ = 4.5, $p < .001$, 95% CI = [8.5%, 23.7%]; Cohen's d = 1.0).

Spatial distributions of participants' responses were concentrated around the target (Figure 2.2A). To correct for small errors in localization, we computed the rate of correct responding with a tolerance of two positions (+/- 36°) surrounding the target location. In subjects displaying above-chance blindsight (chance = 25%; $p < .05$ in a χ^2-test; n = 13), we estimated the precision of working memory as the standard deviation of the distribution within this tolerance interval (Methods). Performance was better on seen than on unseen trials, both in terms of rate of correct responding ($F(1, 16) = 198.5$, $p < .001$; partial η^2 = .925) and precision ($F(1, 12) = 36.7$, $p < .001$; partial η^2 = .754). There was neither an effect of the distractor on these measures

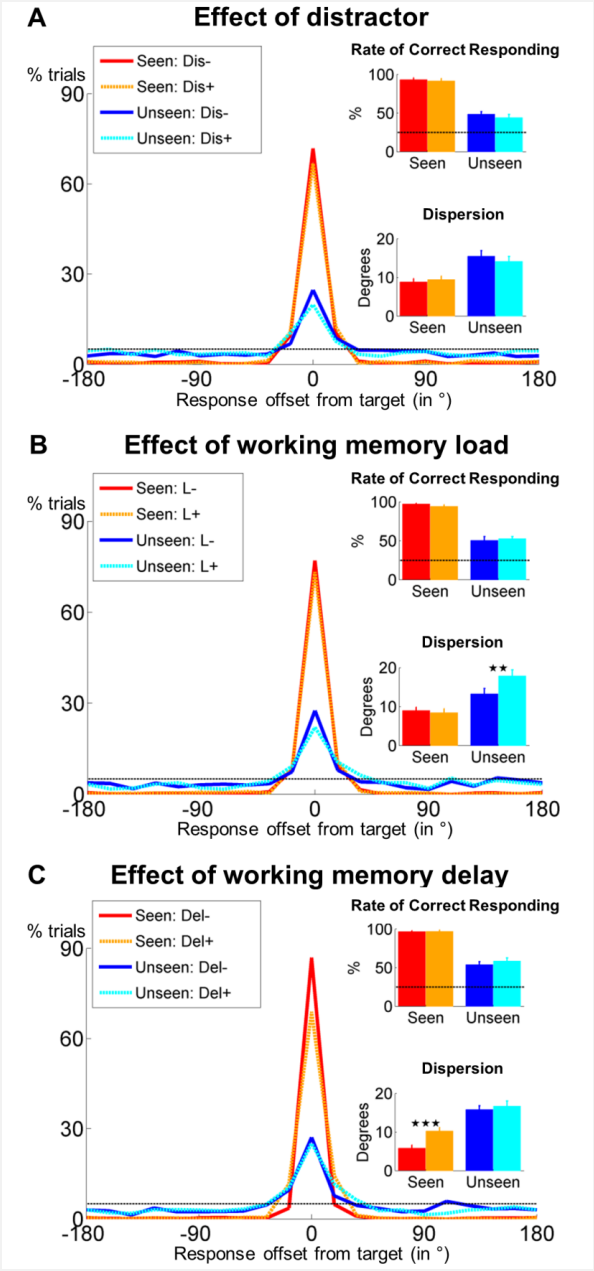


Figure 2.2

Behavioral evidence for non-conscious working memory.

Spatial distributions of responses (0 = correct target location; positive = clockwise offset) as a function of visibility and distractor presence **(A)**, conscious working memory load **(B)** and delay duration **(C)**. Insets show rate of correct responding (within +/- 2 positions of actual location) and precision of working memory representations separately for seen and unseen trials. Error bars represent standard error of the mean (SEM) across subjects and horizontal, dotted line indicates chance-level (5%). *$p < .05$, **$p < .01$, and ***$p < .001$ in a paired samples t-test. Del = delay, Dis = distractor, L = load.

(all *p*s > .079), nor any significant interactions between distractor and visibility (all *p*s > .251), indicating that distractor presence did not affect retention for seen or unseen targets. Restricting the analyses to trials within one position of the actual target location (+/- 18°) or to the subgroup of 13 subjects included in the MEG analyses did not change these findings qualitatively.

While target detection *d'* exceeded chance-level (*M* = 1.5, *SD* = 0.7; *t*(16) = 8.9, *p* < .001, 95% CI = [1.2, 1.9]; Cohen's *d* = 2.1) and correlated with accuracy and the rate of correct responding on seen trials (both Pearson *r*s > .762, both *p*s < .001), there was no relationship between our participants' sensitivity to the target and any of our performance measures on the unseen trials (all Pearson *r*s < .342, all *p*s > .179; Figure 2.2 - Figure Supplement 1A). Thus, target visibility predicted performance in the objective working memory task only on seen trials, but not on unseen trials.

Overall, these results confirm, with much higher non-conscious performance, the observations of previous studies (Soto et al., 2011): Non-conscious information may be maintained for up to 4 s and successfully shielded against distraction from a salient visual stimulus, independently of overall subjective visibility.

2.3.2 Resistance to conscious working memory load and delay duration

To probe the similarity between conscious working memory and the observed long-lasting blindsight effect, in a second behavioral experiment with 21 subjects, we examined whether imposing a load on conscious working memory (remembering digits) affected non-conscious performance. On each trial, 1 (low load) or 5 (high load) digits were simultaneously shown for 1.5 s, followed by a 1s-fixation period and the same sequence of events (target and mask) as in experiment 1. After a variable delay (0 or 4 s), participants had to (1) localize the target, (2) recall the digits in the correct order, and (3) rate target visibility.

Subjects again chose the exact target position with high accuracy on seen trials ($M_{correct}$ = 77.8%, $SD_{correct}$ = 13.9%) and remained above chance on unseen trials ($M_{correct}$ = 25.6%, $SD_{correct}$ = 11.8%; chance = 5%; *t*(18) = 7.6, *p* < .001, 95% CI = [14.9%, 26.3%]; Cohen's *d* = 1.7). While, as in experiment 1, cue detection *d'* was greater than chance (*M* = 1.7, *SD* = 0.8; *t*(20) = 10.2, *p* < .001, 95% CI = [1.4, 2.1]; Cohen's *d* = 2.2), no correlations were observed with objective task performance on the unseen trials (all Pearson *r*s < .366, all *p*s > .115; seen trials: all Pearson *r*s > .443, all *p*s < .051; Figure 2.2 - Figure Supplement 1B). As expected, participants were better at recalling 1 rather than 5 digits in the correct order (*M* = 93.3% vs. 89.5%, *F*(1, 17) = 4.7, *p* = .045), irrespective of target visibility or delay duration (all *p*s > .135).

Analyzing only the trials with correctly recalled digits, we observed an impact of load on the precision with which target location was retained (*F*(1, 13) = 7.3, *p* = .018; partial η^2 = .360). Crucially, load modulated the relationship between precision and visibility (interaction *F*(1, 13) = 8.7, *p* = .011; partial η^2 = .400), with no effect on seen (*t*(13) = 0.6, *p* = .561) and a strong reduction of precision on unseen trials (*t*(13) = -3.6, *p* = .004). There was no effect of working memory load on the rate of correct responding (all *p*s > .229; Figure 2.2B).

Delay duration (0 or 4 s) also did not influence the rate of correct responding (all *p*s > .082; Figure 2.2C). It did, however, affect overall precision (*F*(1, 15) = 9.3, *p* = .008; partial η^2 = .383) and the relationship between precision and visibility (interaction *F*(1, 15) = 5.2, *p* = .037; partial η^2 = .259). This interaction was driven by higher precision on no-delay than on 4s-delay trials, exclusively when subjects had seen the target (*t*(15) = -5.7, *p* < .001; unseen trials: *t*(15) = -0.6, *p* = .559).

Overall, these results highlight the replicability and robustness of the long-lasting blindsight effect and suggest that it does not just constitute a prolonged version of iconic memory: Even in the presence of a concurrent conscious working memory load, unseen stimuli could be maintained, with no detectable decay as a function of delay. However, the systems involved in the short-term maintenance of conscious and

non-conscious stimuli interacted, because a conscious verbal working memory load diminished the precision with which non-conscious spatial information was maintained.

2.3.3 Similarity of conscious perception and conscious working memory

To tackle our second objective – a detailed examination of the link between conscious perception and conscious working memory –, we turned to our MEG data and first ensured that the mechanisms underlying conscious perception were stable across experimental conditions. The subtraction of the event-related fields (ERFs) evoked by unseen trials from those evoked by seen trials revealed similar topographies for the perception and working memory task (Figure 2.3A): Starting at ~300 ms and extending until ~500 ms after target onset, a response emerged over right parieto-temporal magnetometers. This divergence resulted primarily from a sudden increase in activity on seen trials ("ignition") in the perception ($p_{FDR} < .05$ from 384–416 ms and from 504–516 ms) and working memory task ($p_{FDR} < .05$ from 328–364 ms and from 396–404 ms; Figure 2.3B). The observed topographies and time courses fall within the time window of typical neural markers of conscious perception, including the P3b (e.g., Del Cul et al., 2007; Salti et al., 2015; Sergent et al., 2005). Consciously perceiving the target stimulus therefore involved comparable neural mechanisms, irrespective of task.

We next directly probed the relationship between conscious perception and information maintenance in conscious working memory. Does the latter reflect a prolonged conscious episode, or does it involve a distinct set of processes recruited only during the retention phase? If conscious working memory can indeed be likened to conscious perception, one might expect the same patterns that index such perception to be sustained throughout the working memory maintenance period. Linear multivariate pattern classifiers were trained to predict visibility (seen or unseen) from MEG signals separately for each task. Classification performance was assessed during an early time period (100–300 ms), the critical P3b time window (300–600 ms), and the first (0.6–1.55 s) and second part (1.55–2.5 s) of the delay period.

Decoding of the visibility effect was comparable in the two tasks (Figure 2.3C and online Table 1): Classification performance rose sharply between 100 and 300 ms and peaked during the P3b time window (all *p*s < .007, except 100–300 ms in the working memory task, where $p = .066$). It then decayed slowly from ~1 s onwards in both tasks, yet remained above chance during the 0.6–1.55 s interval (all *p*s < .001). Similar time courses were also observed when training in one task and testing for generalization to the other. Though rapidly dropping to chance-level after ~1 s, classifiers trained in the perception task performed above chance during the first three time windows on working memory trials (and vice versa; all *p*s < .014), indicating that, early on, both tasks recruited similar brain mechanisms.

Temporal generalization analyses (King and Dehaene, 2014) were used to evaluate the onset and duration of patterns of brain activity. If working memory were just a prolonged conscious episode, classifiers trained at time points relevant to conscious perception (e.g., the P3b window) should generalize extensively, potentially spanning the entire delay. Our findings supported this hypothesis only in part. The temporal generalization matrix for the working memory task presented as a thick diagonal, suggesting that brain activity was mainly characterized by changing, but long-lasting patterns. Though failing to achieve statistical significance over the entire 0.6–1.55 s interval (all *p*s > .101), at a more lenient, uncorrected threshold, classifiers trained during the P3b time window (300–600 ms) in the working memory task remained weakly efficient until ~692 ms (AUC = 0.54 +/- 0.02, $p_{uncorrected} = .023$). Similarly, classifiers trained during the same time period in the perception task and tested in the working memory task persisted up to ~860 ms (AUC = 0.53 +/- 0.01, $p_{uncorrected} = .028$). Brain processes deployed for the conscious representation of the target were thus partially sustained during the working memory delay. The reverse analysis, in which we trained classifiers during the retention period in the working memory task (0.8–2.5 s), did not reveal any generalization to the P3b time window in the perception task ($p = .101$).

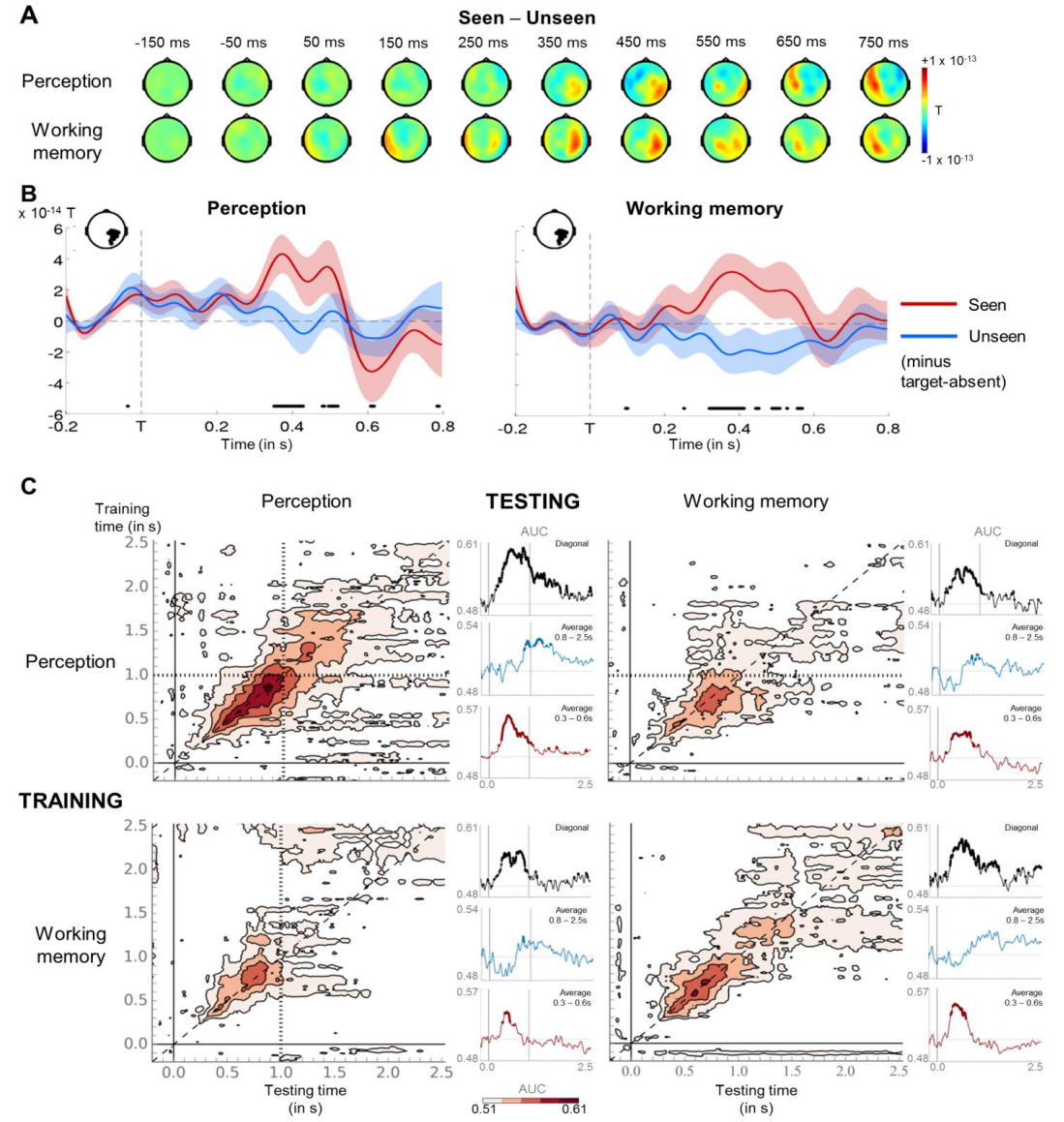


FIGURE 2.3

NEURAL SIGNATURES OF CONSCIOUS PERCEPTION AND MAINTENANCE IN WORKING MEMORY.

(A) Sequence of brain activations (-200–800 ms) evoked by consciously perceiving the target in the perception (top) and working memory (bottom) task. Each topography depicts the difference in amplitude between seen and unseen trials over a 100 ms time window centered on the time points shown (magnetometers only).
(B) Average time courses of seen and unseen trials (-200–800 ms) after subtraction of target-absent trials in a group of parietal magnetometers in the perception (left) and working memory (right) task. Shaded area illustrates standard error of the mean (SEM) across subjects. Significant differences between conditions are depicted with a horizontal, black line (Wilcoxon signed-rank test across subjects, uncorrected). For display purposes, data were lowpass-filtered at 8 Hz. T = target onset.
(C) Temporal generalization matrices for decoding of visibility category as a function of training and testing task. In each panel, a classifier was trained at every time sample (y-axis) and tested on all other time points (x-axis). The diagonal gray line demarks classifiers trained and tested on the same time sample. Please note the event markers in any panel involving the perception task: Mean reaction time (target-present trials) for the visibility response is indicated as vertical and/or horizontal, dotted lines. Any classifier beyond this point only reflects post-visibility processes. Time courses of diagonal decoding and of classifiers averaged over the P3b time window (300–600 ms) and over the working memory maintenance period (0.8–2.5 s) are shown as black, red, and blue insets. Thick lines indicate significant, above-chance decoding of visibility (Wilcoxon signed-rank test across subjects, uncorrected, two-tailed except for diagonal). For display purposes, data were smoothed using a moving average with a window of eight samples. AUC = area under the curve.

These results confirm that seeing the target entailed a similar unfolding of neural events in two task contexts: Conscious perception primarily consisted in a dynamic series of partially overlapping information-processing stages, each characterized by temporary, metastable patterns of neural activity. The same neural codes appeared to be recruited at the beginning of the maintenance period (up to ~1 s). As such, these findings corroborate previous accounts linking conscious perception to an “ignition” of brain

activity (Del Cul et al., 2007; Gaillard et al., 2009; Salti et al., 2015; Sergent et al., 2005) and suggest that, in part, working memory implies the prolongation of a conscious episode, and, in part, a succession of additional processing steps.

2.3.4 A SUSTAINED DECREASE IN ALPHA/BETA POWER DISTINGUISHES CONSCIOUS WORKING MEMORY

Our focus so far has been on evoked brain activity. However, other reliable neural signatures of conscious perception have been identified in the frequency domain (Gaillard et al., 2009; Gross et al., 2007; King et al., 2016; Wyart and Tallon-Baudry, 2009). We thus turned to time-frequency analyses and first contrasted seen trials with both our target-absent control condition as well as unseen trials in both tasks (Figure 2.4A and Figure 2.4 - Figure Supplement 1A). In order to qualify as a signature of conscious perception, any candidate characteristic should exist in the perception-only control condition (without any working memory requirement) and be specific to seen trials. Cluster-based permutation analyses singled out a desynchronization in the alpha band (8–12 Hz) as the principal correlate of conscious perception in the perception task (seen – target-absent: p_{clust} = .004; seen – unseen: p_{clust} = .009), with seen trials displaying a strong decrease in power (relative to baseline) compared to either the target-absent or the unseen trials. Initially left-lateralized in centro-temporal sensors, this effect moved to fronto-central channels and extended between ~300 and 1700 ms. A similar, albeit later (500–1700 ms) and more bilateral fronto-central, desynchronization was also observed in the beta band (13–30 Hz; seen – target-

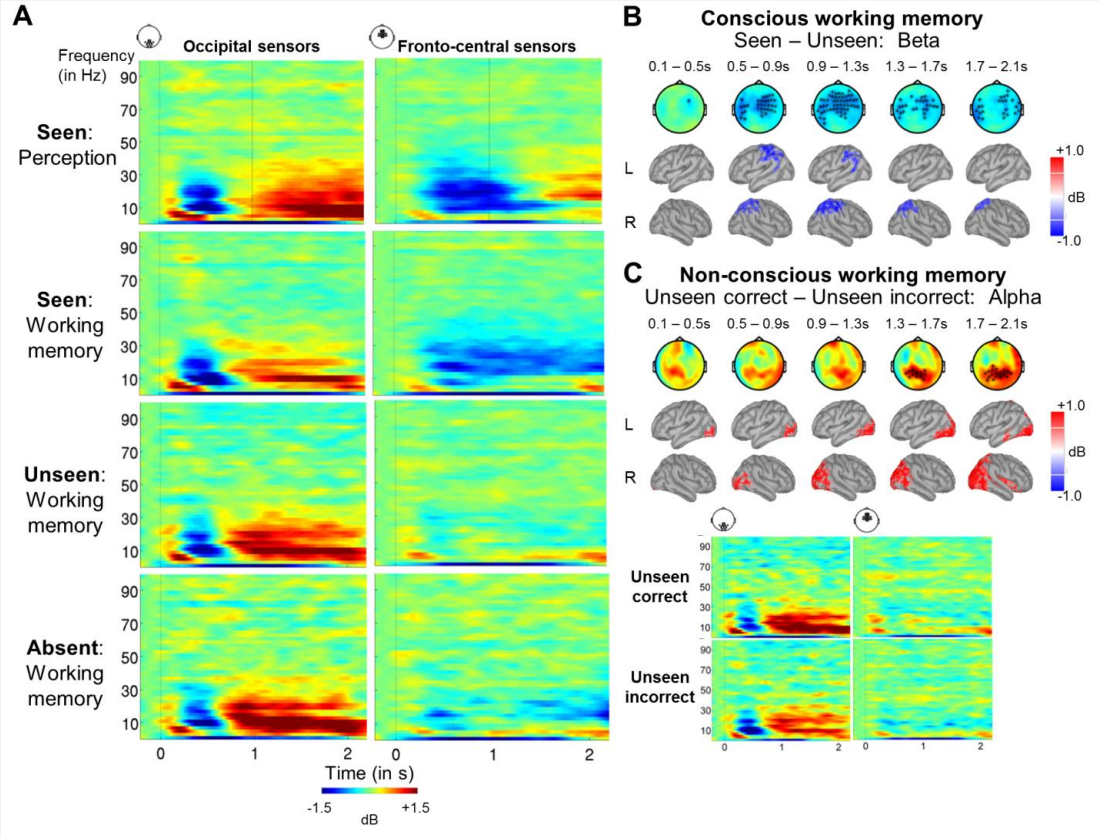


FIGURE 2.4

A SUSTAINED DECREASE IN ALPHA/BETA POWER AS A MARKER OF CONSCIOUS WORKING MEMORY.

(A) Average time-frequency power relative to baseline (dB) as a function of task and visibility category in a group of occipital (left) and fronto-central (right) magnetometers. Mean reaction time (target-present trials) for the visibility response in the perception task is indicated as a vertical, dotted line.
(B) Beta band activity (13–30 Hz; 0–2.1 s) related to conscious working memory (seen – unseen trials) as shown in magnetometers (top) and source space (bottom; in dB relative to baseline). Black asterisks indicate sensors showing a significant difference as assessed by a Monte-Carlo permutation test.
(C) Same as in (A) and (B) but for unseen correct and unseen incorrect trials in the alpha band (8–12 Hz).

absent: p_{clust} < .001; seen – unseen: p_{clust} = .01). No differences between the unseen and target-absent trials were found in the alpha (p_{clust} > .676) or beta band (p_{clust} > .226, apart from a short-lived, weak difference between ~0.9 and 1.3 s, where p_{clust} = .020), suggesting that unseen trials strongly resembled trials without a target.

Most importantly, when comparing seen and target-absent/unseen trials in the working memory task, we again observed a similar, but now temporally sustained, pattern of alpha/beta band desynchronization (Figure 2.4B and Figure 2.4 - Figure Supplement 1B). Starting at ~300 to 500 ms, seen targets evoked a power decrease in central, temporal/parietal, and frontal regions in the alpha (seen – target-absent: p_{clust} = .003; seen – unseen: p_{clust} = .003) and beta band (seen – target-absent: p_{clust} = .009; seen – unseen: p_{clust} < .001). Crucially, this desynchronization spanned the entire delay period and was specific to seen trials (Figure 2.4A), with no differences in power between the unseen and target-absent trials in either band (alpha: p_{clust} > .729; beta: p_{clust} > .657) and only a couple of interspersed periods of residual desynchronization persisting in the target-absent control trials. No task- or visibility-related modulations in power spectra were found in occipital areas, and the desynchronization originated primarily from a parietal network of brain sources (Figure 2.4A and B). In conjunction with the afore-mentioned results, these findings imply that alpha/beta desynchronization is a correlate of conscious perception (Gaillard et al., 2009) and a neural state common to conscious perception and conscious working memory.

2.3.5 A DISTINCT NEUROPHYSIOLOGICAL MECHANISM FOR NON-CONSCIOUS WORKING MEMORY

Having identified markers of conscious perception and working memory in both multivariate and time-frequency analyses, we can now test the reality of non-conscious working memory by confronting it with several alternative hypotheses. The miscategorization hypothesis suggests that the long-lasting blindsight resulted from a small set of seen trials erroneously labeled as unseen. Unseen correct trials should thus display similar neural signatures as seen trials, including a shared discriminative decoding axis and a desynchronization in the alpha/beta band. An analogous reasoning holds for the conscious maintenance hypothesis, according to which the observed blindsight effect arises from the conscious maintenance of an early guess: Conscious processing would occur on unseen trials and we should thus find a sustained decrease in alpha/beta power similar to the one on seen trials. Conversely, a clear distinction between brain responses on seen trials and on unseen (correct) trials would suggest that blindsight resulted from a distinct non-conscious mechanism of information maintenance.

We first probed the alternative hypotheses with the ERF data. Training a decoder to distinguish seen from unseen trials in the perception task and applying it to the unseen correct and incorrect trials in the working memory task, we directly assessed the classifier's ability to generalize from seen to unseen correct trials (accuracy decoder). If, indeed, the latter had actually been seen, such a decoder should look similar to the above-described generalization analysis, in which a classifier had been trained on seen/unseen trials in the perception task and tested on the same labels in the working memory task (visibility decoder). As shown in Figure 2.4 - Figure Supplement 2A, this was not the case. Whereas the temporal generalization matrix for the visibility decoder presented as a thick diagonal, no discernable pattern emerged for the accuracy decoder. The time courses of diagonal decoding were also quite dissimilar. For the visibility decoder (see also above), classification performance first rose above chance at ~148 ms (AUC = 0.54 +/- 0.01, p_{FDR} = .023), peaked at ~640 ms (AUC = 0.58 +/- 0.02, p_{FDR} = .001), and then decayed rapidly by ~1 s (first three time windows: all *p*s < .001). In contrast, classification for the accuracy decoder was erratic and transient: It first sharply peaked at ~180 ms (AUC = 0.55 +/- 0.01, $p_{uncorrected}$ = .037), dropped to chance-level, and then exceeded chance between ~372 and 724 ms with a peak at 444 ms (AUC = 0.57 +/- 0.02, $p_{uncorrected}$ = .007). Much unlike any of the previous decoders involving the perception task, long after the visibility response, it rose a third time between ~1.44 and 1.74 s, peaking with similar magnitude as before at ~1.58 s (AUC = 0.57 +/- 0.02, $p_{uncorrected}$ = .010; P3b and last time window: all *p*s < .023). Although the level of noise evident in the accuracy decoder thus precludes any definitive conclusion, the visibility and

accuracy decoders had little in common, rendering it unlikely for the unseen correct trials to have simply been mislabeled.

We next turned to time-frequency analysis. When averaging over all unseen trials in the working memory task, there was no indication of a desynchronization remotely comparable to the one on seen trials (Figure 2.4A and Figure 2.4 - Figure Supplement 1C). Indeed, Bayesian statistics indicated that, on the unseen trials, evidence for the null hypothesis (i.e., no relative change in alpha/beta power) was at least similar (at the very end of the epoch) or stronger than evidence for the alternative hypothesis. By contrast, on seen trials, evidence for the alternative hypothesis was always strongly favored (Figure 2.4 - Figure Supplement 3). Even when analyzing the unseen correct trials separately, there was no appreciable trace of any alpha/beta desynchronization (Figure 2.4C and Figure 2.4 - Figure Supplement 3). Only one short-lived effect, reversed relative to conscious trials, was observed in the alpha band (p_{clust} = .040) in a set of posterior central sensors, corresponding to primarily occipital sources: Starting at ~1.5 s and extending until ~1.9 s, unseen correct trials exhibited a stronger *increase* in alpha power than their incorrect counterparts. Given the difference in performance on these two types of unseen trials, such small variations are not surprising and could, perhaps, reflect a stronger suppression of interference from the distractor on the unseen correct trials. Unseen correct trials thus appeared to be nearly indistinguishable from the unseen incorrect and target-absent trials.

As multivariate analyses might be more sensitive than univariate ones in detecting similarities between conditions, we also performed the above decoding analysis separately for average alpha (8–12 Hz) and beta (13–30 Hz) power. Overall, these analyses confirmed our previous findings, albeit more clearly so in the alpha than in the beta band. A visibility decoder trained on alpha power to distinguish seen from unseen trials in the perception task and tested in the working memory task again exhibited a thick diagonal, with above-chance decoding between ~180 ms and 1.18 s (first three time windows: all *p*s < .016). There was no evidence for any generalization to the unseen correct trials (Figure 2.4 - Figure Supplement 2B; all time windows: *p*s > .211). Similarly, a visibility decoder trained on average beta power entirely failed to generalize to the unseen correct trials (Figure 2.4 - Figure Supplement 3C; all time windows: *p*s > .191). Considering the weak, although statistically significant (all four time windows: *p*s <= .05), initial generalization from the perception to the working memory task, probably due to the slightly later onset of the beta desychronization in the former, this failure is less informative than the one observed in the alpha band and should be replicated in future investigations.

Taken together, we found a clear distinction in the brain responses of seen and unseen (correct) trials. Converging evidence from our decoding analyses in the ERFs and alpha/beta band suggests that there was no apparent discriminative axis shared between the seen and the unseen correct trials. Similarly, the desynchronization in alpha/beta power characterizing the seen targets did not emerge on the unseen (correct) trials. These findings therefore argue against the miscategorization and conscious maintenance hypotheses and instead suggest that non-conscious working memory is a genuine phenomenon, distinct from conscious working memory.

2.3.6 Contents of conscious and non-conscious working memory can be tracked transiently

We next set out to identify the neural mechanisms supporting both conscious and non-conscious working memory and first determined where and how the specific contents of working memory were stored. Circular-linear correlations between the amplitude of the ERFs and target location (across all working memory trials) revealed a strong and focal association (relative to a permuted null distribution) over posterior channels, starting at ~120 ms and lasting until 904 ms (early and P3b time windows: all *p*s < .001; all BFs > 109.60; Figure 2.5A and online Tables 2 and 3). Similarly, distractor position could be tracked between ~194 and 570 ms after its presentation (early and P3b time windows: all *p*s < .009; all BFs > 14.47). The position of our stimuli could thus be faithfully retrieved in visual areas.

In a subsequent step, we investigated how target location would be maintained in the context of conscious and non-conscious working memory (Figure 2.5B). Target position was transiently encoded via slowly decaying activity in occipital as well as bilateral temporo-occipital cortex from ~120 to 800 ms on

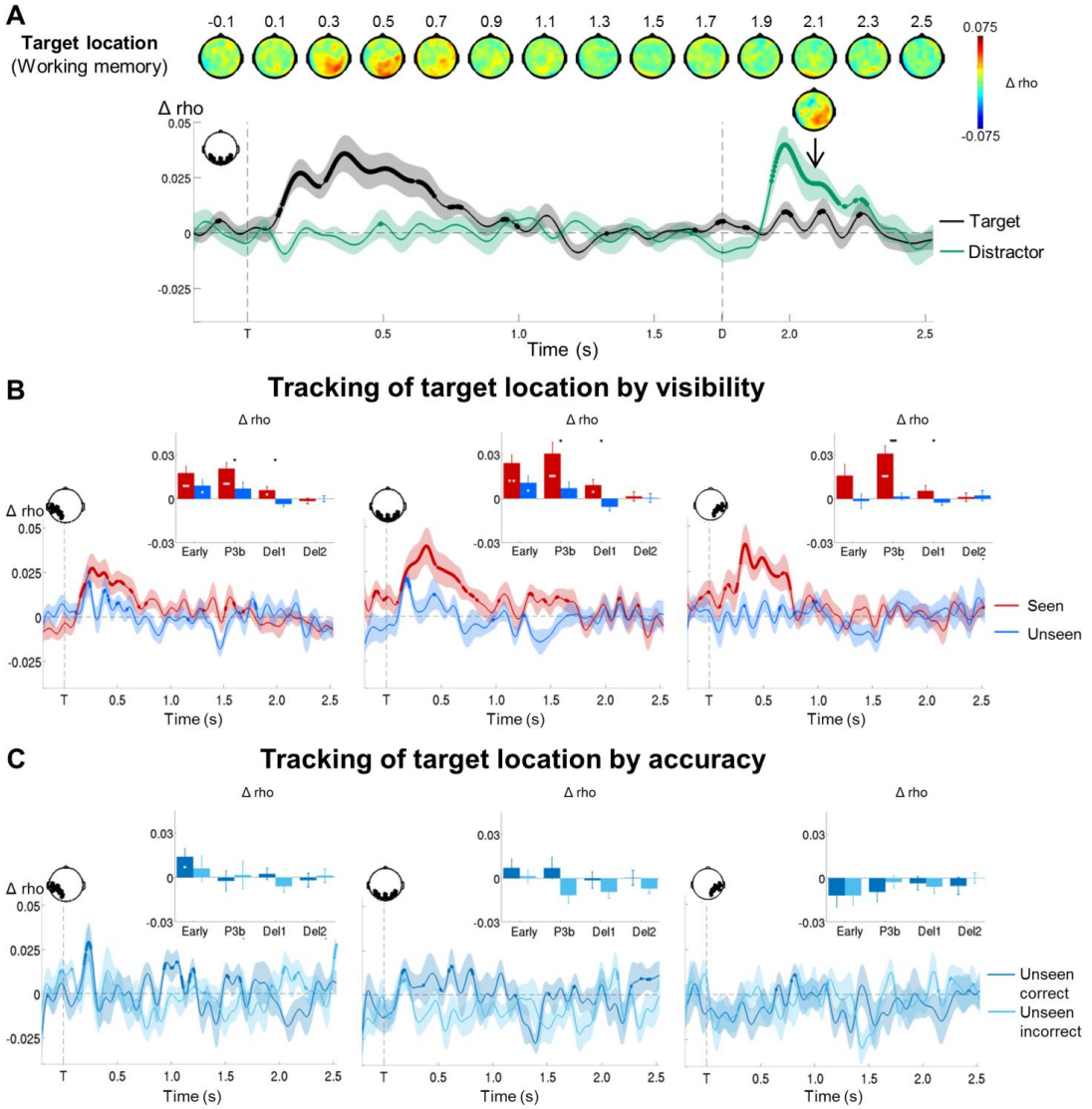


FIGURE 2.5

TRACKING THE CONTENTS OF CONSCIOUS AND NON-CONSCIOUS WORKING MEMORY.

(A) Topographies (top) and time courses (bottom; -0.2–2.5 s) of average circular-linear correlations between the amplitude of the MEG signal (gradiometers) and target/distractor location. Shaded area demarks standard error of the mean (SEM) across subjects. Thick line represents significant increase in correlation coefficient as compared to an empirical baseline (one-tailed Wilcoxon signed-rank test across subjects, uncorrected).
(B) Average time courses (-0.2–2.5 s) of circular-linear correlation coefficients between amplitude of the ERFs and target location as a function of visibility in the working memory task in a group of left temporo-occipital (left), occipital (middle), and right temporo-occipital (right) gradiometers. Shaded area demarks standard error of the mean (SEM) across subjects. Thick line represents significant increase in correlation coefficient as compared to an empirical baseline (one-tailed Wilcoxon signed-rank test across subjects, uncorrected). Insets show average correlation coefficients (relative to an empirical baseline) in four time windows: 100–300 ms (early), 300–600 ms (P3b), 0.6–1.55 s (Del1), and 1.55–2.5 s (Del2). White asterisks denote significant differences to baseline (one-tailed Wilcoxon signed-rank test across subjects), black asterisks significant differences between conditions (two-tailed Wilcoxon signed-rank test across subjects). For display purposes, data were lowpass-filtered at 8 Hz. *p < .05, **p < .01, and ***p < .001. Del1= first part of delay, Del2 = second part of delay, T = target onset.
(C) Same as in (B), but as a function of accuracy on the unseen trials (correct = within +/-2 positions of the target).

seen trials (early and P3b time windows: all ps < .001 and all BFs > 24.07, with the exception of the 100–300 ms period in right temporo-occipital channels, where p = .064 and BF = 2.31) and in occipital and left temporo-occipital brain areas from ~180 to 504 ms on unseen trials (early time window: all ps < .047; all

BFs > 2.58). A clear correlation with target location was therefore found for both seen and unseen trials. In fact, although it was more short-lived on the latter, it was of comparable magnitude as the one observed on the seen trials during the early time window (occipital/left temporo-occipital channels: all *p*s > .110 when directly comparing the correlation scores of seen and unseen trials in a Wilcoxon signed-rank test). In the case of seen trials, both occipital and left temporo-occipital cortex also maintained the target representation at least throughout the first part of the delay period (all *p*s < .024; all BFs > 3.77), though, intriguingly, this was not accompanied by continuously sustained activity. Target "decodability" instead waxed and waned, appearing and disappearing periodically. No such activity was observed for the maintenance of unseen targets (first and second part of the delay: all *p*s > .446; all BFs < .047). This absence of "decodability" during the maintenance period persisted, even when considering unseen correct and unseen incorrect trials separately (Figure 2.5C). There was only a trace of residual decoding of target location on unseen correct trials in left temporo-occipital areas during the delay period, but this did not reach significance, potentially due to the low number of trials in this condition. Note that in the perception task, seen targets could be retrieved similarly to their counterparts in the working memory task between ~232 and 1184 ms in occipital and bilateral temporo-occipital regions (all *p*s > .068, except for the 100–300 ms time window in occipital channels where *p* = .008, when directly comparing the correlation scores of seen targets in both tasks in a Wilcoxon signed-rank test; Figure 2.5 - Figure Supplement 1).

Given the univariate nature of the circular-linear correlations, one might again wonder whether a multivariate strategy would be more sensitive in detecting subtle associations between the MEG data and target location. We therefore used linear support vector regressions (SVR) to predict target angle from the MEG signal as a function of visibility (Methods). As can be seen in Figure 2.5 - Figure Supplement 2, this method resulted in similar, albeit more noisy, time courses as the ones obtained with the circular-linear correlations: Seen targets were again encoded and maintained intermittently between ~268 ms and 1.4 s (P3b time window and first part of the delay: *p*s < .05). No statistically significant decoding emerged for unseen target locations. Due to the fact that subjects responded correctly on approximately half of all unseen trials (see online Table 4 for average trial counts), we attempted to evaluate the dynamics of the encoding and maintenance of unseen correct and incorrect target locations by training the regression model on the strongest case, the seen correct trials, and applying it separately to the unseen correct and incorrect trials. We again observed no evidence for any generalization at all (Figure 2.5 - Figure Supplement 3A), though this likely reflects the sensitivity of the analysis more so than any meaningful effect.

Taken together, in line with previous research (Harrison and Tong, 2009; King et al., 2016), these results suggest that posterior sensory regions may initially encode seen and unseen memoranda via slowly decaying neural activity. In the case of conscious working memory, these then seem to be maintained by those same areas through an intermittently reactivated, neural code (Fuentemilla et al., 2010). In contrast, no such periodically resurfacing activity appears to accompany non-conscious working memory.

2.3.7 Further evidence against the conscious maintenance hypothesis

The correlation between target location and brain activity affords an additional way to interrogate the conscious maintenance hypothesis. If subjects quickly guessed the location of an unseen target and then held it in conscious working memory, in addition to observing a signature of conscious processing on the unseen trials, we should observe a correlation with the location of their response long before it occurs. Potentially, remembering the response might recruit brain systems completely different from the ones representing the target.

Circular-linear correlations rendered this prediction unlikely. Associations between response location and the MEG signal were again primarily confined to posterior channels, with more frontal areas being recruited preferentially at the time of the response (Figure 2.6A). As such, the topographical patterns were highly similar to the ones observed for the correlation with target location. Importantly, no additional

regions were identified on the unseen trials and none of these areas showed any appreciable correlation before the presentation of the response screen (Figure 2.6 - Figure Supplement 1).

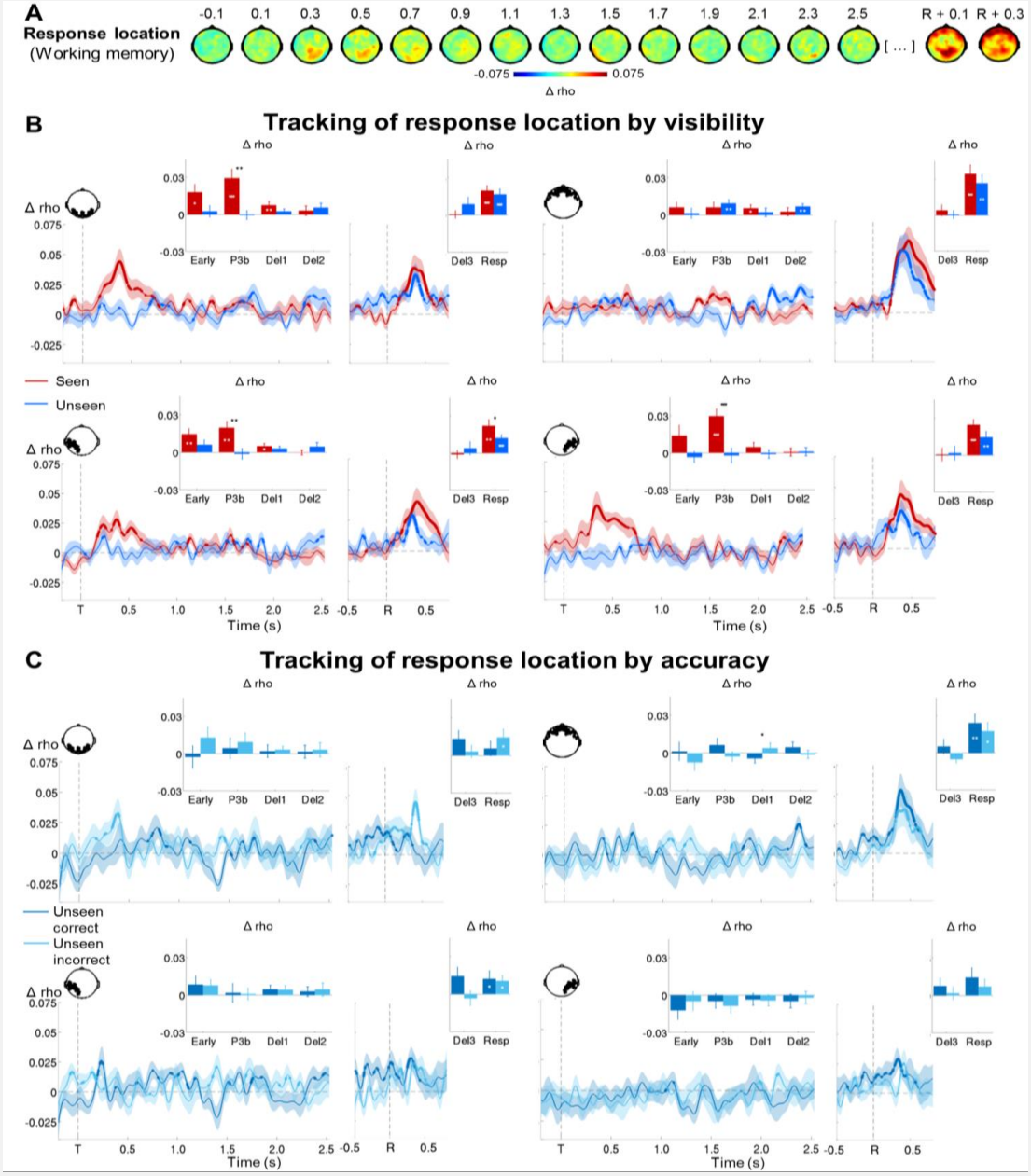


FIGURE 2.6

TRACKING RESPONSE LOCATION IN CONSCIOUS AND NON-CONSCIOUS WORKING MEMORY.

(A) Topographies of average circular-linear correlations between the amplitude of the MEG signal (gradiometers) and response location. R = onset of the response screen.
(B) Average time courses (left: stimulus-locked, -0.2–2.5 s; right: response-locked, -0.5–0.8 s) of circular-linear correlation coefficients between amplitude of the ERFs and response location as a function of visibility in the working memory task in a group of occipital (top, left), frontal (top, right) left temporo-occipital (bottom, left) and right temporo-occipital (bottom, right) gradiometers. Shaded area demarks standard error of the mean (SEM) across subjects. Thick line represents significant increase in correlation coefficient as compared to an empirical baseline (one-tailed Wilcoxon signed-rank test across subjects, uncorrected). Insets show average correlation coefficients (relative to an empirical baseline) in four stimulus-locked time windows, 100–300 ms (early), 300–600 ms (P3b), 0.6–1.55 s (Del1), and 1.55–2.5 s (Del2), and two response-locked time windows, -0.5–0.0 s (Del3) and 0.0–0.8 s (Resp). White asterisks denote significant differences to baseline (one-tailed Wilcoxon signed-rank test across subjects), black asterisks significant differences between conditions (two-tailed Wilcoxon signed-rank test across subjects). For display purposes, data were lowpass-filtered at 8 Hz. *$p < .05$, **$p < .01$, and ***$p < .001$. Del1= first part of delay, Del2 = second part of delay, Del3 = last 500 ms before response screen, R = response screen onset, T = target onset.
(C) Same as in (B), but as a function of accuracy on the unseen trials (correct = within +/-2 positions of the target).

This suggests that, irrespective of stimulus visibility, common brain networks supported memories for the target stimulus and the ensuing decision and that, in the case of non-conscious working memory, these did not come online until the response.

The time courses of the circular-linear correlations further solidified this interpretation (Figure 2.6B). On seen trials, response position was maintained throughout the majority of the epoch in occipital and left temporo-occipital brain areas (first three time windows: all *p*s < .020; all BFs > 4.16). This was not the case on the unseen trials: No correlation patterns appeared in any of the posterior channels during the course of the epoch (all time windows: all *p*s > .064; all BFs < 1.32). In contrast, a strong correlation emerged for both seen and unseen trials during the response period (0–800 ms with respect to the onset of the letter cue). Response location could be tracked with similar time courses and magnitude on seen and unseen trials in occipital, bilateral temporo-occipital, and frontal channels (all *p*s < .024; all BFs > 13.73; when directly comparing the correlation scores of seen and unseen targets in a Wilcoxon signed-rank test: all *p*s > .216, except for left temporo-occipital channels, where *p* = .040). When we further distinguished unseen correct from unseen incorrect trials, the results remained similar, though much noisier (Figure 2.6C): There was no clear correlation pattern before the onset of the response screen on either the unseen correct or the unseen incorrect trials (all *p*s > .096; all BFs < 1.47). Only after the appearance of the letter cues did we observe a correlation with response location.

Multivariate decoding analyses confirmed this picture: Whereas response location for seen targets could be tracked similarly to actual target location at least throughout the first part of the delay period (P3b time window and first part of the delay: *p*s < .05; Figure 2.6 - Figure Supplement 2), no such pattern was observed on the unseen trials (all *p*s > .153). This absence of decodability persisted on the unseen correct and incorrect trials, even when training the regression model on the seen correct trials (Figure 2.5 - Figure Supplement 3B).

Overall, these results are incompatible with the hypothesis that the long-lasting blindsight is only due to the conscious maintenance of an early guess, as, in this case, brain responses linked to the subjects' response should have been observed shortly after the presentation of the target stimulus.

2.3.8 Short-term synaptic change as a neurophysiological mechanism for conscious and non-conscious working memory

What mechanism might permit above-chance recall without any continuously sustained brain activity? Recent modelling suggests that sustained neural firing may not be required to maintain a representation in conscious working memory. Mongillo, Barak, and Tsodyks (2008) proposed a theoretical framework for working memory, in which information is stored in calcium-mediated short-term changes in synaptic weights, thus linking the active cells coding for the memorized item. Once these changes have occurred, the cell assembly may go dormant during the delay, while the synaptic weights are slowly decaying. At the end of the delay period, a non-specific read-out signal may then suffice to reactivate the assembly. Furthermore, reactivation of the assembly may also occur spontaneously during the retention phase, similar to the rehearsal process postulated by Baddeley (2003), thus refreshing the weights and permitting the bridging of longer delays. Could this 'activity-silent' mechanism also constitute a plausible neural mechanism for non-conscious working memory?

To test this hypothesis, we simulated our experiments using a one-dimensional recurrent continuous attractor neural network (CANN) based on Mongillo et al. (2008). The CANN encoded the angular position of the target and was composed of neurons aligned according to their preferred stimulus value (Figure 2.7A). Transient short-term plasticity between the recurrent connections, with a 4s-decay constant, was implemented as described by Mongillo and colleagues (2008; Figure 2.7B). Timing of the simulated events was comparable to the experimental paradigm: A target signal was briefly presented at a random location, followed by a mask signal to all neurons and a non-specific recall signal after a 3s-delay.

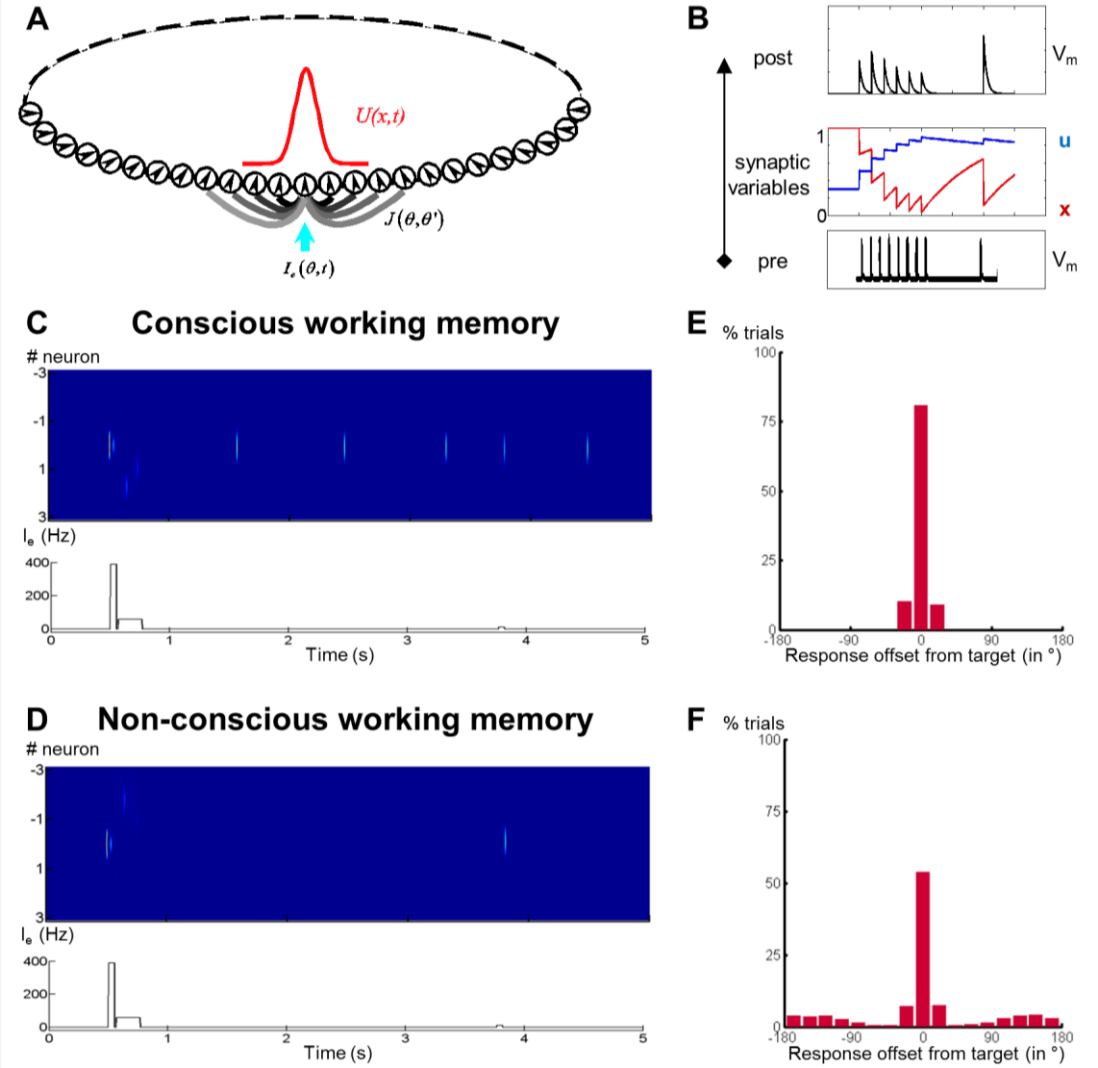


FIGURE 2.7

ACTIVITY-SILENT NEURAL MECHANISMS UNDERLYING CONSCIOUS AND NON-CONSCIOUS WORKING MEMORY.

(A) Structure of a one-dimensional continuous attractor neural network (CANN). Neuronal connections $J(\vartheta, \vartheta')$ are translation-invariant in the space of the neurons' preferred stimulus values $(-\pi, \pi)$, allowing the network to hold a continuous family of stationary states (bumps). An external input $I_e(\vartheta, t)$ containing the stimulus information triggers a bump state (red curve) at the corresponding location in the network.
(B) Model of a synaptic connection with short-term potentiation. In response to a presynaptic spike train (bottom), the neurotransmitter release probability u increases and the fraction of available neurotransmitter x decreases (middle), representing synaptic facilitation and depression. Effective synaptic efficacy is proportional to ux (top).
(C) Firing rate of neurons (top) and sequence of events (bottom; target and mask signal) when simulating conscious working memory with A_{mask} = 50 Hz $< A_{critical}$.
(D) Same as in (C) for non-conscious working memory when A_{mask} = 65 Hz > $A_{critical}$.
(E, F) Performance of the network (distribution of responses) when mask amplitude was near the critical level, A_{mask} = 62 Hz ~$A_{critical}$, and noise had been added to the system. Out of 4000 trials, 2035 resulted in the conscious **(E)** and the remainder in the non-conscious regime **(F)**. In both cases, performance remained above chance with the responses concentrated around the initial target location.

If the activity-silent mechanism constituted a plausible neurophysiological correlate of conscious and non-conscious working memory, these simulations should capture our principal findings. A stimulus presented at threshold should entail one of two different maintenance regimes: a first distinguished by near-perfect recall with spontaneous reactivations of the memorized representation throughout the retention period (thus resembling the prolonged, yet fluctuating, "decodability" of seen target locations), and a second characterized by above-chance objective performance in the almost complete absence of delay activity (thereby portraying the time course of the circular-linear correlations for the unseen stimuli).

In a noiseless model, there indeed existed a critical value of mask amplitude, $A_{critical}$, which separated two distinct regimes: Just as was the case for our seen trials, when $A_{mask} < A_{critical}$, the neural assembly

coding for the target spontaneously reactivated during the delay (Figure 2.7C). However, when $A_{mask} > A_{critical}$, the system evolved into a state without spontaneous activation of target-specific neurons, yet with a reactivation in response to a non-specific recall signal, mimicking our unseen trials (Figure 2.7D). When fixing mask amplitude near $A_{critical}$ and adding noise continuously or just to the inputs, the network exhibited both types of regimes in nearly equal proportions: 50.8% of trials were characterized by an activity-silent delay interspersed with spontaneous reactivations and 49.2% by an entirely activity-silent delay period. Reminiscent of our behavioral results, sorting the trials according to the existence or absence of these reactivations and computing the histograms of recalled target position relative to true location produced two distributions of objective working memory performance: one, in which target position was nearly accurately stored (Figure 2.7E), and one, in which performance remained above chance despite a higher base rate of errors (Figure 2.7F). These simulations replicate our experimental findings (in particular Figure 2.2 and Figure 2.5) and suggest the activity-silent framework as a likely candidate mechanism for both conscious and non-conscious working memory.

2.4 Discussion

Conscious perception and working memory are thought to be intimately related, yet recent evidence challenged this assumption by proposing the existence of non-conscious working memory (Soto et al., 2011). The present results may reconcile these views. Both conscious perception and conscious working memory shared similar signatures, including an alpha/beta power decrease, the latter spanning the entire delay on working-memory trials. However, participants remained able to localize a subjectively invisible target after a 4s-delay. We found no evidence that this long-lasting blindsight could simply be explained by erroneous visibility reports or by the conscious maintenance of an early guess. It thus likely reflects genuine non-conscious working memory. Despite the inherent differences in subjective experience for conscious and non-conscious working memory, a single, activity-silent mechanism might support both conscious and non-conscious information maintenance. We now discuss these points in turn.

2.4.1 Shared brain signatures underlie conscious perception and conscious working memory

Consistent with introspective reports and research on visual awareness and working memory (Baddeley, 2003; Dehaene et al., 2014), we observed a close relationship between conscious perception and maintenance in conscious working memory. In both tasks, classifiers trained to separate seen and unseen trials resulted in thick diagonals up to ~1 s after target onset, even when generalizing from one task to the other. Such long diagonals have repeatedly been observed in recent studies and are thought to reflect sequential processing (King and Dehaene, 2014; Marti et al., 2015; Salti et al., 2015; Stokes et al., 2015; Wolff et al., 2015). Irrespective of context, conscious perception and early parts of conscious maintenance thus involve a similar series of partially overlapping processing stages.

Time-frequency decompositions reinforced and extended this conclusion. Seen trials in the perception task were distinguished from both a target-absent control condition and unseen trials by a prominent decrease in alpha/beta power over fronto-central sensors, corresponding to a distributed network centered on parietal cortex. A similar desynchronization, sustained throughout the retention period, was also observed for conscious working memory. Alpha/beta band desynchronizations such as these have previously been linked with conscious perception (Gaillard et al., 2009; Wyart and Tallon-Baudry, 2009) and working memory (Lundqvist et al., 2016). Modelling suggests that the memorized item is encoded by intermittent gamma bursts, which interrupt an ongoing desynchronized beta default state (Lundqvist et al., 2011). Such a decreased rate of beta bursts, once averaged over many trials, would have resulted in the apparently sustained power decrease we observed. Increases in gamma power have also been shown in some studies on conscious perception (e.g., Gaillard et al., 2009), but we failed to detect it here, perhaps because our targets were brief, peripheral, and low in intensity.

Circular-linear correlations further highlighted the similarity between conscious perception and working memory. Location information could be tracked for ~1 s on perception-only trials and for at least 1.5 s of the working-memory retention period. The mental representation formed during conscious perception was therefore either maintained or repeatedly replayed during conscious working memory.

2.4.2 Long-lasting blindsight effect reflects genuine non-conscious working memory

Even when subjects indicated not having seen the target, they still identified its position much better than chance up to 4 s after its presentation. This long-lasting blindsight effect was replicated in two independent experiments and exhibited typical properties of working memory, withstanding salient visible distractors and a concurrent demand on conscious working memory. Those results corroborate previous research showing that information can be maintained non-consciously (e.g., Bergström and Eriksson, 2014; Bergström and Eriksson, 2015; Dutta et al., 2014; Soto et al., 2011). However, these prior findings could have arisen due to errors in visibility reports. If, for example, a participant had been left with a weak impression of the target (and, consequently, its location), he or she might not have had adequate internal evidence to refer to this perceptual state as seen, thus incorrectly applying the label unseen. A small number of such errors would have produced above-chance responding. Another explanation could have been the conscious maintenance of an early guess, whereby subjects would have ventured a prediction as to the correct target position immediately after its presentation and then consciously maintained this hunch.

The MEG results provide evidence against these possibilities. First, whereas seen trials were characterized by a sustained desynchronization in the alpha/beta band in parietal brain areas, no comparable desynchronization was observed on unseen trials, even when subjects correctly identified the target location. On the contrary, the only, short-lived, difference between unseen correct and unseen incorrect trials emerged around the time of the distractor and was reversed in direction: Unseen correct trials were accompanied by an increase in power in the alpha band with respect to their incorrect counterpart, an effect that might relate to a successful attempt to reduce interference from the distractor (Cooper et al., 2003; Jensen and Mazaheri, 2010). Otherwise, unseen correct and incorrect trials were indistinguishable in their power spectra and similar to the target-absent control condition. Second, there was no clear evidence for a shared discriminative decoding axis between the seen and the unseen correct trials: Generalization was entirely unsuccessful when the classifier was trained on the time-frequency data, and highly dissimilar from the original visibility decoder when trained on the ERFs. While it is impossible to draw definitive conclusions just from the current dataset and future research should replicate these results, the majority of our evidence thus points against an interpretation, in which the unseen correct trials constituted either just a subset of seen trials, or arose from the conscious maintenance of an early guess. Instead, inasmuch as the observed desynchronization serves as a faithful indicator of conscious processing, it argues in favor of a differential state of non-conscious working memory with a distinct neural signature.

Circular-linear correlations as well as multivariate regression models between the amplitude of the MEG signal and response location support this interpretation. On seen trials, response position was coded akin to target location: Initially maintained via slowly decaying neural activity in posterior brain areas, the response code subsequently resurfaced intermittently in the same as well as more frontal regions. There was no detectable evidence for such a code on the unseen trials. Only during the very last part of the delay, right before the response, did response-related neural activity emerge and ramp up to the same level as on seen trials during the response period. As such, the absence of any prior delay-period activity does not appear to be an artifact attributable to low statistical power or an increase in noise on the unseen trials. Instead, in conjunction with the absence of any signature of conscious processing on these trials, these

findings imply that subjects did not consciously maintain an early guess and rather relied on genuine non-conscious working memory to perform the task.

In this context, an interesting avenue for future investigations might be to delineate the boundary conditions of such non-conscious working memory. Although the short-term maintenance of information certainly lies at the heart of most theories of working memory (Eriksson et al., 2015), there exist additional criteria for working memory that were not investigated in the present study. It is thus an interesting empirical question whether these other working memory processes may also occur without subjective awareness. Is it, for example, possible to manipulate information non-consciously? Though speculative, in light of the proposed activity-silent code for non-conscious maintenance (without any spontaneous reactivations; see below), it seems unlikely. Being an entirely passive process, it is not clear how stored representations could be transformed without being persistently activated and thus becoming conscious. Future research is, however, needed to provide a definitive answer.

2.4.3 A theoretical framework for 'activity-silent' working memory

Target-related activity was not continuously sustained throughout the delay period, even when the target square had been consciously perceived. It instead fluctuated, disappearing and reappearing intermittently. This feature was even more pronounced on the unseen trials, with no evidence for any such retention-related activity beyond ~1 s. We presented a theoretical framework, based on Mongillo et al. (2008) and the concept of 'activity-silent' working memory (Stokes, 2015), that may provide a plausible explanation for maintenance without sustained neural activity. According to this model, short-term memories are retained by slowly decaying patterns of synaptic weights. A retrieval cue presented at the end of the delay may then serve as a non-specific read-out signal capable of reactivating these dormant representations above chance-level. Support for this model comes from experiments in which non-specific, task-irrelevant stimuli (Wolff et al., 2017, 2015), neutral post-cues (Sprague et al., 2016), or transcranial magnetic stimulation (TMS) pulses (Rose et al., 2016) presented during a delay restore the decodability of representations. Direct physiological evidence for the postulated short-term changes in synaptic efficacies also exists (Fujisawa et al., 2008).

The present non-conscious condition provides further support for such an activity-silent mechanism. In this framework, a stimulus that fails to cross the threshold for sustained activity and subjective visibility may still induce enough activity in high-level cortical circuits to trigger short-term synaptic changes. Such transient non-conscious propagation of activity has indeed been simulated in neural networks (Dehaene and Naccache, 2001) and measured experimentally in temporo-occipital, parietal, and even prefrontal cortices (van Gaal and Lamme, 2012; Salti et al., 2015). In the present work, we indeed observed some residual, transiently decodable activity over left occipito-temporal sensors on unseen correct trials. The memory of target location could therefore have arisen from posterior visual maps (Roelfsema, 2015), although future research should test this prediction further. Note that activity-silent mechanisms need not apply solely to prefrontal cortex as originally proposed by Mongillo et al. (2008), but constitute a generic mechanism that may be replicated in different areas, possibly with increasingly longer time constants across the cortical hierarchy (Chaudhuri et al., 2014). Only some of these areas/spatial maps may be storing the information on unseen trials.

A key feature of Mongillo et al.'s (2008) model and the present simulations is that, even for above-threshold ('seen') stimuli, delay activity is not continuously sustained. Occasional bouts of spontaneous reactivation instead refresh the synaptic weights and maintain the memory for an indefinite time. The time courses of the circular-linear correlations and of the multivariate decoding we observed on seen trials match this description: While target location was encoded and maintained in temporo-occipital areas, target "decodability" was not constantly sustained, but waxed and waned throughout the delay. Fuentemilla et al. (2010) also observed that, during a delay period, decodable representations of memorized images recurred at a theta rhythm. More recently, single-trial analyses of monkey

electrophysiological recordings in a working memory task have confirmed the absence of any continuous activity and instead identified the presence of discrete gamma bursts, paired with a decrease in beta-burst probability (Lundqvist et al., 2016). Such periodic refreshing of otherwise activity-silent representations could potentially serve as the neural correlate of conscious rehearsal, a central feature of working memory according to Baddeley (2003). It also suggests, however, that even consciously perceived items may not always be "in mind." Future research might attempt to more directly simulate activity-silent mechanisms in the context of conscious and non-conscious perception by, for example, relying on more elaborate models capturing decreases in alpha/beta power (Lundqvist et al., 2011).

In conjunction with prior evidence (King et al., 2016; Salti et al., 2015), our findings therefore indicate that there may be two successive mechanisms for the short-term maintenance of conscious and non-conscious stimuli: an initial, transient period of ~1 s, during which the representation is encoded by active firing with a slowly decaying amplitude, and an ensuing activity-silent maintenance via short-term changes in synaptic weights, during which activity either intermittently resurfaces (conscious case) or vanishes (non-conscious case). Such activity-silent retention need not necessarily be specific to working memory. Recent investigations have, for instance, demonstrated the existence of recognition memory for invisible cues (Chong et al., 2014; Rosenthal et al., 2016). As delay periods ranged in the order of minutes rather than seconds, persistent neural activity seems to be an unlikely candidate mechanism of maintenance. Activity-silent codes might have been at play, though they probably depended on mechanisms with longer time constants than the relatively rapidly decaying patterns of synaptic weights discussed in the context of the present experiments. Nevertheless, activity-silent representations may constitute a general mechanism for maintenance across the whole spectrum of temporal delays (from seconds over minutes/hours to days/weeks/decades), thus forming a generic property of memory.

2.4.4 Limitations and future perspectives

Our study presents limitations that should be addressed by future research. Due to the nature of the current investigation (a working memory task with long trials and subjectively determined variables), a relatively small number of unseen trials was acquired, thus making it difficult to detect subtle effects. While our conclusions are supported by Bayes' Factor analyses, converging evidence from univariate and multivariate techniques, and similar results obtained with larger samples in the domain of activity-silent conscious working memory (e.g., Rose et al., 2016; Wolff et al., 2017), a number of our observations are based on null effects, and it remains a possibility that we missed some target- and/or response-related activity on the unseen trials. Future research should thus aim at replicating the present findings with larger datasets or with more sensitive techniques, such as intracranial recordings. In particular, it might be interesting to further probe the relationship between seen, unseen correct, and unseen incorrect targets: A specific prediction of the proposed model is that unseen correct trials should possess enough activity to modify synaptic weights in high-level cortical circuits, yet without crossing the threshold for sustained activity and consciousness ("failed ignition"). Unseen correct trials should thus share some of the processes that are found on seen trials and future research is necessary to directly test this hypothesis.

2.4.5 Conclusion

In contrast to a widely held belief, our findings support the existence of genuine working memory in the absence of either conscious perception or sustained activity. Our proposal is that, following a transient encoding phase via active firing, non-conscious stimuli may be maintained by 'activity-silent' short-term changes in synaptic weights without any detectable neural activity, allowing above-chance retrieval for several seconds. Similar activity-silent codes also subserve conscious maintenance, though in this case periodic refreshing appears to stabilize the stored representations throughout the delay. Our findings thus highlight the need to refine our understanding of working memory, and to continuously challenge the limits of non-conscious processing.

2.5 METHODS

2.5.1 SUBJECTS

38 healthy volunteers participated in the present study (experiment 1: $N = 17$, $M_{age} = 23.3$ years, $SD_{age} = 2.8$ years, 10 men; experiment 2: $N = 21$, $M_{age} = 24.3$ years, $SD_{age} = 3.8$ years, 9 men). They gave written informed consent and received 80 or 15€ as compensation for the imaging and behavioral paradigms. Due to noisy recordings, only 13 of the 17 subjects in experiment 1 were retained for the MEG analyses. Although sample size had not specifically been estimated for our study, it thus was reasonable given typical experiments in the field.

2.5.2 EXPERIMENTAL PROTOCOL

Participants performed variations of a spatial delayed-response task, designed to assess retention of a target location under varying levels of subjective visibility (Figure 2.1A). Each trial began with the presentation of a central fixation cross (500 ms), displayed in white ink on an otherwise black screen. In experiment 1, a faint gray target square (RGB: 89.25 89.25 89.25) was flashed for 17 ms in 1 out of 20 equally spaced, invisible positions along a circle centered on fixation (radius = 200 pixels; 8 repetitions/location). Another fixation cross (17 ms) preceded the display of the mask (233 ms). Mask elements were composed of four individual squares (two right above and below, and two to the left and right of the target stimulus), arranged to tightly surround the target square without overlapping it. They appeared simultaneously at all possible target locations. Mask contrast was adjusted on an individual basis in a separate calibration procedure (see below). A variable delay period with constant fixation followed the mask (experiment 1: 2.5, 3.0, 3.5, or 4.0 s). On 50% of the trials in experiment 1, an unmasked distractor square, randomly placed and with the same duration as the target, was presented 1.5 s into the delay period.

After the delay, 20 letters – drawn from a subset of lower-case letters of the alphabet (excluded: *e*, *j*, *n*, *p*, *t*, *v*) – were randomly presented in the 20 positions (2.5 s). Participants were asked to identify the target location by speaking the name of the letter presented at the location. They were instructed to always provide a response, guessing if necessary. A trial ended with the presentation of the word *Vu?* (French for *seen*) in the center of the screen (2.5 s), cueing participants to rate the visibility of the target on the 4-point Perceptual Awareness Scale (PAS; *1*: no experience of the target, *2*: brief glimpse, *3*: almost clear experience, *4*: clear experience; Ramsøy and Overgaard, 2004) using the index, middle, ring, or little finger of their right hand (five-button non-magnetic response box, Cambridge Research Systems Ltd., Fiber Optic Response Pad). We instructed subjects to reserve a visibility rating of *1* for those trials, for which they had absolutely no perception of the target. The target square was also replaced by a blank screen on 20% of the trials, in order to obtain an objective measure of participants' sensitivity to the presence of the target. The inter-trial interval (ITI) lasted 1 s. Subjects completed a total of 200 trials of this working memory task, divided into four separate experimental blocks. They also undertook two blocks of 100 trials each of a perception-only control paradigm, identical to the working memory task in all respects except that the delay period and target localization screen were omitted, such that the presentation of the mask immediately preceded subjects' visibility ratings. Task order (perception vs. working memory) was counterbalanced across participants.

Experiment 2 was designed to investigate the impact of a conscious working memory load on non-conscious working memory. Apart from the following exceptions, it was identical to experiment 1: A screen with either 1 (low load) or 5 (high load) centrally presented digits (1.5 s) – randomly drawn (without replacement) from the numbers 1 through 9 – as well as a 1s-fixation period were shown prior to the presentation of the target square. Following either a 0s- or a 4s-delay period, subjects first identified the target location by typing their responses on a standard AZERTY keyboard (4 s). The French word for

numbers (*Numéros?*) then probed participants to recall the sequence of digits in the correct order. Responses were again logged on the keyboard during a period of 4.5 s. Subjects last rated target visibility as in experiment 1 (3 s). The ITI varied between 1 and 2 s. Participants completed two experimental blocks of 100 trials each.

2.5.3 Calibration task

Prior to the experimental tasks, each participant's perceptual threshold was estimated in order to ensure roughly equal proportions of seen and unseen trials. Subjects completed 150 (experiment 1: 3 blocks) or 125 (experiment 2: 5 blocks) trials of a modified version of the working memory task (no distractor, delay duration: 2 s in experiment 1 and 0 s in experiment 2), during which mask contrast was either increased (following a visibility rating of 2, 3, or 4) or decreased (following a visibility rating of 1) on each target-present trial according to a double-staircase procedure. Individual perceptual thresholds to be used in the main tasks were derived by averaging the mask contrasts from the last four switches from seen to unseen (or vice versa) in each staircase.

2.5.4 Behavioral analyses

We analyzed our behavioral data in Matlab R2014a (MathWorks Inc., Natick, MA) and SPSS Statistics Version 20.0 (IBM, Armonk, NY), using repeated-measures analyses of variance (ANOVAs). Only meaningful trials without missing responses were included in any analysis. Distributions of localization responses were computed for visibility categories with at least five trials per subject. Objective working memory performance was quantified via two complementary measures. The *rate of correct responding* was defined as the proportion of trials within two positions (i.e.,+/- 36°) of the actual target location and served as an index of the amount of information that could be retained. Because 5 out of 20 locations were counted as correct, chance on this measure was 25%. The *precision* of working memory was estimated as the dispersion (standard deviation) of spatial responses. In particular, we modeled the observed distribution of responses *D(n)* as a mixture of a uniform distribution (random guessing) and an unknown probability distribution *d* ("true working memory"):

$$(1)\ D(n) = \frac{p}{N} + (1-p)d(n)$$

where *p* refers to the probability that a given trial is responded to using random guessing; *N* to the number of target locations (*N* = 20); and *n* is the deviation from the true target location. We assumed that $d(n) = 0$ for deviations beyond a fixed limit *a* (with *a* = 2). This hypothesis allowed us to estimate *p* from the mean of that part of the distribution *D* for which one may safely assume no contribution of working memory:

$$(2)\ \hat{p} = \frac{\sum D(n) \mid n \text{ outside } [-a, a]}{(N-2a-1)} * N$$

where the model is designed in such a way as to ensure that $\hat{p} = 1$ if *D* is a uniform distribution (i.e., 100% of random guessing) and $\hat{p} = 0$ if *D* vanishes outside the region of correct responding (i.e., 0% of random guessing). There needs to be at least chance performance inside the region of correct responding, so

$$(3)\ \sum D(n) \mid n \in [-a, a] \geq \frac{2a-1}{N}$$

which ensures $0 \leq \hat{p} \leq 1$. This is the reason why, when computing precision, we included only subjects whose rate of correct responding for unseen trials, collapsed across all experimental conditions, significantly exceeded chance performance (i.e., 25%) in a χ^2-test ($p < .05$). An estimate of *d,* $\hat{d}$, can then be derived in two steps from Equation 1 as

$$(4)\ \delta(n) = \frac{D(n) - \frac{\hat{p}}{N}}{1 - \hat{p}}$$

(5) $\hat{d}(n) = \frac{\delta(n)|n\in[-a, a]}{\sum\delta(n)|n\in[-a, a]}$.

We note that the distribution δ has residual, yet negligible, positive and negative mass (due to noise) outside the region of correct responding. In order to obtain $\hat{d}$, we therefore restricted the distribution δ to [-*a*, *a*], set all negative values to 0, and renormalized its mass to 1. The precision of the representation of the target location in working memory was then defined as the standard deviation of that distribution.

2.5.5 MEG RECORDINGS AND PREPROCESSING

In experiment 1, we recorded MEG with a 306-channel (102 sensor triplets: 1 magnetometer and 2 orthogonal planar gradiometers), whole-head setup by ElektaNeuromag® (Helsinki, Finland) at 1000 Hz with a hardware bandpass filter between 0.1 and 330 Hz. Eye movements as well as heart rate were monitored with vertical and horizontal EOG and ECG channels. Prior to installation of the subject in the MEG chamber, we digitized three head landmarks (nasion and pre-auricular points), four head position indicator (HPI) coils placed over frontal and mastoïdian skull areas, and 60 additional locations outlining the participant's head with a 3-dimensional Fastrak system (Polhemus, USA). Head position was measured at the beginning of each run.

Our preprocessing pipeline followed Marti et al. (2015). Using MaxFilter Software (ElektaNeuromag®, Helsinki, Finland), raw MEG signals were first cleaned of head movements, bad channels, and magnetic interference originating from outside the MEG helmet (Taulu et al., 2004), and then downsampled to 250 Hz. We conducted all further preprocessing steps with the Fieldtrip toolbox (http://www.fieldtriptoolbox.org/; Oostenveld et al., 2011) run in a Matlab R2014a environment. Initially, MEG data were epoched between -0.5 and +2.5 s with respect to target onset for all stimulus-locked, and between -0.5 and +0.8 s with respect to the onset of the response screen for all response-locked analyses. Trials contaminated by muscle or other movement artifacts were then identified and rejected in a semi-automated procedure, for which the variance of the MEG signals across sensors served as an index of contamination. To remove any residual eye-movement and cardiac artifacts, we performed independent component analysis separately for each channel type, visually inspected the topographies and time courses of the first 30 components, and subtracted any contaminated component from the MEG data. Except for analyses requiring higher spatial precision (i.e., circular-linear correlations and decoding), results are presented for magnetometers only.

Further preprocessing steps depended on the nature of the subsequent analysis: Epochs retained for investigations based on evoked responses (i.e., ERFs, decoding, circular-linear correlations) were low-pass filtered at 30 Hz, while time-frequency decompositions relied on entirely unfiltered data. In the latter case, a sliding, frequency-independent Hann taper (window size: 500 ms, step size: 20 ms) was convolved with the unfiltered epochs in order to extract an estimate of power between 1 and 99 Hz (in 2 Hz steps) to identify the neural correlates of conscious and non-conscious perception and working memory in the frequency domain. Prior to univariate or multivariate statistical analysis, data (ERFs, time-frequency power estimates) were baseline corrected using a period between -200 and -50 ms.

2.5.6 CIRCULAR-LINEAR CORRELATIONS

To localize and track the neural representations of target, response, and distractor location, filtered epochs were transformed into circular-linear correlation coefficients. Following King et al. (2016), we combined the two linear correlation coefficients between the MEG signal and the sine and cosine of the angle defining the location in question (i.e., target, distractor, or response). An empirical null distribution was generated for each condition separately by shuffling the labels (i.e., target, distractor, or response location) at the corresponding time points and averaging the resulting distribution from 1000 such permutations.

Due to the spatial nature of our task, there is a possibility that subjects could have systematically moved their eyes after the presentation of the target, thus contaminating the correlation analyses. However, several lines of evidence suggest that this was not the case: First, participants were carefully instructed not to move their eyes. A close inspection of the EOG traces confirmed that subjects successfully implemented this request and did not display any strategic eye movements. Second, we carefully removed any trials contaminated by such movements as part of our preprocessing procedure. Third, the topographical patterns of the correlations show that the signal primarily originated in occipital and parietal channels. Eye movements therefore unlikely have driven the circular-linear correlations.

2.5.7 Sources

Individual anatomical magnetic resonance images (MRI), obtained with a 3D T1-weighted spoiled gradient recalled pulse sequence (voxel size: 1 * 1 * 1.1 mm; repetition time [TR]: 2,300 ms; echo time [TE]: 2.98 ms; field of view [FOV]: 256 * 240 * 176 mm; 160 slices) in a 3T Tim Trio Siemens scanner, were first segmented into gray/white matter as well as subcortical structures with FreeSurfer (https://surfer.nmr.mgh.harvard.edu/). We then reconstructed the cortical, scalp, and head surfaces in Brainstorm (http://neuroimage.usc.edu/brainstorm; Tadel et al., 2011) and co-registered these anatomical images with the MEG signals, using the HPI coils and the digitized head shape as a reference. Current density distributions on the cortical surface were subsequently estimated separately for each condition and subject. Specifically, we employed an analytical model with overlapping spheres to compute the leadfield matrix and modeled neuronal current sources with an unconstrained (dipole orientation loosening factor: 0.2) weighted minimum-norm current estimate (wMNE; depth-weighting factor: 0.5) and a noise covariance obtained from the baseline period of all trials. Average time-frequency power in the alpha (8–12 Hz) and beta (13–30 Hz) band was then estimated with complex Morlet wavelets using the Brainstorm default parameters, the resulting transformations projected onto the ICBM 152 anatomical template (Fonov et al., 2009, 2011), and the contrasts between the conditions of interest computed. Group averages for spatial clusters of at least 150 vertices are shown in dB relative to baseline and were thresholded at 60% of the maximum amplitude (cortex smoothed at 60%).

2.5.8 Multivariate pattern analyses

We employed the Scikit-Learn package (Pedregosa et al., 2011) as implemented in MNE 0.13 (Gramfort, 2013; Gramfort et al., 2014) in order to conduct our multivariate pattern analyses (MVPA). Following Marti et al. (2015) and King et al. (2016), we fit linear estimators at each time sample within each participant to isolate the topographical patterns best differentiating our experimental conditions. Support vector machines (Chang and Lin, 2011) were trained in the case of categorical data (i.e., visibility/accuracy) and a combination of two linear support vector regressions was used for circular data (i.e., target/response location) to estimate an angle from the arctangent of the separately predicted sine and cosine of the labels of interest.

A 5- (for categorical variables) or, due to the much larger number of labels, 2-fold (for circular variables), stratified cross-validation procedure was used in order to avoid overfitting: MEG data were first split into five (two) sets of trials with the same proportion of samples for each class. Within each fold, four (one) of these sets served as the training data and the remainder as the testing data. Model fitting, including all preprocessing steps, was exclusively performed on the training set. 50% of the most informative features (i.e., channels) were selected by means of a simple, univariate analysis of variance to reduce the dimensionality of the data (Charles et al., 2014; Haynes and Rees, 2006), the remaining channel-time features *z*-score normalized, and a weighting procedure applied in order to counteract the effects of any class imbalances. The classifier was then trained on the resulting data and applied to the left-out trials in order to identify the hyperplane (i.e., topography) best suited to separate the classes. This sequence of events (univariate feature selection, normalization, training and testing) was repeated five (two) times, ensuring that each trial would be included in the test set once.

Within the same cross-validation loop, we also evaluated the ability of each classifier to discriminate the experimental conditions of interest at all other time samples (i.e., generalization across time). This kind of MVPA results in a temporal generalization matrix, in which each entry represents the decoding performance of each classifier trained at time point *t* and tested at time point *t'*, and in which the diagonal corresponds to classifiers trained and tested on the same time points (King and Dehaene, 2014). Importantly, when interrogating the capacity of our classifiers to generalize across tasks or labels (e.g., from the perception to the working memory task, or from seen to unseen correct target locations), we modified the aforementioned cross-validation procedure to capitalize on the independence of our training and testing data. As such, classifiers from each training set were directly applied to the entire testing set and the respective predictions averaged.

Classifiers for categorical data generated a continuous output in the form of the distance between the respective sample and the separating hyperplane for each test trial. In order to be able to compare classification performance across subjects, we then applied a receiver operating characteristic analysis across trials within each participant and summarized overall effect sizes with the area under the curve (AUC). Unlike average decoding accuracy, the AUC serves as an unbiased measure of decoding performance as it represents the true-positive rate (e.g., a trial was correctly categorized as seen) as a function of the false-positive rate (e.g., a trial was incorrectly categorized as seen). Chance performance, corresponding to equal proportions of true and false positives, therefore leads to an AUC of 0.5. Any value greater than this critical level implies better-than-chance performance, with an AUC of 1 indicating a perfect prediction for any given class. In contrast, classifiers for circular data were first summarized by computing the mean absolute difference between the predicted and the actual angle (range: 0 to π; chance: π/2) and then transformed into an "accuracy" score (range: -π/2 to π/2; chance: 0). To facilitate comparability between different conditions, an additional baseline correction was then performed.

2.5.9 Statistical analyses

We performed statistical analyses across subjects. For the ERF and time-frequency data, cluster-based, non-parametric *t*-tests with Monte Carlo permutations were used to identify significant differences between experimental conditions (Maris and Oostenveld, 2007). Further planned comparisons of ERF time courses (seen vs. unseen) in a-priori defined spatio-temporal regions of interest (i.e., P3b time window: 300–600 ms) were conducted with non-parametric signed-rank tests ($p_{uncorrected} < .05$). A correction for multiple comparisons was then applied with a false discovery rate ($p_{FDR} < .05$).

Non-parametric signed-rank tests ($p_{uncorrected} < .05$) were also employed to evaluate decoding performance and the strength of circular-linear correlations. Specifically, we assessed whether classifiers could predict the trials' classes better than chance (categorical data: AUC > 0.5; circular data: rad > 0) and whether circular-linear correlation coefficients deviated from an empirical baseline (Δrho > 0). We report temporal averages over four a-priori time bins, corresponding to an early perceptual period (0.1–0.3 s), the P3b time window (0.3–0.6 s), and the first (0.6–1.55 s) and second (1.55–2.5 s) part of the delay period. To capitalize on the increased spatial selectivity of gradiometers, averaged time courses of these two channels are shown for circular-linear correlations.

Bayesian statistics, based on either two- (time-frequency analyses) or one-sided (circular-linear correlations) *t*-tests, were also computed when appropriate with a scale factor of $r = .707$ (Rouder et al., 2009).

2.5.10 Simulations

A one-dimensional, recurrent continuous attractor neural network (CANN) model (Mongillo et al., 2008) was adapted in order to simulate the experimental findings (Figure 2.7A). Individual neurons were aligned according to their preferred stimulus value, enabling the network to encode angular position of a

target stimulus (range: -π to π; periodic boundary condition). The dynamics of this system were determined by the synaptic currents of each neuron given by

$$(6)\ \tau\frac{\partial h_E(\theta, t)}{\partial t} = -h_\theta+\rho\int_{-\pi}^{\pi} J(\theta, \theta')u(\theta', t)x(\theta', t)R_E(\theta', t)d\theta' - J_{EI}R_I+I_b+\delta_1\xi_1(\theta, t)+I_e+\delta_2\xi_2(\theta, t),$$

$$(7)\ \frac{\partial u(\theta, t)}{\partial t} = \frac{U-u(\theta, t)}{\tau_f}+U[1-u(\theta, t)]R_E(\theta, t),$$

$$(8)\ \frac{\partial x(\theta ,t)}{\partial t} = \frac{1-x(\theta, t)}{\tau_d}-u(\theta, t)x(\theta, t)R_E(\theta, t), \text{ and}$$

$$(9)\ \tau\frac{\partial h_I}{\partial t} = -h_I+J_{IE}\int_{-\pi}^{\pi} R_E(\theta, t)\ ,$$

where τ describes the time constant of firing rate dynamics (in the order of milliseconds); ρ refers to neuronal density; $h_E(\vartheta, t)$ and $R_E(\vartheta, t)$ capture the synaptic current to and firing rate of neurons with preference θ at time t respectively; and $R(h) = \alpha \ln(1 + \exp(h/\alpha))$ is the neural gain chosen in the form of a smoothed threshold-linear function. J_{IE} and J_{EI} represent the connection strength between excitatory and inhibitory neurons. All excitatory neurons received a constant background input, I_e, reflecting the arousal signal when the neural system was engaged in a working memory task. $\delta_1\xi_1$ is background noise; I_e, any external stimulus (e.g., target, mask, and recall signal); and $\delta_1\xi_1(t)$ the noise related to those external stimuli. $u(\vartheta, t)$ and $x(\vartheta, t)$ denote the short-term synaptic facilitation (STF) and depression (STD) effects at time t of neurons with preference ϑ, respectively. The short-term plasticity dynamics are characterized by the following parameters: J_1 (absolute efficacy), U (increment of the release probability when a spike arrives), τ_f and τ_d (facilitation and depression time constants). The STF value $u(\vartheta, t)$ is facilitated whenever a spike arrives, and decays to the baseline U within the time τ_f. The neurotransmitter value $x(\vartheta, t)$ is utilized by each spike in proportion to $u(\vartheta, t)$ and then recovers to its baseline, 1, within the time τ_d.

$J(\vartheta, \vartheta')$ is the interaction strength from neurons at ϑ to neurons at ϑ' and is chosen to be

$$(10)(\theta, \theta') =) J\begin{cases} J_1\cos[B*(\theta-\theta')]-J_0, & \text{if } B*(\theta-\theta')\in[-\arccos(-J_0/J_1), \arccos(-J_0/J_1)], \\ -J_0, & \text{else} \end{cases}$$

where J_0, J_1, and B are constants which determine the connection strength between the neurons. Note that $J(\vartheta, \vartheta')$ is a function of $\vartheta - \vartheta'$, i. e., the neuronal interactions are translation-invariant in the space of neural preferred stimuli. The other parameters of the system were as follows: $\tau = 0.008$ s, $\tau_f = 4$ s, $\tau_d = 0.3$ s, $J_1 = 12$, $J_0 = 1$, $J_{EI} = 1.9$, $J_{IE} = 1.8$, $I_b = -0.1$ Hz, $\delta_1 = 0.3$, $\delta_2 = 9$, $N = 100$, $\alpha = 1.5$, $B = 2.2$.

During our simulations, we first presented a target signal with an amplitude of $A_{target} = 390$ Hz at a random location (50 ms), waited for 17 ms, and then applied a mask signal to all the neurons in the system (200 ms). The amplitude of the mask signal was initially varied in order to determine a critical value which would produce two distinct maintenance patterns, but was then fixed at a threshold of $A_{mask} = 62$ Hz. At the end of a 3s-delay period, a non-specific recall signal was given for 50 ms with $A_{recall} = 10$ Hz. Remembered target position was calculated as the population vector angle during this time period.

2.6 Acknowledgements

We gratefully acknowledge Henrik Ueberschär, Leila Azizi, and Virginie Van Wassenhove for their invaluable daily support and stimulating discussion.

2.7 Supplementary Figures

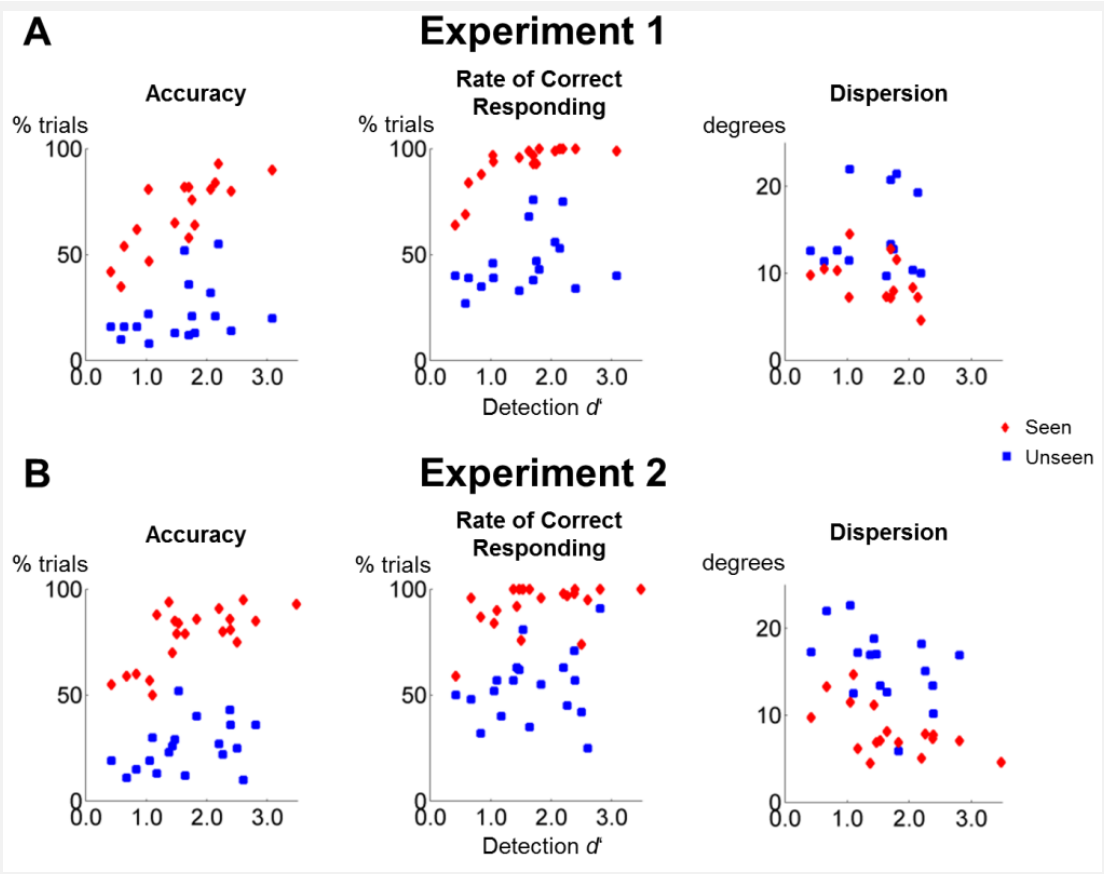


Figure 2.2 - Figure Supplement 1

Perceptual sensitivity does not correlate with working memory performance on unseen trials.

(A) Scatter plots depicting the relationship between detection *d'* and accuracy (left), the rate of correct responding (middle), and precision (right) in the working memory task of experiment 1 as a function of visibility.
(B) Same as in (A), but for experiment 2.

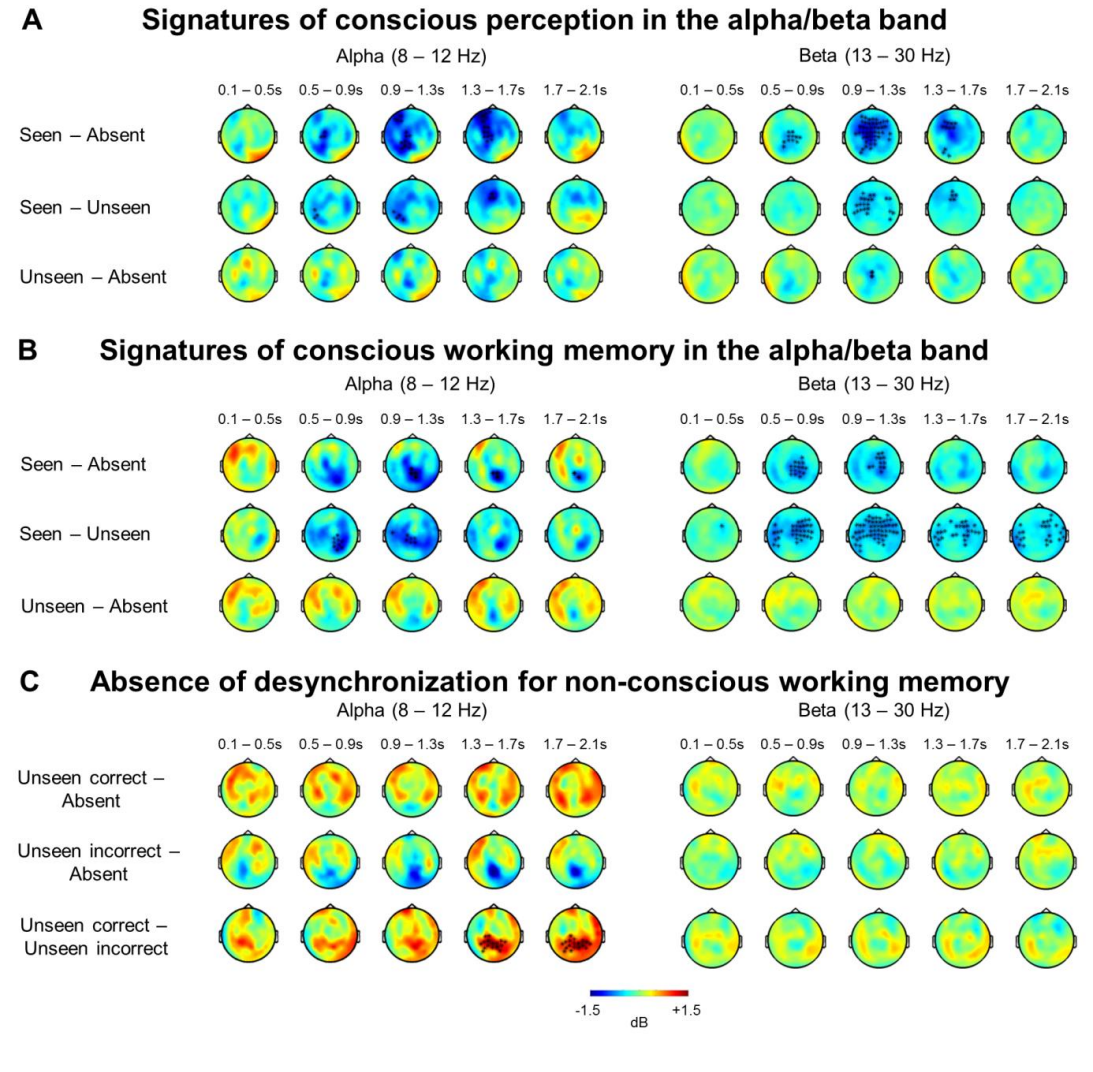


FIGURE 2.4 - FIGURE SUPPLEMENT 1

ALPHA- AND BETA-BAND DESYNCHRONIZATIONS SERVE AS A GENERAL SIGNATURE OF CONSCIOUS PROCESSING AND CONSCIOUS WORKING MEMORY.

(A) Perception task: Topographies represent the power difference (magnetometers) for seen vs target-absent trials (top), seen vs unseen trials (middle), and unseen vs target-absent trials (bottom) in the alpha (8–12 Hz) and beta (13–30 Hz) frequency bands as a function of time (0–2.1 s). Black asterisks indicate sensors showing a significant difference as assessed by a cluster-based permutation test.
(B) Working memory task: Topographies and panels are as in (A).
(C) Working memory task: Topographies represent the power difference (magnetometers) for unseen correct vs target-absent trials (top), unseen incorrect vs target-absent trials (middle), and unseen correct vs unseen incorrect trials (bottom) in the alpha (8–12 Hz) and beta (13–30 Hz) frequency bands as a function of time (0–2.1 s). Black asterisks indicate sensors showing a significant difference as assessed by a cluster-based permutation test.

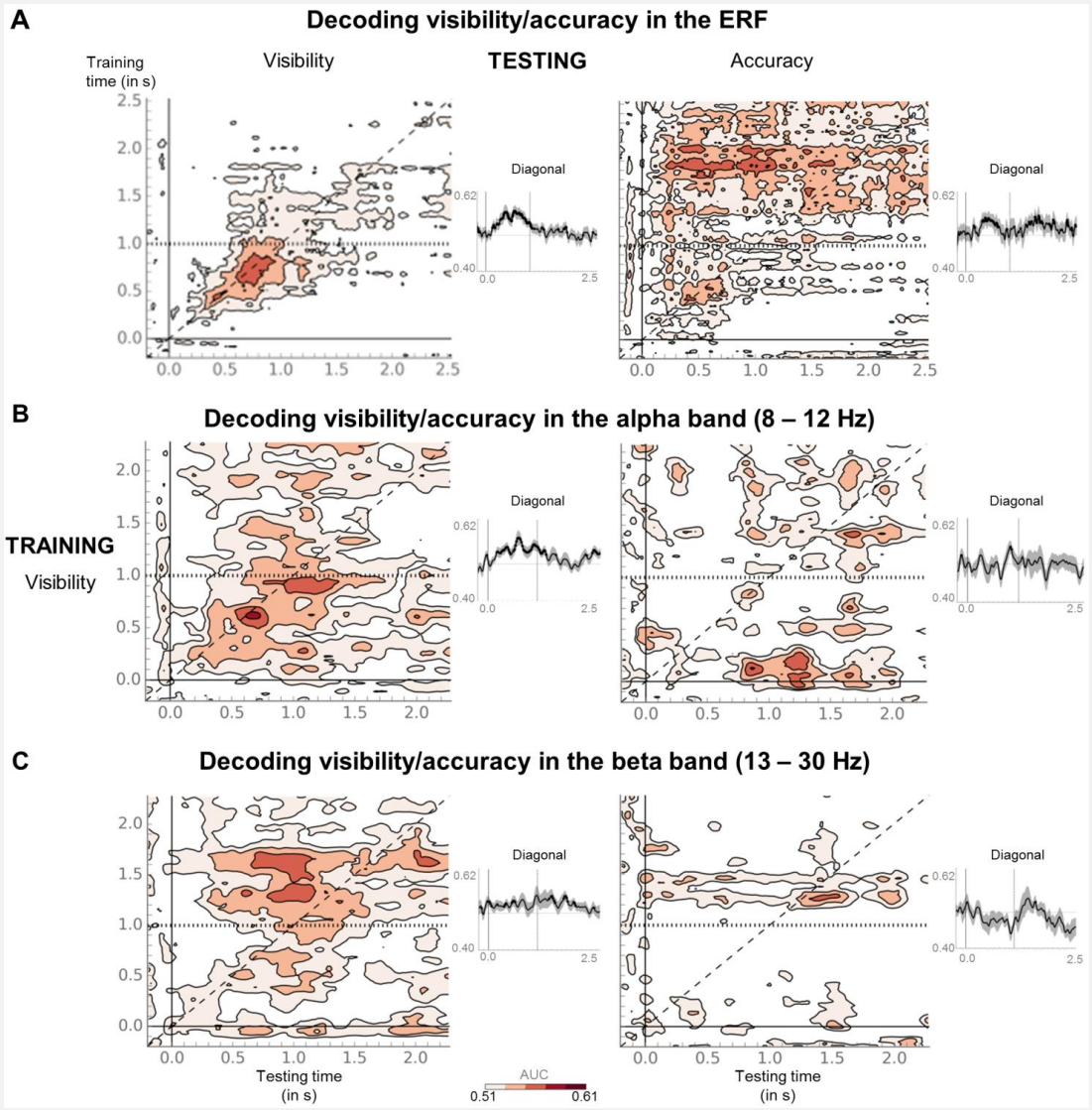


FIGURE 2.4 - FIGURE SUPPLEMENT 2

SEEN AND UNSEEN CORRECT TRIALS DO NOT SHARE THE SAME DISCRIMINATIVE DECODING AXIS.

(A) Temporal generalization matrices for a decoder trained on the ERFs to distinguish seen from unseen trials in the perception task and tested in the working memory task, either with the same labels (visibility decoder; left) or the unseen correct and incorrect trials (accuracy decoder; right). In each panel, a classifier was trained at every time sample (y-axis) and tested on all other time points (x-axis). The diagonal gray line demarks classifiers trained and tested on the same time sample. Please note the additional event marker: Mean reaction time (target-present trials) for the visibility response is indicated as a horizontal, dotted line. Any classifier beyond this point only reflects post-visibility processes. Time courses of diagonal decoding are shown as black insets. Thick lines indicate significant, above-chance decoding (Wilcoxon signed-rank test across subjects, uncorrected, one-tailed). For display purposes, data were smoothed using a moving average with a window of eight samples. AUC = area under the curve.
(B) Same as in (A), except that the decoder was trained and tested on average power (relative to baseline) in the alpha band (8–12 Hz). For display purposes, data were smoothed using a moving average with a window of one sample.
(C) Same as in (B), except that the decoder was trained and tested on average power (relative to baseline) in the beta band (13–30 Hz).

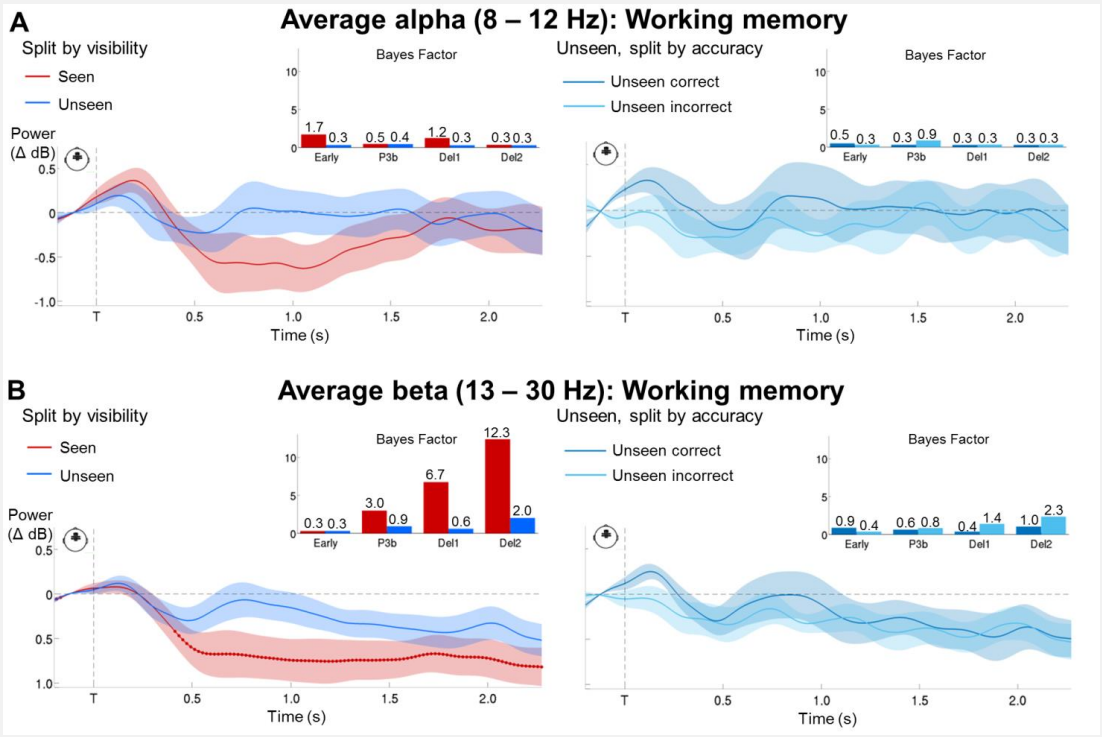


FIGURE 2.4 - FIGURE SUPPLEMENT 3

BAYESIAN STATISTICS FOR THE TIME-FREQUENCY ANALYSES.

(A) Time courses of average alpha band activity (8–12 Hz; -0.2 –2.1 s) in a group of frontal sensors as a function of visibility (left) and accuracy on the unseen trials (right; correct = within +/- 2 positions of the actual target location). Shaded area demarks standard error of the mean (SEM) across subjects. Insets show Bayes Factors (as assessed in a two-tailed *t*-test) in four time windows: 0.1–0.3 s (early), 0.3–0.6 s (P3b), 0.6–1.55 s (Del1), and 1.55–2.1 s (Del2). Del1 = first part of the delay, Del2 = second part of the delay, T = target onset.
(B) Same as in (A), but for average beta band (13–30 Hz) activity.

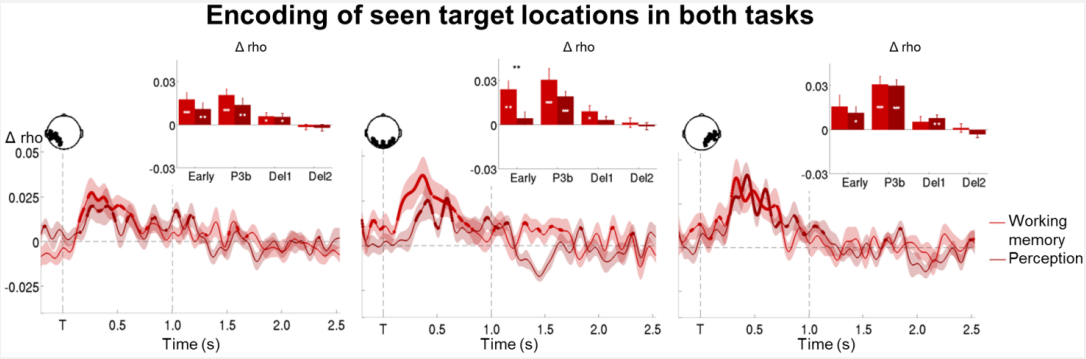


FIGURE 2.5 - FIGURE SUPPLEMENT 1

REPRESENTATION OF SEEN TARGET LOCATIONS DURING CONSCIOUS PERCEPTION AND WORKING MEMORY.

Average time courses of circular-linear correlation coefficients between amplitude of the ERFs and target location on seen trials as a function of task (perception and working memory) in a group of left temporo-occipital (left), occipital (middle), and right temporo-occipital (right) gradiometers. Shaded area demarks standard error of the mean (SEM) across subjects. Mean reaction time (target-present trials) for the visibility response in the perception task is indicated as a vertical, dotted line. Thick line represents significant increase in correlation coefficient as compared to an empirical baseline (one-tailed Wilcoxon signed-rank test across subjects, uncorrected). Insets show average correlation coefficients (relative to baseline) over four time windows: 0.1–0.3 s (early), 0.3–0.6 s (P3b), 0.6–1.55 s (Del1), and 1.55–2.5 s (Del2). White asterisks denote significant differences to baseline (one-tailed Wilcoxon signed-rank test across subjects), black asterisks significant differences between conditions (two-tailed Wilcoxon signed-rank test across subjects). For display purposes, data were lowpass-filtered at 8 Hz. *p < .05, **p < .01, and ***p < .001. Del1 = first part of the delay period, Del2 = second part of the delay period, T = target onset.

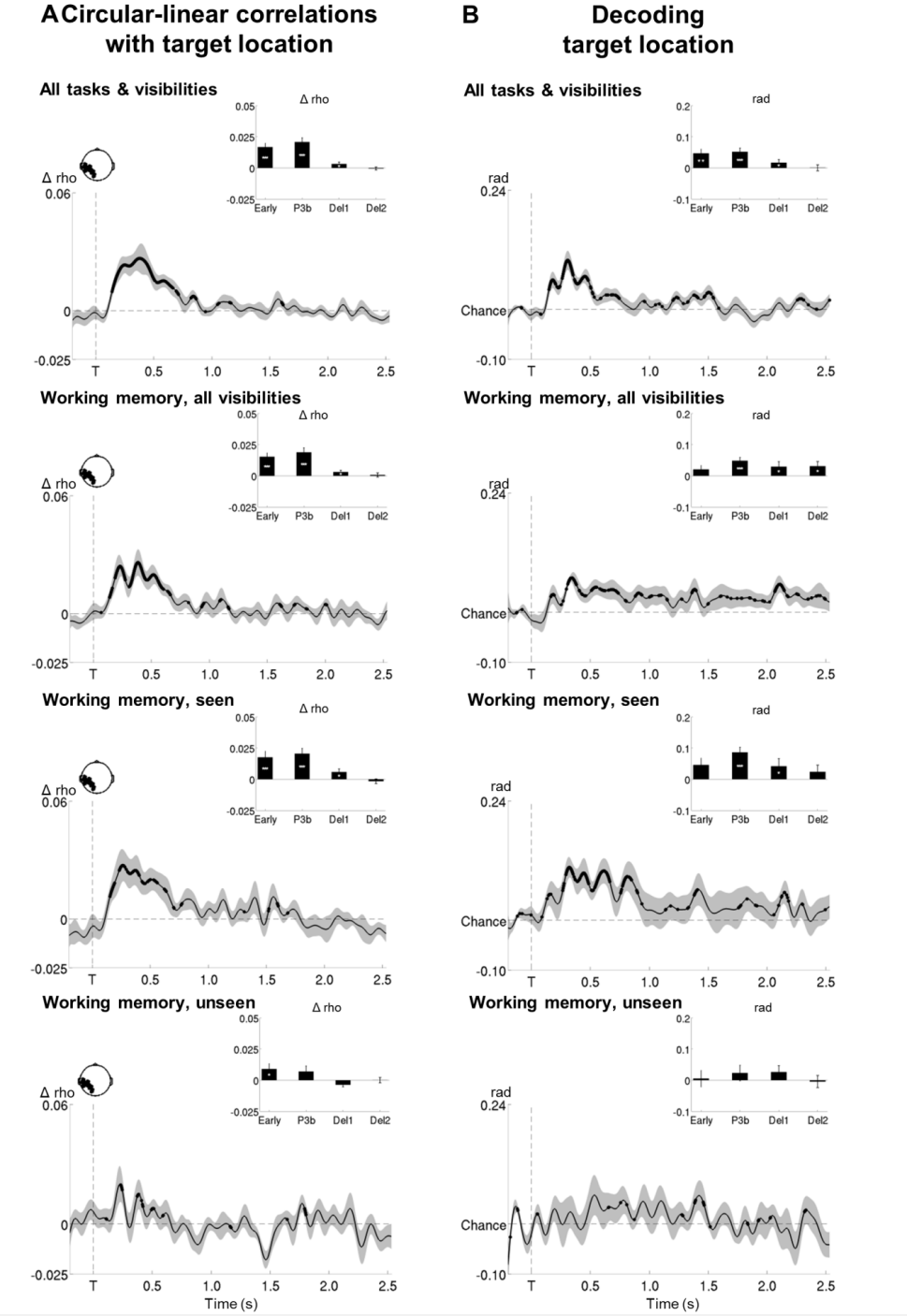


FIGURE 2.5 - FIGURE SUPPLEMENT 2

CIRCULAR-LINEAR CORRELATIONS AND MULTIVARIATE DECODING REVEAL SIMILAR TIME COURSES FOR TARGET LOCATION.

(A) Average time courses of circular-linear correlation coefficients between amplitude of the ERFs and target location as a function of task (perception and working memory) and visibility (seen and unseen) in a group of left temporo-occipital gradiometers. Shaded area demarks standard error of the mean (SEM) across subjects. Thick line represents significant increase in correlation coefficient as compared to an empirical baseline (one-tailed Wilcoxon signed-rank test across subjects, uncorrected). Insets show average correlation coefficients (relative to baseline) over four time windows: 0.1–0.3 s (early), 0.3–0.6 s (P3b), 0.6–1.55 s (Del1), and 1.55–2.5 s (Del2). White asterisks denote significant differences to baseline (one-tailed Wilcoxon signed-rank test across subjects). For display purposes, data were lowpass-filtered at 8 Hz. *$p < .05$, **$p < .01$, and ***$p < .001$. Del1 = first part of the delay period, Del2 = second part of the delay period, T = target onset.
(B) Average time courses of a linear support vector regression trained to predict target angle as a function of task (perception and working memory) and visibility (seen and unseen). Thick line represents significant increase in decoding accuracy (in radians) as compared to a baseline (one-tailed Wilcoxon signed-rank test across subjects, uncorrected). Insets show average correlation coefficients (relative to baseline) over four time windows: 0.1–0.3 s (early), 0.3–0.6 s (P3b), 0.6–1.55 s (Del1), and 1.55–2.5 s (Del2). White asterisks denote significant differences to baseline (one-tailed Wilcoxon signed-rank test across subjects). For display purposes, data were lowpass-filtered at 8Hz. *$p < .05$, **$p < .01$, and ***$p < .001$. Del1 = first part of the delay period, Del2 = second part of the delay period, T = target onset.

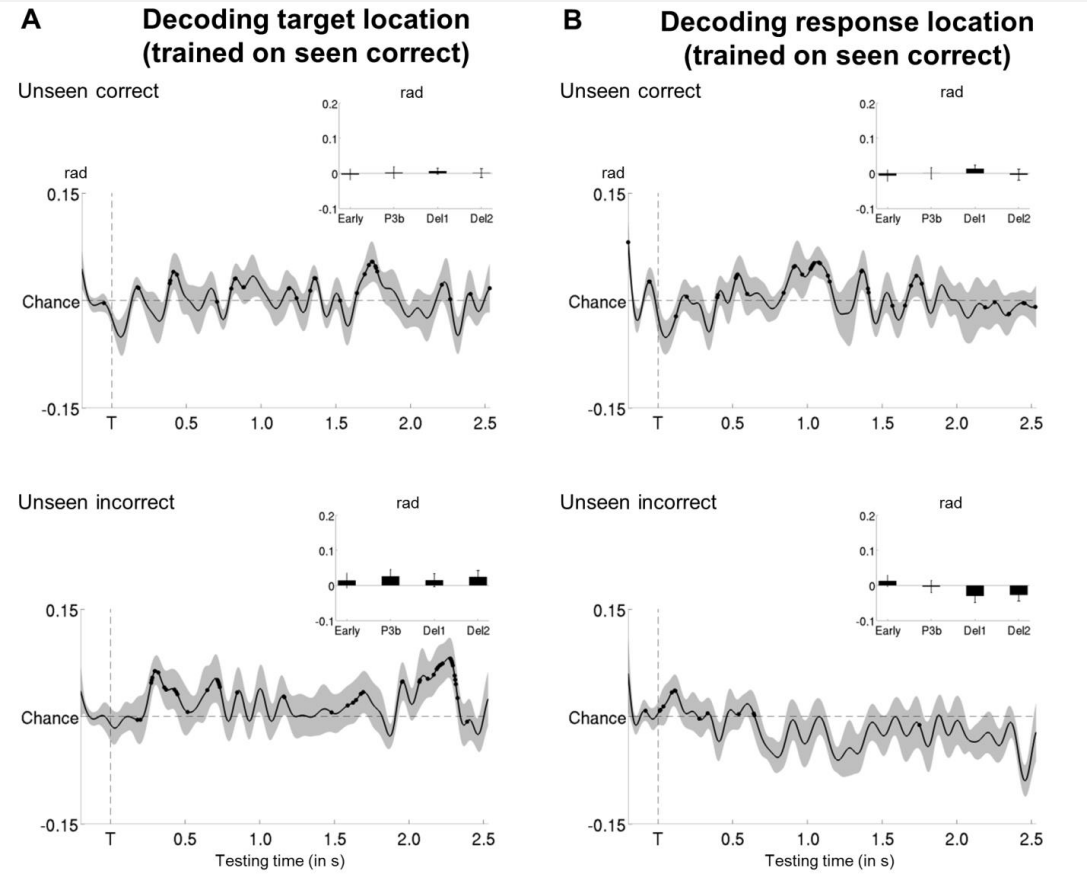


FIGURE 2.5 - FIGURE SUPPLEMENT 3

TRACKING TARGET/RESPONSE LOCATION ON UNSEEN CORRECT AND INCORRECT TRIALS WITH MULTIVARIATE DECODING.

(A) Average time courses of a linear support vector regression trained on seen correct trials to predict target angle on the unseen correct (top) and unseen incorrect (bottom) trials. Thick line represents significant increase in decoding accuracy (in radians) as compared to a baseline (one-tailed Wilcoxon signed-rank test across subjects, uncorrected). Insets show average correlation coefficients (relative to baseline) over four time windows: 0.1–0.3 s (early), 0.3–0.6 s (P3b), 0.6–1.55 s (Del1), and 1.55–2.5 s (Del2). White asterisks denote significant differences to baseline (one-tailed Wilcoxon signed-rank test across subjects). For display purposes, data were lowpass-filtered at 8Hz. *p < .05, **p < .01, and ***p < .001. Del1 = first part of the delay period, Del2 = second part of the delay period, T = target onset.
(B) Same as in (A), but for response location.

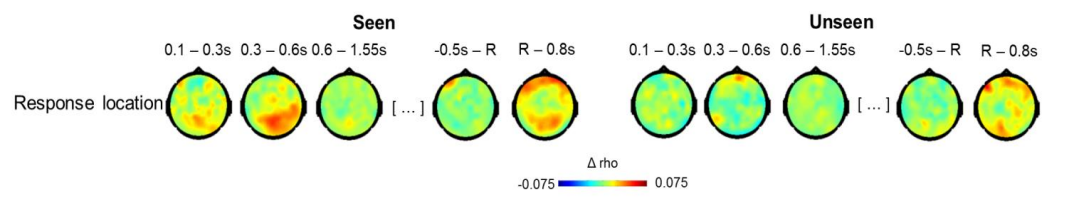


FIGURE 2.6 - FIGURE SUPPLEMENT 1

TOPOGRAPHIES FOR CIRCULAR-LINEAR CORRELATIONS WITH RESPONSE LOCATION AS A FUNCTION OF VISIBILITY.

Topographies of circular-linear correlations with response location as a function of time for seen (left) and unseen (right) trials. The first three time bins are relative to target, the last two relative to response screen onset. R = response screen onset.

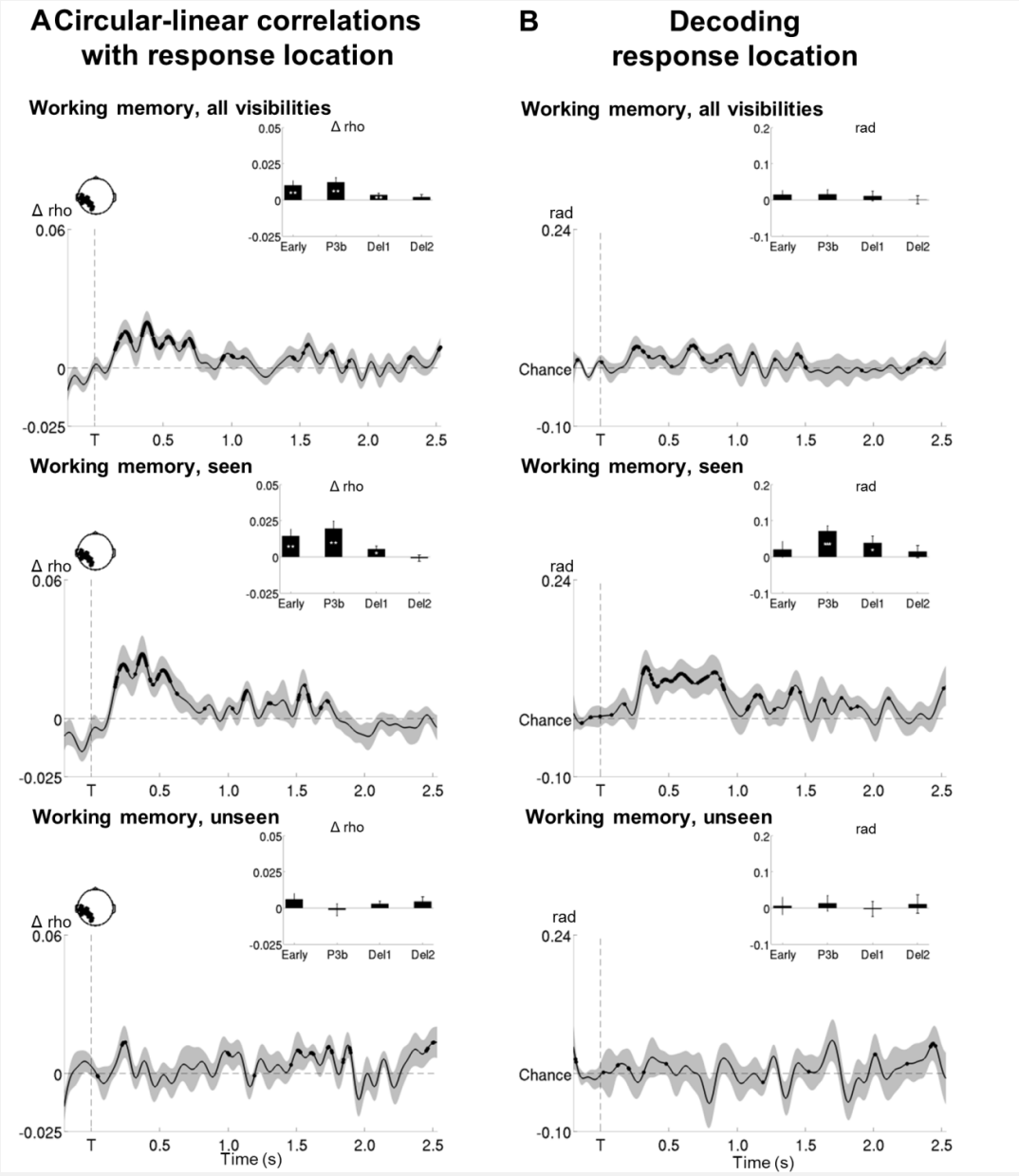


FIGURE 2.6 - FIGURE SUPPLEMENT 2

CIRCULAR-LINEAR CORRELATIONS AND MULTIVARIATE DECODING REVEAL SIMILAR TIME COURSES FOR RESPONSE LOCATION.

(A) Average time courses of circular-linear correlation coefficients between amplitude of the ERFs and response location as a function of task (perception and working memory) and visibility (seen and unseen) in a group of left temporo-occipital gradiometers. Shaded area demarks standard error of the mean (SEM) across subjects. Thick line represents significant increase in correlation coefficient as compared to an empirical baseline (one-tailed Wilcoxon signed-rank test across subjects, uncorrected). Insets show average correlation coefficients (relative to baseline) over four time windows: 0.1–0.3 s (early), 0.3–0.6 s (P3b), 0.6–1.55 s (Del1), and 1.55–2.5 s (Del2). White asterisks denote significant differences to baseline (one-tailed Wilcoxon signed-rank test across subjects). For display purposes, data were lowpass-filtered at 8 Hz. *p < .05, **p < .01, and ***p < .001. Del1 = first part of the delay period, Del2 = second part of the delay period, T = target onset.
(B) Average time courses of a linear support vector regression trained to predict response angle as a function of task (perception and working memory) and visibility (seen and unseen). Thick line represents significant increase in decoding accuracy (in radians) as compared to a baseline (one-tailed Wilcoxon signed-rank test across subjects, uncorrected). Insets show average correlation coefficients (relative to baseline) over four time windows: 0.1–0.3 s (early), 0.3–0.6 s (P3b), 0.6–1.55 s (Del1), and 1.55–2.5 s (Del2). White asterisks denote significant differences to baseline (one-tailed Wilcoxon signed-rank test across subjects). For display purposes, data were lowpass-filtered at 8 Hz. *p < .05, **p < .01, and ***p < .001. Del1 = first part of the delay period, Del2 = second part of the delay period, T = target onset.

CHAPTER 3 –
TEMPORAL-ORDER INFORMATION CAN BE MAINTAINED IN NON-CONSCIOUS WORKING MEMORY

The good thing about science is that it's true
whether or not you believe in it.
- NEIL DEGRASSE TYSON

3.1 ABSTRACT

Classical theories hold conscious perception and working memory to be tightly interwoven. Recent work has challenged this assumption, demonstrating that information may be stored for several seconds without any subjective awareness. Does such non-conscious working memory possess the same functional properties as regular conscious working memory? Here, we probe whether non-conscious working memory can maintain multiple items and their temporal order. In a visual masking task with a delayed response, participants were asked to retain the location and order of presentation of two sequentially flashed spatial positions. Even when they had not seen any of the targets, subjects' objective performance exceeded chance after several seconds. Crucially, participants did not commit swapping errors, first reporting the location of the second target and then the location of the first. Non-conscious working memory may therefore store two items in proper temporal order. These findings are compatible with recent proposals of activity-silent storage in non-conscious working memory.

3.2 INTRODUCTION

Until recently, conscious perception and working memory were thought to be inextricably linked, both enabling the short-term maintenance of information (Baars and Franklin, 2003; Baddeley, 2000, 2003) and relying on elevated, sustained neural activity (e.g., Dehaene et al., 2014; Funahashi et al., 1989; Fuster and Alexander, 1971; Lamme and Roelfsema, 2000). Empirical evidence has challenged these prevailing views. Masked items, that participants decline to have seen consciously, may be encoded into and maintained in working memory (Bergström and Eriksson, 2014; Bergström and Eriksson, 2015; Soto et al., 2011), and working memory itself may operate without subjective awareness (Bona et al., 2013; Hassin et al., 2009). Using magnetoencephalography, we have recently shown the existence of a genuine form of non-conscious working memory storing information in "activity-silent" brain states (Trübutschek et al., 2017), presumably via short-term synaptic plasticity (Mongillo et al., 2008). However, such non-conscious working memory is a recent discovery, and we still know very little about its functional properties. Is it a fully-fledged working memory system, with characteristics similar to the ones of conscious working memory? Or should it be conceived of as a restricted special-purpose system, with limited functionality?

Here, we chose to probe one of the most defining features of conscious working memory: the ability to store multiple items and related temporal-order information. Historically, research on conscious working memory has focused on the short-term maintenance of ordered information, including lists of digits, letters, or words (e.g., Baddeley, 1993; Brown et al., 2000; Burgess and Hitch, 1999; Henson, 1999). Most contemporary conceptualizations of working memory have evolved from these early findings. Any working memory system in the conscious sense should thus be able to accommodate the storage of multiple items in addition to their temporal order. However, whether this is also the case for non-conscious working memory is currently not clear.

Research on non-conscious working memory has almost exclusively investigated the storage of a single sensory item. Participants either had to remember the orientation of a Gabor patch (Bona et al., 2013; Soto et al., 2011), a spatial location (Trübutschek et al., 2017), or a number/letter (Bergström and Eriksson, 2014). To our knowledge, there have only been two studies to date, in which there were either two simultaneously presented memory stimuli (Soto et al., 2011), or in which the initial memorandum consisted of two simultaneously presented feature dimensions (i.e., object identity and its spatial location; Bergström and Eriksson, 2015). However, sample size was extremely small (i.e., $N = 9$) and the delay period fairly short (i.e., 2 s) in the first experiment and, in the latter case, automatic object-based attention may have integrated both features into a single object file, thus reducing the remembered stimulus to a single, bound representation (Bapat et al., 2017). There is thus very little, if any, empirical evidence that would be able to speak directly to the storage of multiple non-conscious items and no such information at all when considering temporal order.

There may also be theoretical reasons to believe that non-conscious working memory may not be able to accommodate the storage of temporal order. An emerging view is that, in contrast to conscious working memory, contents in non-conscious working memory may be maintained in activity-silent brain states (Silvanto, 2017; Soto and Silvanto, 2016; Trübutschek et al., 2017). According to this framework, temporary shifts in synaptic weights may effectively link populations of neurons, thereby allowing networks to go silent, while still leaving behind a transient synaptic memory trace of their previously active configuration for several seconds (Mongillo et al., 2008; Stokes, 2015). Although activity-silent mechanisms had initially been proposed as a specific property of prefrontal cortex (Mongillo et al., 2008; Stokes, 2015), empirically, they have exclusively been observed in posterior sensory and parietal regions of the brain (Christophel et al., 2018; Quentin et al., 2018; Rose et al., 2016; Trübutschek et al., 2017; Wolff et al., 2015, 2017). The maintenance of temporal order, in contrast, has been shown to primarily recruit higher-level regions of the cortex, including parietal, motor, and prefrontal cortices (Marshuetz and Smith, 2006; Roberts et al., 2017). It is thus not immediately evident if activity-silent brain states might account for the storage of temporal-order information in non-conscious working memory.

We here set out to address these key unknowns by confronting non-conscious working memory with multiple, independents items and their temporal order. Specifically, we aimed to determine (1) whether more than a single item may be maintained simultaneously in non-conscious working memory, and (2) if non-conscious storage includes order information. We were able to answer both of these questions in the affirmative: Two subjectively unseen target stimuli as well as their order could be retained for several seconds. As such, these results critically expand the realm of non-conscious working memory, further challenging predominant models of the nature of working memory.

3.3 Results

Participants completed a modified version of a spatial delayed-response paradigm, requiring the short-term maintenance of two sequentially presented spatial locations and their order of appearance (Figure 3.1). Each target was flashed in 1 of 20 possible positions, selected independently of each other with replacement, and immediately masked. After a 2.5 s delay period, subjects first localized both targets (by typing the random letter that appeared at that location), and then rated their subjective visibility for each of the two on a scale from *1* (not seen) to *4* (clearly seen). Critically, all responses were to be given in the order in which the targets had appeared, and irrespective of whether or not they had been consciously perceived. Participants were told to guess the locations of unseen squares. To enable the objective quantification of sensitivity to the presence of the masked targets, 20% of all trials served as a target-absent control condition: Either just the first, just the second, or both target stimuli were replaced by the presentation of a blank screen.

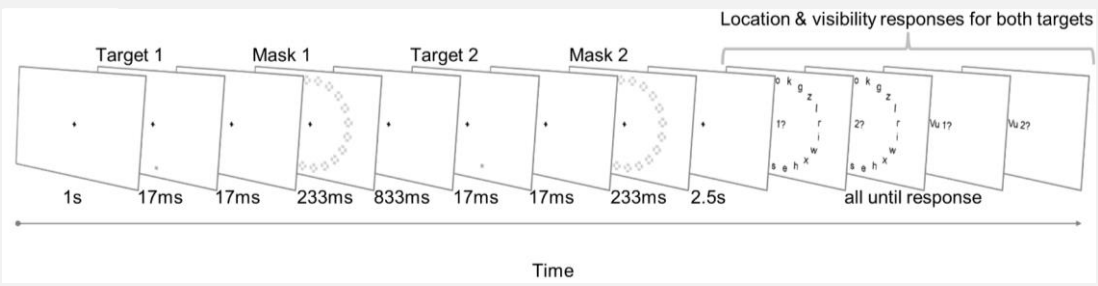


FIGURE 3.1

EXPERIMENTAL DESIGN.

Two individually masked target squares were flashed in 1 out of 20 positions. Each possible combination of angular distance between the two targets occurred once during the course of the experiment, such that, on a small subset of trials (i.e., 5%), successive targets appeared in the exact same spatial location. On 20% of trials, the presentation of either one or both targets was omitted and replaced by the display of a blank screen (target-absent control). Participants were instructed to perform two consecutive tasks after a long delay: First, they had to localize both targets in the order they had appeared. Then, they were to rate their visibility for each target on a 4-point scale. Critically, subjects had to complete both tasks, even when they had not seen the squares. In that case, they simply were to guess a position.

3.3.1 VISIBILITY RATINGS ACCURATELY REFLECT SUBJECTIVE PERCEPTION

We first evaluated our participants' ability to detect the masked squares independently for each target. Subjective visibility ratings for both targets varied as a function of target presence. Target 1 was reported as seen on the majority of target-present trials (visibility > *1*; 80.4 ± 15.0%), but was primarily rated as unseen on the target-absent catch trials (visibility = *1*; 76.6 ± 18.0%). Similarly, subjects indicated having seen a large proportion of target 2 when it was present (80.1 ± 16.5%), and judged most of the target-absent trials as unseen (81.5 ± 14.5%). Detection *d'* exceeded chance in both cases (target 1: 1.87 ± 0.80; $t(37) = 14.35$, $p < .001$; target 2: 2.04 ± 0.83; $t(37) = 15.10$, $p < .001$). Visibility reports nevertheless were not fully independent, as the two targets tended to be either both perceived or both unperceived (Supplement 3.1). Overall, then, participants used the visibility scale appropriately.

3.3.2 BOTH TARGETS CAN BE MAINTAINED NON-CONSCIOUSLY

As in our previous work (Trübutschek et al., 2017), subjects' localization responses for both targets were centered on the correct position (Figure 3.2). Accuracy was high on seen trials for target 1 (72.5 ± 15.4%) and increased monotonically with visibility, from rating *2* to rating *4* (pairwise comparisons: *t*s > 2.15, *p*s < .040). A similar pattern of findings also emerged for target 2. Overall

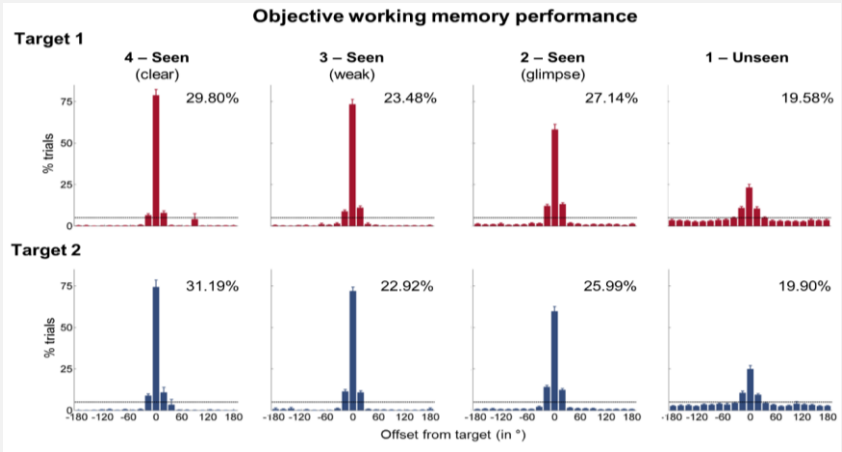


FIGURE 3.2

OBJECTIVE PERFORMANCE FOR BOTH TARGETS.

Spatial distributions of forced-choice localization performance in the working memory task are shown separately for target 1 (red) and target 2 (blue; 0 = correct target location; positive = counter-clockwise offset). Error bars indicate standard error of the mean (SEM) across subjects. The horizontal, dotted line illustrates chance-level at 5%. Percentages show proportion of all available trials, on which the target under consideration had been presented (i.e., combining both fully present as well as partially present trials) and participants had reported the corresponding visibility.

accuracy for seen targets was comparable, though a bit lower, than the one for the first target (70.7% ± 16.2%; $t(37) = 2.28$, $p = .028$). It again varied as a function of visibility (pairwise comparisons: ts > 4.78, ps < .001, with the exception of the contrast between visibility *3* and *4*, where $t(30) = 0.78$ and $p = .440$). Crucially, even when participants had not seen the target in question (rating = *1*), they still identified the correct position much better than chance (chance = 5%; target 1: 23.2 ± 12.1%; $t(37) = 9.25$, $p < .001$, 95% CI = [14.2%, 22.1%]; Cohen's d = 1.69; target 2: 24.9 ± 12.7%; $t(37) = 9.68$, $p < .001$, 95% CI = [15.8%, 24.1%]; Cohen's d = 1.78). This long-lasting blindsight effect was of similar magnitude for both targets ($t(37) = 1.06$, $p = .298$, Bayes' Factor = 0.29). We thus replicated and extended prior results (Bergström and Eriksson, 2015; Soto et al., 2011; Trübutschek et al., 2017), demonstrating that both targets could be retained in non-conscious working memory over a long delay.

3.3.3 Temporal order is maintained for seen and unseen targets

We next set out to evaluate objective performance for target 1 and target 2 as a function of joint visibilities. If indeed it were possible to retain information about temporal order in non-conscious working memory, location reports should remain accurate even if none of the targets had been seen. Furthermore, the size of these effects should be similar to when either only one target had been detected or only one had been presented.

To compare participants' localization across different conditions, we summarized objective working memory performance with two complementary measures (Trübutschek et al., 2017): (1) The rate of correct responding quantified the amount of information that could be stored in working memory and was defined as the proportion of trials within ± 2 positions (i.e., ± 36°) of the actual target location. (2) The precision of subjects' working memory representations was estimated as the standard deviation of that part of the original distribution reflecting genuine working memory (i.e., the spread within the zone of correct responding). It was only computed for those 32 of the 38 participants who exhibited above-chance blindsight for both target stimuli (i.e., chance = 25%; $p < .05$ in a χ^2-test).

Subjects' ability to identify the correct target location depended on visibility, but not on temporal order. Overall, participants retained more and more precise information when they had seen rather than when they had not seen the targets (rate of correct responding: seen = 93.9 ± 9.2%, unseen = 53.9 ± 15.4%, $F(1, 37) = 322.95$, $p < .001$; precision: seen = 9.1 ± 2.4°, unseen = 16.3° ± 3.1°, $F(1, 31) = 114.05$, $p < .001$). Ordinal position, by contrast, did not affect localization reports and there were no significant interactions, suggesting that subjects had maintained the spatial position equally well for the two targets (rate of correct responding: target 1 = 73.8 ± 11.7%, target 2 = 74.0 ± 10.5%, $F(1, 37) = 0.08$, $p = .780$; precision: target 1 = 12.6 ± 2.3°, target 2 = 12.9 ± 2.4°, $F(1, 31) = 0.40$, $p = .534$; interaction effects: both Fs < 0.78, both ps > .384).

Indeed, for the same visibility category (i.e., seen vs. unseen), the distributions of participants' localization responses looked virtually identical, irrespective of whether they pertained to the first or the second target (Figure 3.3A). Whenever subjects had detected both targets, they were near perfect in localizing both of them, with comparably high levels of correct responding and similar precision (Tables 3.1 and 3.2; Figure 3.3A, top left). This performance for seen targets was remarkably stable for different pairings of visibility and target presence (Figure 3.3A and Supplement 3.2A). We observed no systematic effects for any one condition to suggest different working memory performance in terms of storage or precision. Participants' localization responses for seen targets were thus highly reproducible, irrespective of their visibility for the other target.

The picture was even clearer for unseen target squares. Subjective experience of the other target did not modulate objective localization performance at all. When only one of the targets had been seen (Figure 3.3A) or when one of the targets had not been displayed (Supplement 3.2A), subjects still responded correctly to the unseen target much more frequently than predicted by chance (chance = 25%; all rates of

correct responding > 47.0%; all *ps* < .001, all Bayes' Factors > $1.30*10^7$). The size and precision of this blindsight effect was comparable across all conditions (rate of correct responding: *ts* < 2.21, *ps* > .510, Bayes' Factors < 1.56; precision: *ts* < 2.82, *ps* > .135, Bayes' Factors < 5.12).

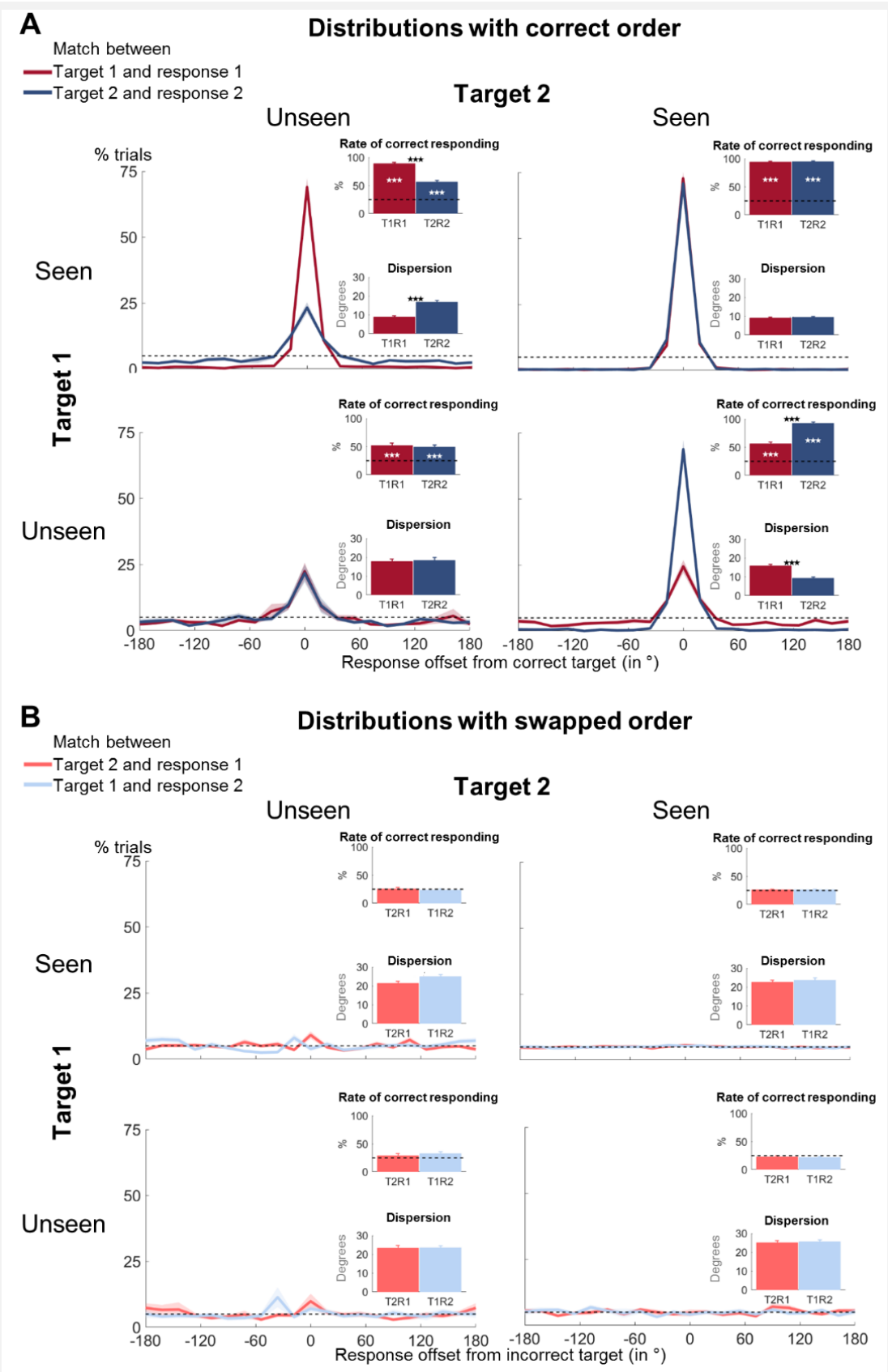


FIGURE 3.3

TEMPORAL ORDER CAN BE MAINTAINED IN NON-CONSCIOUS WORKING MEMORY.

(A) Spatial distributions of forced-choice localization performance in the working memory task on trials with two targets are shown as a function of visibility (i.e., seen vs. unseen) for target 1 (red) and target 2 (blue; 0 = correct target location; positive = counter-clockwise offset). Distributions reflect angular distances between target 1 and response 1 (T1R1) and target 2 and response 2 (T2R2). Insets show rate of correct responding (within ±2 positions of actual location) and precision of working memory representations separately for seen and unseen trials. Error bars indicate standard error of the mean (SEM) across subjects. The horizontal, dotted line illustrates chance-level at 5%. White asterisks show statistical significance when compared to chance (i.e., 25%) and black asterisks when comparing performance for target 1 with performance for target 2.
(B) Same conventions as in (A), except that distributions reflect angular distances between target 1 and response 2 (T1R2, light blue) and target 2 and response 1 (T2R1, light red). *p < .05, **p < .01, *** p < .001 in a paired-samples t-test (Bonferroni-corrected for 8 comparisons when comparing against chance, and for 4 comparisons when comparing performance for target 1 with performance for target 2).

Crucially, even when participants had missed both targets, they still exhibited long-lasting blindsight for both of them (rates of correct responding > 49.7%; ps < .001, Bayes' Factors > $1.29*10^{13}$; Figure 3.3A, bottom left). This effect was equally strong and precise for target 1 and target 2 (rate of correct responding: 52.5% vs. 49.7%; $t(36) = 0.70$, $p = .486$, Bayes' Factor = 0.22; precision: 17.7° vs. 18.5°; $t(30) = -0.55$, $p = .587$, Bayes' Factor = 0.22), and did not differ from the performance in any of the other conditions (rate of correct responding: ts < 2.01, ps > .052, Bayes' Factors < 1.08; precision: ts < 2.03, ps > .052, Bayes' Factors < 1.15). Just as was the case for conscious working memory, the non-conscious maintenance of a target stimulus was therefore unaffected by subjects' visibility of the other target. Moreover, on trials, in which none of the target squares had been seen, the characteristics of the blindsight effect were highly similar for target 1 and target 2, suggesting that it may have been possible for participants to retain more than one target as well as their temporal order non-consciously.

3.3.4 No evidence for swapping errors for seen and unseen targets

Although the existence of a blindsight effect for two unseen targets is a necessary prerequisite, by itself, it is not sufficient to determine whether non-conscious working memory can accommodate the storage of temporal order. It is, for instance, conceivable that, while subjects managed to maintain the order of unseen targets on a subset of trials, they reported an incorrect order on other trials. In our specific paradigm, this would translate into swapping errors, with participants choosing the location of the second target for their first response and vice versa.

To investigate this possibility, we examined the response distributions assuming that subjects had performed swapping errors; that is, we calculated the distance between target 1 and response 2, and between target 2 and response 1, as a function of visibility for both targets. The distributions for both seen and unseen targets were almost entirely flat (Figure 3.3B), indicating the absence of swapping errors. Whenever participants had detected at least one of the targets, there was no discernable above-chance performance (all rates of correct responding < 26.4%, all ps > .120, all Bayes' Factors < 3.31) or differences between conditions in terms of rate of correct responding or precision (rate of correct responding: ts < 2.98, ps > .075, Bayes' Factors < 7.47; precision: ts < 2.78, ps > .135, Bayes' Factors < 4.71). This pattern persisted when one of the targets had been absent (Tables 3 and 4; Supplement 3.2B).

Crucially, for the critical case of two unseen targets, we also obtained similar findings (Figure 3.3B, bottom left). Note that the small peaks one seems to notice within the region of correct responding for both distributions did not cross the threshold for statistical significance (rates of correct responding < 32.5%, ps > .120, Bayes' Factors < 3.27) and, importantly, disappeared when we excluded all trials, in which target 1 and target 2 had the same position (rates of correct responding < 30.4%, ps > .056, Bayes' Factors < 1.10). The "rate of correct responding" and "precision" were also significantly better when classified based on true than on swapped order (rate of correct responding: all ts > 4.12, all ps < .001, all Bayes' Factors > 125.39; precision: all ts > 3.12 , all ps < .004, all Bayes' Factors' > 9.52).

When directly contrasting distributions with the correct versus incorrect temporal order in a single repeated-measures analysis of variance (ANOVA), we observed the expected visibility (i.e., seen vs.

unseen) by type of distribution (i.e., correct vs. incorrect temporal order) interactions (rate of correct responding: $F(1, 37) = 344.86$, $p < .001$; precision: $F(1, 31) = 37.49$, $p < .001$). While visibility modulated responses based on true order, it did not affect the distributions based on swapped order. The distributions based on swapping errors displayed different features than the ones for genuine working memory. Subjects therefore committed very little, if any, temporal-order swapping errors, even on purely unseen trials when they reported not having seen any of the targets.

3.3.5 Long-lasting blindsight effects for both targets can occur on the same trial

We have established that (1) a comparable blindsight effect exists for target 1 and target 2, even when neither had been seen (Figure 3.3A), and that (2) the distributions reflecting genuine working memory of ordered information serve as a better predictor of behavior than do the swapped distributions (Figure 3.3B). Nevertheless, it could still be the case that participants accurately reported target 1 on some trials, and target 2 on other trials, but not both. We need to determine whether, at least on a subset of the trials with two unseen targets, subjects identified the correct target location for both targets on the very same trial. Only if this were the case, could we be sure that participants stored two locations as well as their order simultaneously.

As can be seen in Figure 3.4, this indeed turned out to be so. When restricting our trials to just that subset, on which target 1 had not been seen, yet localized correctly, we still observed above-chance performance for the second target (rate of correct responding: 57.8 ± 26.4%; $t(35) = 7.46$, $p < .001$, Bayes' Factor = $2.59*10^6$). A slightly less pronounced blindsight effect also emerged for target 2 following incorrectly localized first targets (rate of correct responding: 44.0 ± 23.7%; $t(35) = 4.83$, $p < .001$, Bayes' Factor = 1653.02; paired-samples t-test: $t(34) = 2.61$, $p = .013$, Bayes' Factor = 3.32). Both effects were comparable to the non-conscious working memory performance we had observed for target 1 on trials on which neither target had been detected (all ts < 1.65, all ps > .108, all Bayes' Factors < 0.61) or on which only the first had been presented (all ts < 1.06, all ps > .298, Bayes' Factors < 0.30). They also did not differ from the rate of correct responding for target 2, when the first target had been omitted (all ts < 1.37, all ps > .181, Bayes' Factors < 0.43). As such, there indeed existed a subset of trials, on which our participants had been able to non-consciously maintain two targets simultaneously. Non-conscious working memory may therefore accommodate the storage of two pieces of information in proper order.

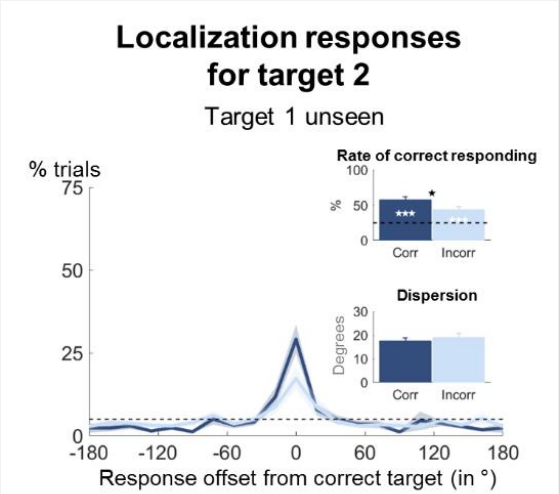


Figure 3.4

Long-lasting blindsight effect may occur simultaneously.

Spatial distributions of forced-choice localization performance for unseen target 2 as a function of whether the position for unseen target 1 had been identified correctly (dark blue) or not (light blue; 0 = correct target location; positive = counter-clockwise offset). Distributions reflect angular distances between target 2 and response 2. Insets show rate of correct responding (within ±2 positions of actual location) and precision of working memory representations separately for trials on which target 1 had been unseen correct and trials on which it had been unseen incorrect. Error bars indicate standard error of the mean (SEM) across subjects. The horizontal, dotted line illustrates chance-level at 5%. White asterisks show statistical significance when compared to chance (i.e., 25%) and black asterisks when comparing performance as a function of accuracy for target 1. *p < .05, **p < .01, *** p < .001 in a paired-samples t-test.

3.4 Discussion

Previous research on non-conscious working memory has almost exclusively focused on the maintenance of single, sensory items (Bergström and Eriksson, 2014; Bergström and Eriksson, 2015; Soto et al., 2011; Trübutschek et al., 2017). We here replicated and critically extended this earlier work: Subjects were able to not only store two unseen targets simultaneously, but also to retain their temporal order.

Consistent with prior findings, we observed a long-lasting blindsight effect for both targets. Even when participants had reported not having seen the target stimulus they identified its location much better than predicted by chance – up to ~ 4 s after its presentation. The magnitude and precision of this above-chance objective performance in the absence of subjective awareness remained constant throughout the entire experiment: It neither varied as a function of the number of targets presented (i.e., one vs. two), nor as a function of the ordinal position of the target (i.e., first vs. second). Crucially, it also persisted when neither of the two targets had been detected and, importantly, still occurred for the second target when the first target had been localized correctly. Subjects were thus clearly able to simultaneously retain two target locations in non-conscious working memory, at least on a subset of trials. On the flipside, they also committed virtually no swapping errors, almost exclusively reporting the location of the two targets in the proper order. In addition to maintaining the identity of multiple items, participants therefore also stored their temporal order. Taken together, these results suggest that the competencies of non-conscious working memory may reach much further than previously shown and, within the realm of features addressed in the present experiment, share important commonalities with conscious working memory.

Our work also raises further important questions. The first concerns capacity limits. Perhaps the most defining characteristic of working memory is its capacity-limited nature. In stark contrast to other forms of short-term memory, such as iconic (Sperling, 1960) or fragile memory (Pinto et al., 2013), only about ~4 to 7 items may concurrently be stored in working memory (Constantinidis and Klingberg, 2016). In the current experiment, we also demonstrated that both the amount of information as well as the precision with which this information could be retained in non-conscious working memory was largely unaffected by the number of items originally encoded into non-conscious working memory. That is, the long-lasting blindsight effect remained the same, irrespective of whether subjects missed the only target present (on partial target-absent trials), failed to detect one of the two targets, or did not see either one. Even the maximum amount encoded non-consciously appears to fall well within the capacity limits of conscious working memory. Another important future test might therefore consist in directly evaluating the capacity of non-conscious working memory. In light of the proposed activity-silent mechanism underlying maintenance in non-conscious working memory (Trübutschek et al., 2017), we speculate that, if any such limits do exist, they might be more closely related to the number and/or quality of the memory traces laid down during encoding or accessed during retrieval than to the quantity and/or precision of the stored representations themselves (as is assumed to be the case for conscious working memory).

This leads us to the second question. Which mechanism might have permitted participants to non-consciously maintain multiple items in addition to their temporal order? A first possibility is that subjects did not rely on non-conscious working memory at all, instead either accidentally miscategorizing some seen targets as unseen, or guessing the target positions immediately after their presentation and then consciously maintaining their ordered identity. In the context of the present experiment, we cannot fully reject this possibility. However, we deem it unlikely for several reasons. First, we carefully instructed subjects on the appropriate use of the visibility rating scale, stressing that a rating of *1* should be reserved exclusively for those trials, on which they thought the target to be absent. Second, having combined an almost identical paradigm with magnetoencephalography recordings, we have previously shown that non-conscious maintenance is genuine: Participants neither erroneously miscategorized their visibility ratings nor consciously maintained an early guess (Trübutschek et al., 2017). Indeed, when comparing the size of the blindsight effect (i.e., rate of correct responding) obtained in our previous study with the ones obtained

for target 1 and target 2 in the present experiment, we found no evidence for any differences (independent samples *t*-test: both *t*s < 1.64, both *p*s > .108, both Bayes' Factors < 0.85). Note that, based on an analysis of reaction time data, other groups have similarly argued in favor of genuine non-conscious working memory (Bergström and Eriksson, 2015). Taken together, this evidence supports the hypothesis of a long-lasting blindsight effect in the present experiment.

On the theoretical level, the non-conscious maintenance of multiple representations with temporal order is fully compatible with the view of activity-silent non-conscious working memory. Indeed, the original computational model of synaptic working memory already included simulations of the storage of multiple items (Mi et al., 2017; Mongillo et al., 2008). The key idea is straightforward: Individual memories are retained by item-specific patterns of synaptic facilitation. If more than a single item is to be stored, the neuronal networks coding for the individual contents reactivate consecutively in brief bursts of activity separated by long activity-silent periods, thereby enabling the short-term maintenance of several representations. Moreover, recent models of serial-order representations in working memory assume that prefrontal neurons may code conjunctively for item and order (Botvinick and Watanabe, 2007). Here, we propose that temporal order may automatically be retrieved: A basic assumption of the activity-silent framework is that the neural response of any network will be patterned according to previous input. Any non-specific signal, such as our recall cue, should reactivate the population coding for a specific item, thereby allowing downstream systems to read out the stored information. If the stored pattern itself contains information about item identity as well as temporal order, retrieving one piece of information would imply retrieving the other. Both multiple representations and their temporal order may thus be maintained in activity-silent brain states. Future research might directly test this proposal at the brain level.

3.4.1 Conclusion

Recently, there has been growing interest and evidence for the notion of a genuine non-conscious working memory. However, the precise nature of this phenomenon is still unclear. Our work critically expands our understanding of this long-lasting blindsight effect. Combining a masking paradigm with a spatial delayed-response task, we demonstrated that non-conscious working memory may accommodate the storage of multiple items as well as their temporal order. We further propose that these capacities are fully aligned with activity-silent mechanisms, believed to support maintenance in non-conscious working memory. As such, our results highlight the similarities between conscious and non-conscious working memory and continue to challenge current conceptualizations of working memory based on conscious processing and sustained neural activity.

3.5 Methods

3.5.1 Subjects

We recruited a total of 40 healthy volunteers (24 women; M_{age} = 24.85 years, SD_{age} = 4.20 years). All subjects had normal or corrected-to-normal vision, presented themselves without a history of neurological or psychiatric antecedents, and gave written informed consent prior to participation. They received €20 as compensation for their time and effort. Due to non-compliance with task instructions, we excluded 2 participants, resulting in a final dataset of 38 subjects.

3.5.2 Working memory task

Participants performed a variant of our masked, spatial-delayed response protocol (Trübutschek et al., 2017) to evaluate the short-term maintenance of sequences of memoranda as well as order information in conscious and non-conscious working memory (Figure 3.1). The experiment was programed and

presented using Psychtoolbox software (http://psychtoolbox.org/), run in a Matlab R2017 environment. Each trial began with a 1 s fixation period, followed by the presentation of the first target stimulus: A faint, gray square was briefly displayed in 1 out of 20 circular locations (17 ms). After a short inter-stimulus interval (ISI) of 17 ms, a visual mask, whose contrast had been calibrated on an individual basis to produce roughly equal proportions of seen and unseen targets (see below), appeared in all possible positions (233 ms), effectively camouflaging the target location. This sequence of events (i.e., target, ISI, mask) was then repeated a second time, separated from the initial one by an 833 ms delay. Importantly, we drew locations for both targets independently of each other (such that, on a small subset of trials, successive targets could appear in the same position), and ensured a fully counterbalanced design, with all possible dependencies between target 1 and target 2 occurring with equal probability. Target-absent catch trials, on which the presentation of the target square was replaced by a blank screen, were also included to allow for an objective quantification of our subjects' sensitivity to the target stimuli: While 4% of all trials contained no target square at all, an additional 16% omitted either just the first or the second target.

A given trial then terminated with two successive responses: Participants first identified the spatial locations of the two targets in the order they had appeared, and then rated their subjective visibility of both target squares on the 4-point Perceptual Awareness Scale (Ramsøy and Overgaard, 2004). Both types of responses were entered on a standard AZERTY keyboard. On each trial, a subset of lower-case letters of the alphabet (excluded: *b*, *c*, *j*, *n*, *p*, *t*) was randomly placed in the 20 positions, permitting subjects to simply type the letter corresponding to the location in question. The number pad keys were used to indicate visibility. Crucially, target localization was required on all trials and under all circumstances. Even when participants had not seen a given target square, we instructed them to choose a position, guessing it if necessary. Moreover, subjects were to only declare a target as unseen (i.e., visibility = *1*), if they had not perceived it at all; in case of the slightest doubt, they had to rate it as seen (i.e., visibility > *1*). The inter-trial interval (ITI) was jittered between 333 ms and 666 ms. Background color of the screen was set to black (RGB: 1, 1, 1), and all other stimuli, with the exception of the target and mask, were shown in white (RGB: 255, 255, 255). We constantly presented a central fixation cross in order to aid participants in orienting their gaze and attention onto the center of the screen throughout the entire experiment. Overall, subjects completed 500 trials of this task, split into 10 blocks of 50 trials each, and presented on a flat screen computer monitor (viewing distance ~ 60 cm) in a dimly lit testing cabinet.

3.5.3 Calibration task

Just before the main experimental task, participants also completed 100 trials of a separate calibration procedure, designed to estimate the mask contrast needed for roughly equal proportions of seen and unseen targets. Up until (and including) the presentation of the first mask, this calibration was strictly identical to the main working memory paradigm. It then, however, diverged, requiring an immediate rating of subjective visibility without the need to maintain multiple targets or the order of their presentation. Crucially, we applied a double-staircase technique to adjust mask contrast at the single-trial level. Whenever subjects had rated a target-present trial as unseen (visibility = *1*), mask contrast was reduced by $1/20^{th}$ on the subsequent trial. By contrast, it was increased by the same amount whenever participants had reported a target-present trial as seen (visibility > *1*). Initial values for the two staircases were set to RGB values of 12.75, 12.75, 12.75 and 242.5, 242.5, 242.5, respectively, and one of the two staircases was randomly selected at the beginning of each trial. In case of target-absent trials, the previous mask contrast from a randomly chosen staircase was re-used without having been updated. We then computed individual mask contrasts to be used in the main task by taking the grand average over the last four switches (i.e., from seen to unseen or vice versa) across the two staircases. The same contrast was applied to the two targets.

3.5.4 Data analyses and statistics

In analogy to our previous approach (Trübutschek et al., 2017), we summarized objective working memory performance with three complementary measures. For each subject, target, and condition of interest, we computed (1) the accuracy, (2) the rate of correct responding, and (3) the precision of forced-choice localization responses. Whereas the former two both capture the quantity of information that may be retained, the latter serves as an estimate of the quality of the underlying working memory representations. Details on how exactly to derive all of these indices have already been provided in our previous open-access publication (Trübutschek et al., 2017), so we will only focus on the main elements here.

Accuracy simply corresponds to that proportion of trials for which participants had identified the exact target location, leading to a chance level of 5% (i.e., 1/20). The rate of correct responding, in contrast, takes into account small errors in localization performance, having been defined as that proportion of trials within close spatial proximity (i.e., ± 2 positions) of the actual target location. Chance for this measure is 25% (i.e., 5/20). For participants displaying sufficient blindsight for both targets (i.e., $p < .05$ in a χ^2-test against chance, collapsed across all other conditions), we also estimated the precision of that part of the distribution within the zone of correct responding corresponding to genuine working memory, after having accounted for random guessing (see Trübutschek et al., 2017 for all details).

We submitted all of these indices to appropriate statistical tests, being either (1) one-sample *t*-tests (for comparisons against chance), (2) paired samples *t*-tests (for all comparisons requiring identification of which specific conditions might have differed), or (3) repeated-measures analyses of variance (ANOVAs; for comparisons aiming at identifying just any overall effect). The statistical threshold for significance was set to $p < .05$, and, in the case of multiple comparisons, a Bonferroni correction was applied. In addition, where appropriate, we also provide Bayes' Factors based on one one- or two-sided *t*-tests ($r = .707$; Rouder et al., 2009).

3.6 Acknowldegements

This work was funded by INSERM, CEA, Collège de France, ERC, and Fondation Roger de Spoelberch. D.T. was funded by a graduate fellowship from the Ecole des Neurosciences de Paris (ENP) and Fondation Schneider Electric.

3.7 Tables

Visibility	Target 1		Target 2		Paired-samples *t* test		
	M	*SD*	*M*	*SD*	*t*	*p*	BF
SeenSeen	94.2%	7.7%	94.9%	9.2%	-1.49	.145	0.48
SeenUnseen	89.6%	13.4%	56.5%	18.4%	9.97	**<.001**	$1.33*10^9$
UnseenSeen	56.1%	19.4%	93.2%	13.3%	-11.02	**<.001**	$2.58 * 10^{10}$
UnseenUnseen	52.5%	24.9%	49.7%	21.0%	0.70	.486	0.22
SeenAbsent	94.8%	13.7%	-	-	-	-	-
UnseenAbsent	51.1%	24.6%	-	-	-	-	-
AbsentSeen	-	-	93.9%	16.8%	-	-	-
AbsentUnseen	-	-	47.0%	32.4%	-	-	-

Table 3.1. Summary statistics for the rate of correct responding for target 1 and target 2 as a function of visibility for the two targets. Statistical comparison was done between the two targets. BF = Bayes' Factor.

Visibility	Target 1		Target 2		Paired-samples *t* test		
	M	*SD*	*M*	*SD*	*t*	*p*	BF
SeenSeen	9.0°	2.7°	9.6°	2.4°	-2.59	**.014**	3.25
SeenUnseen	8.9°	3.6°	16.9°	4.5°	-7.21	**<.001**	344,125.89
UnseenSeen	15.8°	4.8°	9.2°	3.8°	5.54	**<.001**	4,300.18
UnseenUnseen	17.7°	7.2°	18.5°	7.2°	-0.55	.587	0.22
SeenAbsent	7.5°	3.6°	-	-	-	-	-
UnseenAbsent	20.0°	8.0°	-	-	-	-	-
AbsentSeen	-	-	8.9°	5.2°	-	-	-
AbsentUnseen	-	-	17.8°	10.1°	-	-	-

Table 3.2. Summary statistics for the precision for target 1 and target 2
as a function of visibility for the two targets. Statistical comparison was done between the two targets. BF = Bayes' Factor.

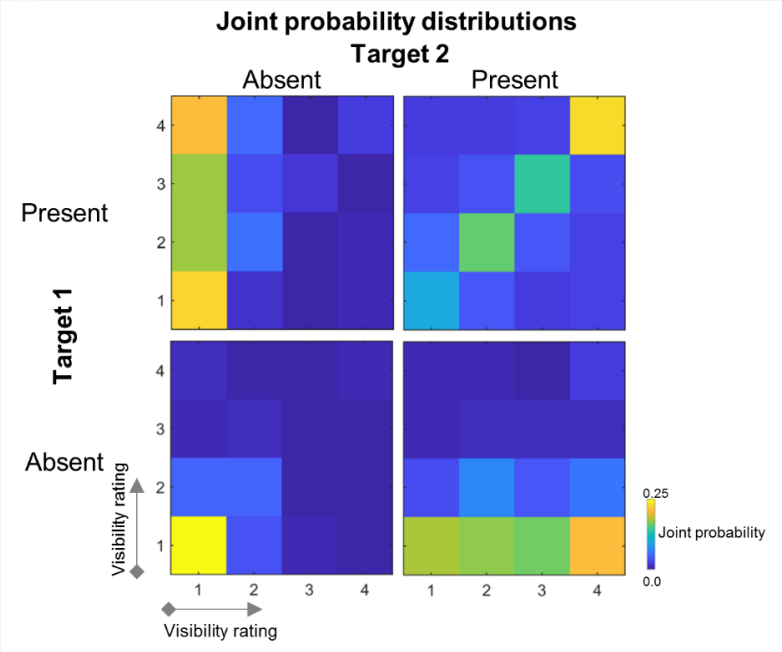
Visibility	Target 1 – Response 2		Target 2 – Response 1		Paired-samples *t* test		
	M	*SD*	*M*	*SD*	*t*	*p*	BF
SeenSeen	26.4%	3.7%	26.1%	3.6%	-0.75	.456	0.23
SeenUnseen	23.6%	12.4%	25.9%	15.1%	1.16	.253	0.28
UnseenSeen	22.0%	8.5%	23.1%	12.2%	0.47	.644	0.19
UnseenUnseen	32.5%	20.1%	29.4%	19.1%	-0.84	.408	0.25
SeenAbsent	23.5%	17.7%	-	-	-	-	-
UnseenAbsent	32.5%	21.9%	-	-	-	-	-
AbsentSeen	-	-	27.5%	22.1%	-	-	-
AbsentUnseen	-	-	28.9%	28.1%	-	-	-

Table 3.3. Summary statistics for the rate of swapping errors for target 1 and target 2
as a function of visibility for the two targets. Statistical comparison was done between the two targets. BF = Bayes' Factor.

Visibility	Target 1 – Response 2		Target 2 – Response 1		Paired-samples *t* test		
	M	*SD*	*M*	*SD*	*t*	*p*	BF
SeenSeen	23.6°	7.6°	22.6°	5.9°	-0.52	.608	0.22
SeenUnseen	24.8°	5.7°	21.3°	6.9°	-2.78	**.009**	4.71
UnseenSeen	25.6°	5.2°	25.2°	5.5°	-0.43	.670	0.21
UnseenUnseen	23.5°	5.9°	23.3°	7.6°	-0.28	.781	0.21
SeenAbsent	23.1°	6.0°	-	-	-	-	-
UnseenAbsent	22.7°	7.6°	-	-	-	-	-
AbsentSeen	-	-	25.6°	5.9°	-	-	-
AbsentUnseen	-	-	22.3°	7.6°	-	-	-

Table 3.4. Summary statistics for the precision related to the rate of swapping errors for target 1 and target 2
as a function of visibility for the two targets. Statistical comparison was done between the two targets. BF = Bayes' Factor.

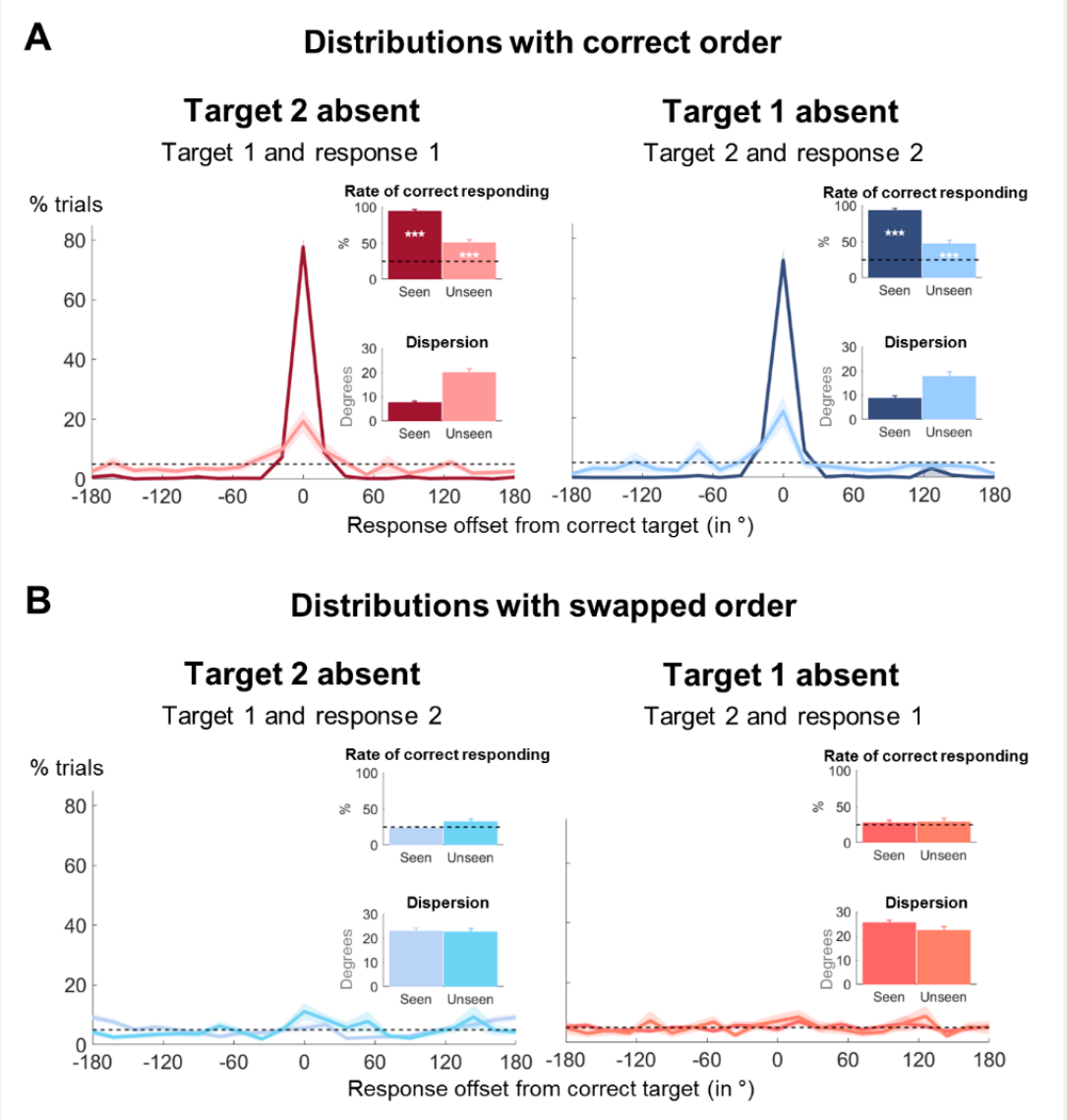
3.8 Supplementary figures



Supplement 3.1

Visibility ratings for the two targets are not fully independent.

Joint probability distributions for all combinations of visibility for target 1 (vertical axis) and target 2 (horizontal axis) as a function of target presence and absence. Hotter colors reflect higher probability.



SUPPLEMENT 3.2

TARGET ABSENCE DOES NOT INFLUENCE LOCALIZATION REPORTS FOR THE OTHER TARGET.

(A) Spatial distributions of forced-choice localization performance in the working memory task on trials with one target are shown as a function of visibility (i.e., seen vs. unseen) for target 1 (left, red) and target 2 (right, blue; 0 = correct target location; positive = counter-clockwise offset). Distributions reflect angular distances between target 1 and response 1 and target 2 and response 2. Insets show rate of correct responding (within ±2 positions of actual location) and precision of working memory representations separately for seen and unseen trials. Error bars indicate standard error of the mean (SEM) across subjects. The horizontal, dotted line illustrates chance-level at 5%. White asterisks show statistical significance when compared to chance (i.e., 25%).

(B) Same conventions as in (A), except that distributions reflect angular distances between target 1 and response 2 (light blue) and target 2 and response 1 (light red). *p < .05, **p < .01, *** p < .001 in a paired-samples t-test (Bonferroni-corrected for 4 comparisons when comparing against chance).

CHAPTER 4 –

PROBING THE LIMITS OF ACTIVITY-SILENT NON-CONSCIOUS WORKING MEMORY

Science is not only a disciple of reason
but, also, one of romance and passion.
- STEPHEN HAWKING

4.1 ABSTRACT

Two types of working memory (WM) have recently been proposed: conscious active WM, depending on sustained neural activity, and activity-silent WM, requiring neither conscious awareness nor accompanying neural activity. However, whether both states support identical forms of information processing is unknown. Theory predicts that activity-silent states are confined to passive storage and cannot operate on stored information. To determine whether an explicit reactivation is required prior to the manipulation of information in WM, we evaluated whether participants could mentally rotate brief visual stimuli of variable subjective visibility. Behaviorally, even for unseen targets, subjects reported the rotated location above chance after several seconds. As predicted, however, such blindsight performance was accompanied by neural signatures of conscious reactivation at the time of mental rotation, including a sustained desynchronization in alpha/beta frequency and a decodable representation of participants' guess and response. Our findings challenge the concept of genuine non-conscious "working" memory and argue that activity-silent states merely support passive short-term memory.

4.2 INTRODUCTION

Working memory (WM) serves a critical role in the online storage of information for rapid access, transformation, and flexible use. Until recently, it was thought to depend on conscious, effortful processing (Baars and Franklin, 2003; Baddeley, 2000, 2003) and the maintenance of persistent neural activity (Fuster and Alexander, 1971; Goldman-Rakic, 1995; Kamiński et al., 2017). However, a growing body of evidence suggests that successful WM maintenance may be dissociated from consciousness and persistent delay-period activity. Items subjectively reported as unseen may still be retrieved above chance-level after several seconds (Bergström and Eriksson, 2014, 2015; King et al., 2016; Soto et al., 2011; Trübutschek et al., 2017). Likewise, an uninterrupted chain of persistent neural firing is not always observed during WM maintenance (Watanabe and Funahashi, 2007, 2014) and content-specific delay-period activity may vanish during the maintenance of non-conscious or unattended information (Rose et al., 2016; Trübutschek et al., 2017; Wolff et al., 2015, 2017).

Theories and simulations indicate that such "activity-silent" maintenance in the absence of accompanying neural activity may be supported by short-term changes in synapses temporarily linking populations of neurons coding for the stored items (Mongillo et al., 2008; Stokes, 2015). Later, a non-specific stimulation of the system may reinstate the original neural firing pattern, an effect that was recently observed experimentally (Rose et al., 2016; Wolff et al., 2017). Short-term synaptic changes may thus effectively allow networks to go silent for several seconds while still supporting a delayed information readout.

While the evidence for active versus activity-silent forms of WM is mounting, whether they support identical forms of information processing remains unknown. Beyond maintenance, a defining feature of WM is the ability to manipulate information, for instance during mental rotation (Baddeley, 1992a; Luck

and Vogel, 2013). If non-conscious WM representations are indeed stored via activity-silent short-term synaptic changes, it is unclear whether they might be transformed without first being reinstated into active firing. Neural network models operate by exchanging patterns of spiking activity, and there exists no theory of how computations could unfold solely via transient synaptic changes (Mongillo et al., 2008). Thus, we predicted that, for an activity-silent WM to enter into an information-processing stream, it would first have to be reinstated into an active form.

We evaluated the limits of information processing for active versus activity-silent WM by asking participants to perform a delayed mental rotation task with subjectively seen and unseen stimuli. Our results suggest that this task can be performed even with invisible stimuli, but that such a manipulation of WM involves the reinstatement of consciousness and persistent neural activity, thus suggesting an intrinsic limit to both activity-silent and non-conscious operations.

4.3 Results

We collected behavioral measures in a first set of participants (*n* = 23), then recorded magnetoencephalography (MEG) signals in a second sample (*n* = 30), always employing the same experimental task (Figure 4.1). On each trial, a target square in gray (barely visible target-present trials, 80%) or black ink (target-absent control condition, 20%) was flashed in 1 of 24 possible locations, then masked. Halfway during the ensuing 3 s delay period, a symbolic cue instructed participants to maintain the original target location (no-rotation condition), or to mentally rotate it 120° clockwise or counter-clockwise (rotation condition). Subjects had to comply with these instructions even if they had not seen the target: They were asked to guess the correct final response location if necessary. At the end of a trial, participants rated their subjective visibility of the target using the classical perceptual awareness scale (Ramsøy and Overgaard, 2004), ranging from *1* (no perception whatsoever) to *4* (clearly seen).

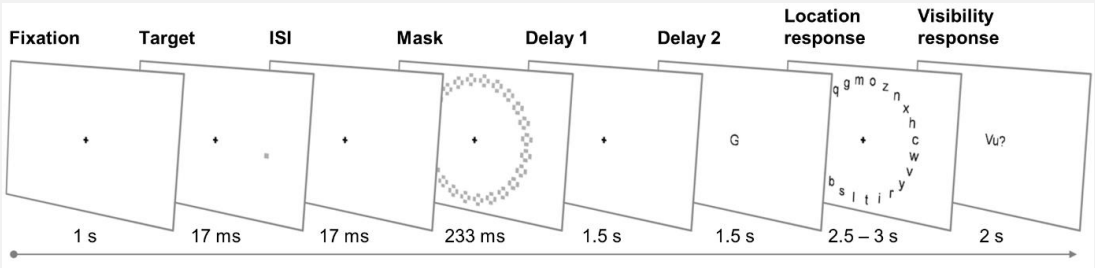


Figure 4.1

Experimental design.

In the behavioral and MEG experiment, participants completed the same spatial delayed-response task. On each trial, a faint target was flashed in 1 out of 24 possible locations and masked. A letter cue presented halfway through a 3 s delay period instructed subjects on the specific task to be performed: (1) Following an equal-sign (« = »), participants were to report the exact location in which the target had appeared. (2) The letter *D* indicated a 120° clockwise, and (3) the letter *G* a 120° counter-clockwise rotation with respect to the target position. At the end of a trial, subjects rated their subjective visibility of the target on a 4-point scale.

4.3.1 Behavioral evidence for mental rotation of non-conscious stimuli

We first quantified the extent to which subjects could detect, maintain, and manipulate targets in the behavioral experiment. Participants varied their visibility ratings as a function of target presence, reporting the vast majority of target-absent trials as unseen (visibility = *1*; 88.1 ± 3.1%) and ~2/3 of the target-present trials as seen (visibility > *1*; 67.7% ± 3.5%). Target detection *d'* therefore exceeded chance (2.0 ± 0.1; $t(22) = 13.2$, $p < .001$). Task (no-rotation vs. rotation) did not modulate subjects' visibility (task x target presence x visibility interaction: $F(1, 22) = 3.2$, $p = .088$), suggesting that participants used the rating scale similarly in both tasks.

Forced-choice localization performance corroborated this interpretation. On seen trials in the no-rotation condition, accuracy was relatively high (65.8 ± 2.5%; chance = 4.17%) and increased monotonically from glimpsed (visibility = *2*) to clearly seen targets (visibility = *4*; all pair-wise comparisons: $p < .05$, except for the comparison between visibility *2* and *3*, where $p = .296$; Figure 4.2A, top). Accuracy remained high on seen rotation trials (30.1 ± 1.9%), albeit, as anticipated, lower than on no-rotation trials ($t(22) = 12.3$, $p < .001$), and without a clear increase as a function of visibility (all pair-wise comparisons: $p > .180$; Figure 4.2A, bottom). Most crucially, even on the unseen trials, performance was well above chance for the no-rotation and rotation task, irrespective of rotation direction (Table 4.1).

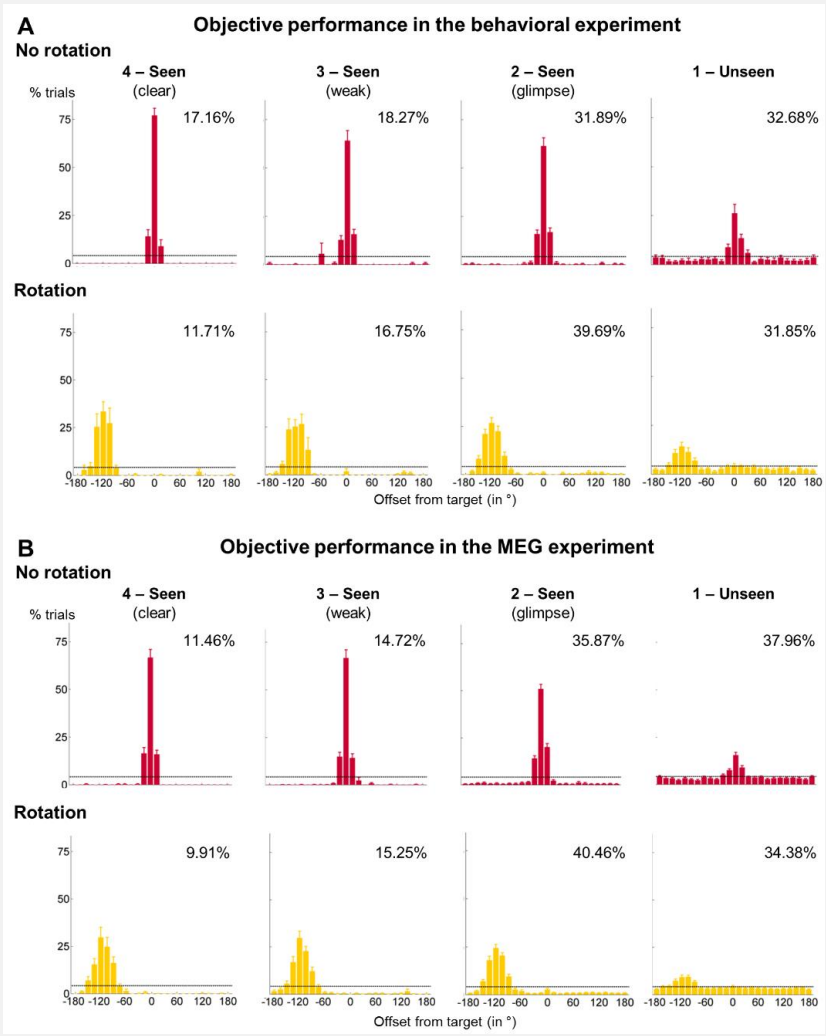


FIGURE 4.2

SPATIAL DISTRIBUTIONS OF FORCED-CHOICE LOCALIZATION PERFORMANCE

in the behavioral **(A)** and MEG **(B)** experiment as a function of task (i.e., no-rotation vs. rotation) and visibility (0° = target location; positive displacement = counter-clockwise offset). The positions at -120° and +120° correspond to the correct locations after clockwise/counter-clockwise rotation. For all analyses and figures, clockwise and counter-clockwise rotations were combined by normalizing all rotation trials into a single rotation condition (i.e., following a counter-clockwise rotation, reflecting a position against 0°). Error bars illustrate the standard error of the mean (SEM) across subjects. The horizontal, dotted lines indicate chance at 4.17%. Percentages in the top right corner of each graph show the grand mean proportion of target-present trials from a given visibility category. Due to low number of trials in visibility ratings *2*, *3*, and *4*, we collapsed these ratings into a *seen* category.

As shown in Figure 4.2A and Figure 4.3A, subjects' responses always surrounded the correct location, yet with greater spread after rotation than no-rotation trials. We separately quantified the rate of approximately correct responding (i.e., correct location ± 30°) and the precision of the spatial representations held in WM (i.e., standard deviation within this tolerance interval; see Methods and Trübutschek et al., 2017). Both task ($F(1, 22) = 9.9$, $p < .001$) and visibility ($F(1, 22) = 151.1$, $p < .001$) affected the rate of correct responding. Participants' responses fell near the correct location more often in the no-rotation (76.5 ± 2.4%) than in the rotation condition (69.4 ± 2.4%), and when having seen (94.1 ± 1.0%) rather than when not having seen the target square (51.9 ± 3.8%). These factors did not interact ($F(1, 22) = 0.2$, $p = .657$; Figure 4.3A, top inset), indicating that decrements in performance following a mental rotation were comparable across seen and unseen targets.

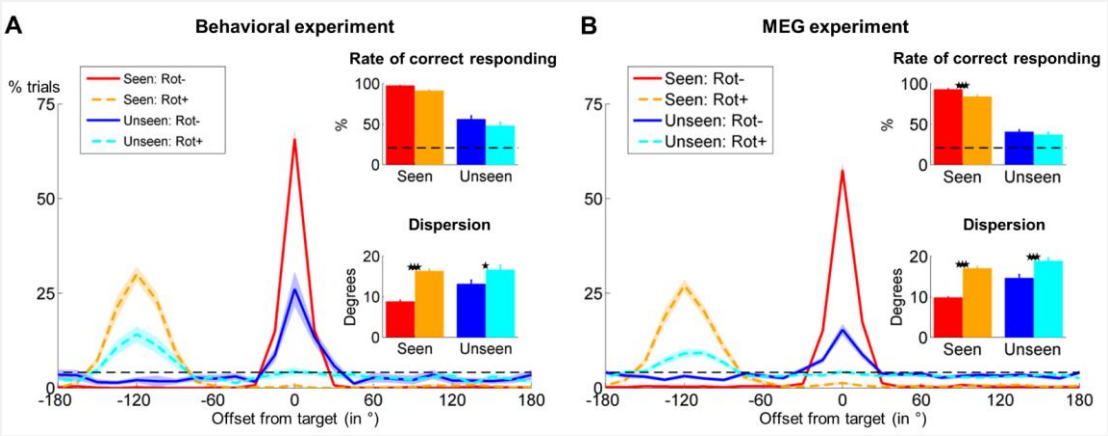


FIGURE 4.3

BEHAVIORAL EVIDENCE FOR MANIPULATION OF NON-CONSCIOUS INFORMATION

in the behavioral **(A)** and MEG **(B)** experiment. Panels depict distributions of participants' localization responses with respect to the target location (0°; positive displacement = counter-clockwise offset) as a function of task (no rotation = solid line, rotation = dotted line) and visibility (seen = warm colors, unseen = cool colors). Insets show the rate of correct responding (proportion of trials within ± 2 positions of correct response location; top) and the precision of working-memory representations in all participants with sufficient blindsight (bottom). Horizontal dotted lines index chance at 4.17% (for single locations) and 20.83% (for the region of correct responding) respectively. Shaded area and error bars represent the standard error of the mean (SEM) across subjects. *p < .05, ** < .01, and *** p < .001 in a paired samples *t*-test.

Analysis of precision reinforced this conclusion: Out of 23 subjects, 19 displayed above-chance blindsight across both rotation directions (chance = 20.83%; $p < .05$ in a χ^2-test) and were thus included here. Task ($F(1, 18) = 34.9$, $p < .001$) and visibility ($F(1, 18) = 10.3$, $p = .005$) again influenced localization performance, but this time also interacted ($F(1, 18) = 8.9$, $p = .008$). Rotating the target location decreased the precision of participants' responses for seen ($t(18) = -11.9$, $p < .001$) and unseen targets ($t(18) = -2.3$, $p = .031$), but this reduction was stronger for seen than unseen trials ($t(18) = -3.0$, $p = .008$; Figure 4.3A, bottom inset). Again, there was therefore no observable detriment to rotating an unseen location.

We replicated these observations in the MEG experiment. Subjects employed the visibility scale meaningfully, rating target-present trials primarily as seen (64.6 ± 3.2%) and target-absent trials as unseen (83.6 ± 2.5%; detection d': 1.7 ± 0.1, $t(29) = 14.2$, $p < .001$) in both tasks (task x target presence x visibility interaction: $F(1, 29) = 2.1$, $p = .159$). Localization accuracy for seen targets was modestly high in the no-rotation condition (57.5 ± 2.2%; Figure 4.2B, top) and reduced following a mental rotation (27.1 ± 1.6%, $t(29) = 14.3$, $p < .001$; Figure 4.2B, bottom). Again, we observed a long-lasting blindsight effect in both tasks and for all rotation directions (Table 4.1). Task and visibility influenced the rate of correct responding (main and interaction effects: all Fs(1, 29) > 4.8, all ps < .036) and precision ($n = 27$; main and interaction effects: all Fs(1, 26) > 8.3, all ps < .008). Mental rotation decreased participants' performance on seen ($t(29) = 5.0$, $p < .001$), but not on unseen trials ($t(29) = 1.8$, $p = .090$; Figure 4.3B, top inset), and also reduced precision more following a rotation with seen ($t(26) = -15.9$, $p < .001$) than unseen targets ($t(26) = -3.9$, $p < .001$; Figure 4.3B, bottom inset).

These findings show that, even when failing to perceive the target, subjects succeeded in manipulating it. However, there exist at least three possible explanations for this long-lasting blindsight effect. First, it may have been the product of a genuine non-conscious manipulation. Second, it may have resulted from a fraction of *seen* trials miscategorized as *unseen*, yet still yielding correct performance; this interpretation, although rejected in our previous experiment without rotation (Trübutschek et al., 2017), needs to be re-examined here. Third, subjects may have recovered the information from non-conscious WM around the time of the cue, transformed it into a conscious, active representation (forced-choice retrieval) and thereafter consciously manipulated this early guess. To resolve these possibilities, we turned to our MEG data, focusing on five a-priori time windows: early brain responses (0.1 – 0.3 s), the P3b time window

previously shown to be critical for conscious perception (0.3 – 0.6 s), the delay period before (0.6 – 1.76 s) and after (1.76 – 3.26 s) the rotation cue, and the response period (3.26 – 3.5 s).

4.3.2 Long-lasting blindsight does not arise from miscategorization of seen trials

Above-chance objective performance for unseen targets could have resulted from the erroneous mislabeling of some seen targets as unseen. If this were the case, the unseen correct trials should display the same neural signatures of conscious processing as seen trials (Trübutschek et al., 2017). There should be an amplification of brain activity during the P3b time window, and a classifier trained to distinguish accuracy on the unseen trials should resemble a standard visibility decoder (i.e., seen vs. unseen). By contrast, the classification of seen versus unseen correct trials should produce a different pattern of results or fail entirely.

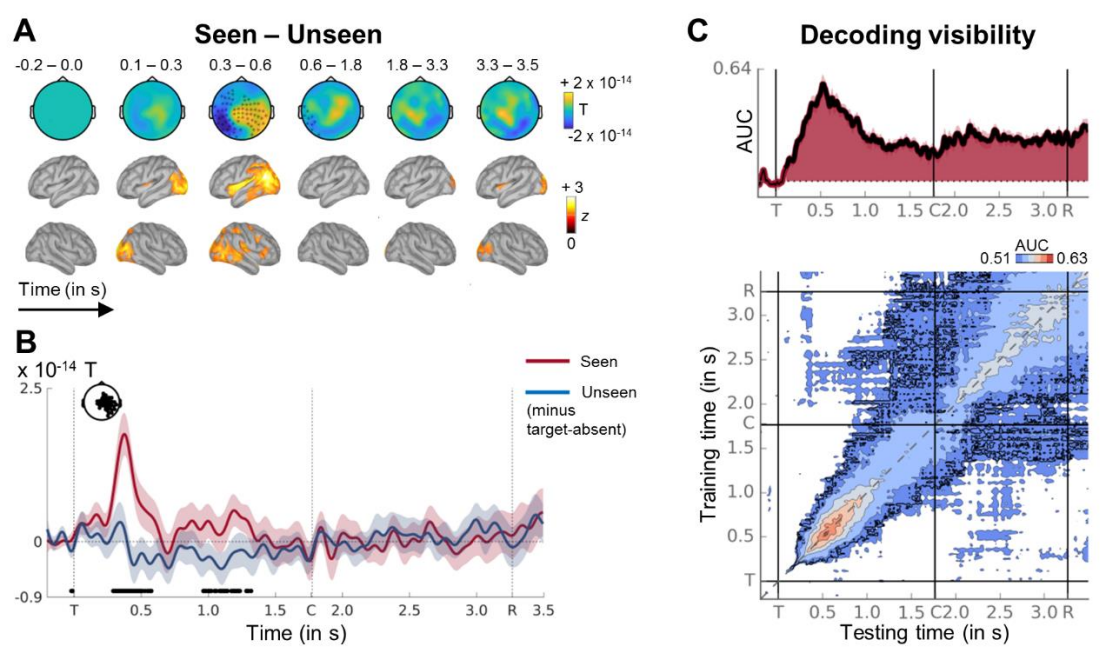


Figure 4.4

Typical neural signatures and dynamics of conscious processing for seen targets.

(A) Sequence of brain activations (-0.2 – 3.5 s) evoked by seen targets in both tasks in sensor (top) and source space (bottom). Each topography depicts the difference in amplitude between seen and unseen trials averaged over the time window shown (magnetometers only). Sources reflect z-scores of absolute difference with respect to a pre-stimulus baseline. Black asterisks indicate sensors showing a significant difference between seen and unseen trials at any point during the respective time window as assessed by a Monte-Carlo permutation test.
(B) Average time courses (-0.2 – 3.5 s) of seen (red) and unseen (blue) trials in that subset of magnetometers having shown a significant effect in (A). Shaded area illustrates standard error of the mean (SEM) across subjects. Significant differences between conditions are depicted with thick black line (two-tailed Wilcoxon signed-rank test, uncorrected). Vertical dotted lines index onset of the target (T), symbolic cue (C), and response (R) screens. For display purposes only, data were lowpass-filtered at 8 Hz.
(C) (Top) Average time course of diagonal decoding of visibility (i.e., seen vs. unseen). Thick black line and shaded area denotes above-chance decoding as assessed by a one-tailed cluster-based permutation analysis. Horizontal, dotted line represents chance level at 50%. (Bottom) Temporal generalization matrix of the same visibility decoder. Each horizontal row in the matrix corresponds to an estimator trained at time *t* and tested on all other time points *t'*. The diagonal gray line demarks classifiers trained and tested on the same time points (i.e., the diagonal estimator shown on top). Thick black outline indexes above-chance decoding as evaluated by a two-tailed cluster-based permutation test. In both plots, vertical lines mark onset of the target (T), symbolic cue (C), and response (R) screens. For display purposes, data were smoothed with a moving average of 5 samples (i.e., 40 ms). AUC = area under the curve.

To evaluate this alternative miscategorization hypothesis, we first characterized univariate neural markers tied to conscious perception. Contrasting brain activity on seen and unseen trials revealed typical signatures of conscious processing (Gaillard et al., 2009; Sergent et al., 2005; Trübutschek et al., 2017). Seen targets elicited a strong positive response between ~300 and 600 ms in right-lateralized centro-parietal sensors, corresponding to activations in occipital, temporal, parietal and dorsolateral prefrontal brain areas (p_{clust} = .011; Figure 4.4A). Moreover, brain activity was amplified during the P3b time window

(i.e., ~292 and 576 ms; $p_{uncorrected} < .05$), though further differences with unseen targets also persisted between ~964 and 1320 ms ($p_{uncorrected} < .05$; Figure 4.4B). Importantly, task did not modulate these brain responses (task x visibility interaction: $p_{clust} > .280$).

When contrasting the unseen correct with the unseen incorrect epochs, we observed no evidence for a miscategorization. No significant differences emerged ($p_{clust} > .221$) and there was no sign of any amplification of brain activity (Supplementary Figure 1A), even when considering the time courses in channels most sensitive to divergences in amplitude for seen and unseen targets (Supplementary Figure 1B). Bayesian statistics provided substantial evidence in favor of the null hypothesis (i.e., no difference in MEG amplitude between unseen correct and incorrect trials) for all time windows (all Bayes' Factors < 0.38).

Because chance corresponded to 20.83% (i.e., 5/24 positions), a non-negligible portion of the unseen correct trials might have resulted from guessing, thus potentially obscuring differences between unseen correct and incorrect epochs. To address this possibility, we next estimated neural activity for unseen correct epochs while accounting for chance-responding (cf. Lamy et al., 2009, footnote 2). If these chance-free unseen correct trials resulted from a miscategorization of seen epochs, we should now observe clear signatures of conscious processing. This was not the case. Chance-free brain activity was still indistinguishable from the one on unseen incorrect and unseen correct trials (whole-brain: all $p_{clust} > .252$; critical time courses: all Bayes' Factors < 0.76). Moreover, it remained strikingly different from a synthetic waveform, derived by proportionally mixing the signals from seen and unseen incorrect trials (as would be expected under the miscategorization hypothesis; Supplementary Figure 1B). Those findings allow us to reject the hypothesis of a miscategorization of some seen trials as unseen.

Decoding analyses refined this conclusion. Training a linear multivariate pattern classifier to discriminate seen from unseen trials resulted in above-chance diagonal decoding from ~120 ms to the end of the epoch (all $p_{clust} < .05$; Figure 4.4C, top), quickly peaking at ~528 ms, then first slowly decaying until the cue before being sustained throughout the remainder of the trial (time bins: AUCs > 0.54, $ps_{corr} < .005$). The temporal generalization of each estimator trained at a specific time to all other time points confirmed this picture (Figure 4.4C, bottom): Visibility decoding was primarily confined to a thick diagonal, indicating that conscious perception was associated with a dynamically evolving chain of metastable patterns of brain activity (King and Dehaene, 2014). Similar findings emerged when training and testing a visibility classifier separately in the no-rotation and rotation condition, or when generalizing from one task to the other (Supplementary Figure 2). Multivariate neural signatures of conscious perception were thus stable across experimental tasks and in line with previous observations (Marti et al., 2015; Salti et al., 2015; Trübutschek et al., 2017).

Crucially, we found no discernable pattern when classifying unseen correct versus unseen incorrect trials (all $p_{clust} > .05$; time bins: AUCs < 0.51, $ps_{corr} > .05$; Bayes' Factors < 0.28; Supplementary Figure 1C). However, training a classifier to distinguish the seen from the unseen correct epochs resulted in a similar, albeit weaker, decoding time course and generalization matrix as when directly training on all unseen or even just the unseen incorrect trials (time bins: AUCs > 0.52, all $ps_{corr} < .05$; Bayes' Factors > 2.07; Supplementary Figure 3). As such, this pattern of results is exactly opposite to what one would have expected in the case of a miscategorization. These findings persisted even when including only those subjects with sufficient blindsight ($n = 27$). This replication of our previous work (Trübutschek et al., 2017) thus rules out a miscategorization of unseen correct trials as an alternative explanation for the long-lasting blindsight effect. Instead, it indicates that information was genuinely encoded in non-conscious WM.

4.3.3 Long-lasting blindsight effect results from active, conscious rotation

What process allowed participants to perform a mental rotation on unseen trials? Was it the result of a genuine non-conscious manipulation? Or did subjects perform a conscious manipulation by first

reinstating an active representation of the estimated target position around the time of the rotation cue and then rotating this conscious guess? Disambiguating between these alternatives requires the identification of a neural marker of active, conscious processing. Prior work has pointed towards a rhythmic signal – a suppression of power in the alpha (8 – 12 Hz) and low (13 – 20 Hz) as well as high beta frequency bands (20 – 27 Hz) – as a reflection of such a cognitive state (Backer et al., 2015; Meyniel and Pessiglione, 2014; Trübutschek et al., 2017).

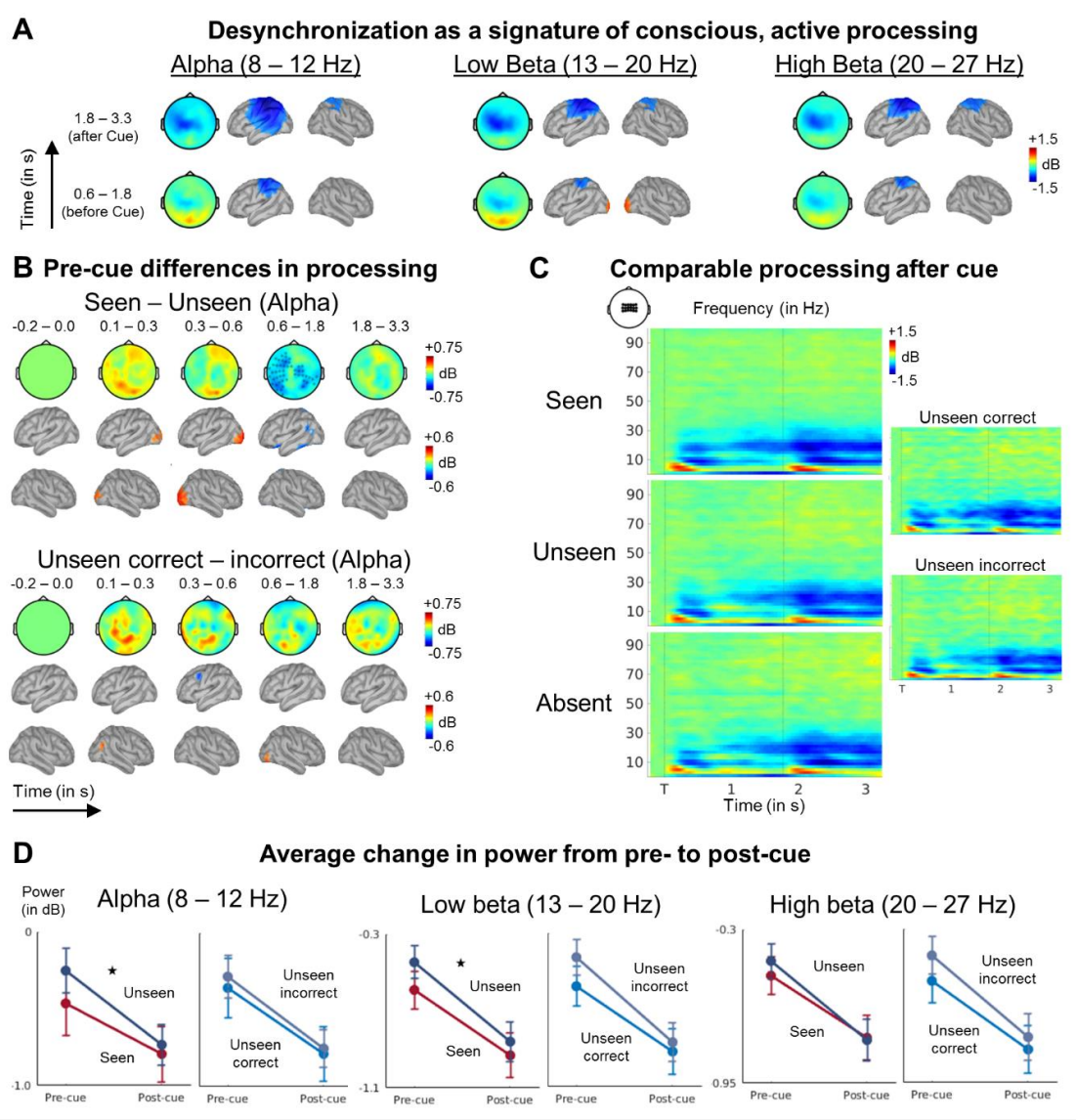


FIGURE 4.5

TIME-FREQUENCY MARKERS OF CONSCIOUS PROCESSING EMERGE AROUND THE TIME OF THE ROTATION CUE ON THE UNSEEN TRIALS.

(A) Average pre-cue (0.6 – 1.8 s; bottom) and post-cue (1.8 – 3.3 s; top) desynchronization in the alpha (8 – 12 Hz; left), low beta (13 – 20 Hz; middle), and high beta (20 – 27 Hz; right) frequency bands in magnetometers and source space (in dB; relative to the pre-stimulus baseline).
(B) (Top) Alpha band activity (8 – 12 Hz) related to consciously perceiving the target square (i.e., seen vs. unseen) is shown in magnetometers and source space (in dB; relative to pre-stimulus baseline). Black asterisks denote cluster of sensors displaying a significant difference at any point in time during the respective time window (as evaluated by a Monte-Carlo permutation test). (Bottom) Same as on top, but for the contrast between unseen correct and unseen incorrect trials.
(C) Average time-frequency power relative to baseline as a function of visibility and target presence in a subset of central magnetometers. Vertical lines demark onset of target (T) and cue presentation.
(D) Plots depict average pre-cue and post-cue power in the same group of sensors as in (C) as a function of frequency (i.e., alpha, low beta, and high beta) and visibility (i.e., seen, unseen, unseen correct and unseen incorrect). Error bars represent standard error of the mean (SEM) across subjects. Asterisks denote significant interaction in a repeated-measures ANOVA at $p < .05$.

Across all trials, we indeed observed a prominent desynchronization in alpha/beta frequencies over an extensive set of central sensors, emanating primarily from parietal brain sources (Figure 4.5A). Cluster-based permutation analyses revealed reliable differences in brain responses in a slightly larger set of channels between seen targets and all other experimental conditions exclusively prior to the presentation of the rotation cue. Power decreased more strongly on seen than on unseen trials between ~580 and 1320 ms in the alpha (p_{clust} = .032), and between ~460 and 1300 ms in the low beta band (p_{clust} = .046; Figure 4.5B, top). Similarly, pre-cue desynchronizations were more pronounced for seen than for target-absent epochs in the low (p_{clust} = .015) and high beta bands (p_{clust} = .030) between ~280 and 940 and ~820 and 2000 ms. There were no discernable differences in the power profiles between (1) unseen and target-absent trials (all p_{clust} > .250) and (2) unseen correct and incorrect epochs (all p_{clust} > .280; Figure 4.5B, bottom). Desynchronization of alpha/beta power may therefore serve as a signature of conscious processing in the current task.

Using this marker, we are now in a position to evaluate the remaining alternatives. If the long-lasting blindsight effect resulted from a genuine, non-conscious rotation, on seen trials, we should observe a sustained desynchronization in the alpha and beta bands throughout the entire epoch, while no (or at least significantly weaker) power decreases should be associated with unseen and target-absent epochs. By contrast, if participants consciously rotated a guess, neural signatures of conscious processing should be highly similar across all experimental conditions after the cue. Differences in desynchronization between seen and unseen/target-absent trials should only exist during the pre-cue phase.

Our results support the latter hypothesis (Figure 4.5C). Following an initial divergence during the early pre-cue maintenance phase (Supplementary Figure 4A-C), differences in spectral profiles between seen, unseen, and target-absent trials vanished by ~1 s. All epochs were characterized by a prominent, sustained desynchronization in the alpha, low and high beta frequencies. This suppression in power varied as a function of subjective visibility (i.e., seen vs. unseen) and time (i.e., pre-cue vs. post-cue delay). It was much more pronounced during the post-cue than the pre-cue maintenance period (i.e., main effect of time: all *F*s > 18.6, all *p*s < .001). Crucially, this difference between pre- and post-cue power was also larger for unseen than for seen targets in the alpha and low beta bands (visibility x time interaction: all *F*s > 4.01, all *p*s <= .05), and marginally so in the high beta band (visibility x time interaction: $F(1, 29) = 2.95$, $p = .097$; Figure 4.5D). No such interaction emerged when contrasting the unseen correct with the unseen incorrect trials (i.e., visibility x time interaction: all *F*s < 2.83, all *p*s > .103; Figure 4.5D), as these conditions displayed largely similar power profiles throughout the entire epoch (Supplementary Figure 4D-F).

We thus observed a reliable distinction between seen and unseen brain states only during the maintenance period preceding the execution of the experimental task up until at least 1 s. Seen targets were accompanied by a significantly larger desynchronization in the alpha and low as well as high beta frequencies. These differences vanished entirely by the time the symbolic rotation cue was presented. The mental rotation task appeared to be solved by reinstating a conscious estimate of a target location.

4.3.4 The location of unseen targets can only be tracked transiently

To further test this conclusion, we used multivariate decoding to track neural activity underlying the encoding, maintenance, manipulation and retrieval of seen and unseen target locations. We first trained a multivariate regression model to predict target angle from participants' brain activity separately for each point in time. In order to maximize statistical power and increase our ability to detect small effects, we fitted the estimator while collapsing target-present trials across rotation and visibility conditions. We then evaluated model performance on left-out subsets of epochs (see Methods for details). Note that, unless explicitly stated, none of the findings changed qualitatively when testing separately on the rotation and no-rotation task (Supplementary Figure 5).

Starting at ~80 ms, estimator performance for seen targets steadily rose until ~264 ms and then slowly decayed towards chance at ~1.46 s (Figure 4.6A). Following the rotation cue, a rebound of position-selective activity was observed and was then fairly sustained for the remainder of the trial, with a short gap between ~2.70 and 3.10 s right before the onset of the response screen (p_{clust} < .05; time bins: *W*s > 417.0, ps_{corr} < .005, Bayes' Factors > 77.93). Thus, in line with previous findings (Trübutschek et al., 2017), seen targets were initially encoded via active neural firing. Then, this representation decayed and was reactivated throughout most of the post-cue delay period.

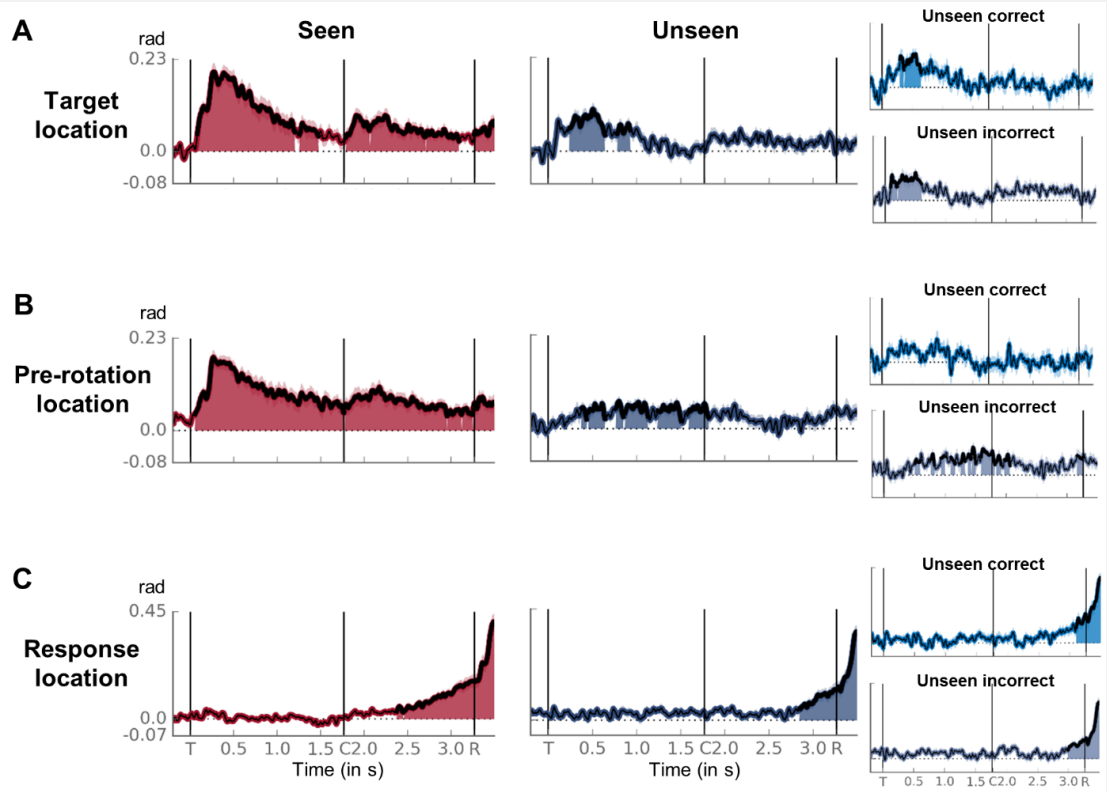


FIGURE 4.6

TRACKING A MENTAL ROTATION ON SEEN AND UNSEEN TRIALS.

(A) Time courses of average decoding of target location on seen (red), unseen (dark blue), unseen correct (light blue) and unseen incorrect (blue) trials. Thick lines and shaded areas represent above-chance performance as assessed by a one-tailed cluster-based permutation test. Horizontal dotted lines index chance. Event markers denote the onset of the target (T), cue (C), and response (R) screens. For illustration purposes, data were smoothed with a moving average of 5 samples (i.e., 40 ms).
(B) Same as in (A), but for pre-rotation location.
(C) Same as in (A), but for response location.

A different picture emerged for unseen targets. While target location was again encoded and actively stored during the early part of the epoch, this representation was weaker than the one for seen targets (paired-samples Wilcoxon signed rank test: pre-cue time bins: *W*s > 370.0, ps_{corr} < .02, Bayes' Factors > 3.42) and decayed much more quickly, vanishing entirely by ~920 ms (p_{clust} < .05; pre-cue time bins: *W*s > 351.0, ps_{corr} < .035, Bayes' Factors > 7.34). During the post-cue delay period, although we found no evidence in favor of an actively coded representation of target location when considering the decoding time course itself (p_{clust} > .05; Figure 4.6A), the estimator's performance over the entire time window remained above chance (rads = 0.03 ± 0.01, *W* = 355.0, p_{corr} =.025, Bayes' Factor = 6.41) and at comparable levels as on seen trials (*W* = 315.0, p_{corr} = .460, Bayes' Factor = 0.86). A more fine-grained analysis with a moving average of 100 ms revealed that this effect was driven primarily by the initial phase of the delay, up to ~2.6 s. We observed no modulation of this pattern of findings by accuracy (time bins: *W*s < 279.0, all ps_{corr} > .950, Bayes' Factors < 0.41; Figure 4.6A, insets).

Overall then, a mixture of two different mechanisms seems to have supported the initial, pre-cue storage of seen and unseen target locations. Whereas seen targets were maintained with persistent albeit

decaying, neural activity, unseen targets elicited weaker position-related activity that also quickly decayed to baseline-level. During the post-cue phase, once participants either actively maintained or manipulated the contents of their WM, the representation of seen targets was reactivated and sustained for the remainder of the epoch. Unseen targets may also have benefitted from a short-lived revival, but this effect was weak and the associated decoding time course much less compelling than the one for seen trials.

4.3.5 An estimate of the location of unseen targets is reinstated prior to the rotation cue

Localization responses on unseen trials did not always follow the actual target position. On more than half of the unseen trials (62.0 ± 2.8%), subjects chose an incorrect location. What determined participants' final response on those trials? According to the activity-silent account of WM, around the time of mental rotation, subjects should have attempted to reinstate an active neural representation of the target, albeit with occasional location errors, and then rotated this guess. To evaluate this prediction, we set out to track the neural representation of participants' location estimates throughout the task. Around the time of the rotation cue, brain signals should contain a decodable representation of the "pre-rotation location", i.e. the spatial location that, given the subjects' response, would have been the location retrieved and then rotated. On no-rotation trials, this location coincided with response location, whereas on rotation trials, it corresponded to the position of participants' response rotated 120° in the direction opposite to what the rotation cue had instructed. Detecting the presence of such a pre-rotation representation on unseen rotation trials would support the results of our time-frequency analyses and the hypothesis that, around the time of the cue, subjects attempted to recover a conscious representation of the target (sometimes an erroneous one) and then consciously rotated this guess. If, however, unseen performance was based on an active manipulation of activity-silent WM, then such decoding should fail.

On seen trials, decoding the pre-rotation location was possible, with a time course strikingly similar to the one for the true position of the target (Figure 4.6B). From ~56 ms onwards, the pre-rotation location was coded in activity-based brain states ($p_{clust} < .05$; time bins: Ws > 408.0, $ps_{corr} < .005$, Bayes' Factors > 517.26), first peaking at ~ 264 ms (rad = 0.18 ± 0.02) and then slowly decaying before being revived by the rotation cue and sustained for the remainder of the epoch.

Crucially, pre-rotation location could also be decoded on unseen trials. Shortly after the presentation of the target, the estimator's performance began to rise and first exceeded chance at ~376 ms (rad = 0.052 ± 0.015). Decoding persisted until ~1.8 s ($p_{clust} < .05$; P3b time window and pre-cue delay: Ws > 382, $ps_{corr} < .005$, Bayes' Factors > 78.83), though estimator performance itself did not drop until ~ 2.5 s. Indeed, a follow-up analysis with narrower 100-ms time windows suggested that the pre-rotation location may have been maintained until ~ 2.2 s ($p < .05$, uncorrected). There was again no evidence for a modulation of this pattern as a function of accuracy (time bins: Ws > 120.0, $ps_{corr} > .600$, Bayes' Factors < 1.44; Figure 4.6B, insets).

As predicted, while the representation of the pre-rotation location was stronger for seen than for unseen targets during the early part of the epoch (early and P3b time window: Ws > 450.0, $ps_{corr} < .005$, Bayes' Factors > 124,688.30), this difference started to diminish during the pre-cue maintenance phase (W = 347.0, $p_{corr} = .085$, Bayes' Factor = 1.76) and vanished entirely by the last second before the rotation cue (moving average of 100 ms: Ws < 359.0, $ps_{corr} > .05$, Bayes' Factor < 1.32). Participants' location estimates were therefore similarly represented on both seen and unseen trials during the last part of the pre-cue maintenance period: Even on unseen trials, the material rotated was an active, conscious guess of a target location.

4.3.6 An active representation of target location is mentally rotated in WM

We last trained and tested a multivariate regression model to decode response location. On seen trials, response location emerged reliably only in the second half of the post-cue delay period (Figure 4.6C).

Starting at ~2.38 s, decoding performance gradually built up until its peak at the very end of the epoch (p_{clust} < .05; post-cue time bins: *W*s > 440.0, ps_{corr} < .005, Bayes' Factors > 21,997.68). There was substantial temporal overlap between the decoding of the target/pre-rotation location and the response position: As the former started to decay around ~2.5 s, the latter slowly began to pick up.

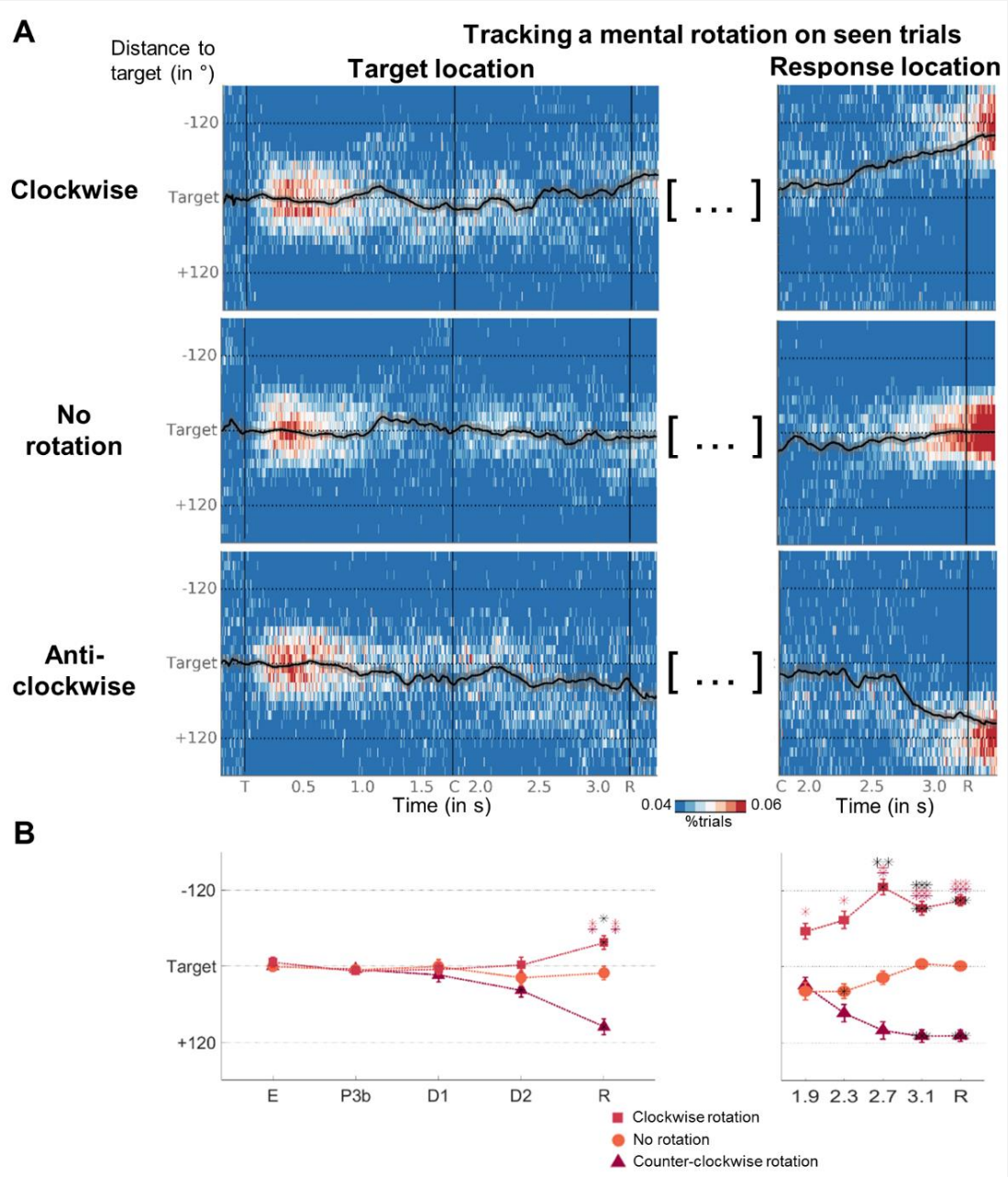


FIGURE 4.7

TRACKING A MENTAL ROTATION ON SEEN TRIALS.

(A) (Left) Time courses of probability density distributions of the angular distance between the estimates of a decoder trained with target angle and actual target location are shown as a function of rotation condition. For display purposes, data were smoothed with a moving average of 12 samples (i.e., 96 ms). Overlaid black line illustrates the evolution of the circular mean of the individual distributions (also smoothed). Shaded area reflects circular standard variation across subjects. Vertical event markers denote the onset of the target (T), cue (C), and response (R) screens, horizontal markers index correct response positions after rotation. (Right) Same as in the left panels, except for angular distance between the estimates of a decoder trained with response angle and actual target location.
(B) Circular means of the above distributions as a function of rotation condition and time bin (i.e., E = 100 – 300 ms, P3b = 300 – 600 ms, D1 = 0.6 – 1.76 s, D2 = 1.76 – 3.26 s, R = 3.26 – 3.5 s). Error bars reflect circular standard deviation. Asterisks inside markers denote significant deviation from mean direction of 0 (as assessed by a circular equivalent of a one-sample *t*-test), asterisks on top significant differences in median direction between conditions (as assessed by a circular equivalent to the Kruskal-Wallis test; black = clockwise vs. counter-clockwise; red = clockwise vs. no rotation; violet = counter-clockwise vs. no rotation). * p < .05, ** p < .01, *** p < .001.

Figure 4.7 further shows the probability density distributions for decoded target and response locations. On seen trials, prior to the rotation cue, decoder estimates for target angle were strongly concentrated around the actual target location, irrespective of rotation condition and direction (resultant vector lengths > .41; Rayleigh tests for non-uniformity: zs > 5.09, *p*s < .005; non-parametric multi-sample test for equal medians: *p*s > .302). This picture changed following the rotation cue. While angle estimates on no-rotation trials continued to stay fairly centered on the original target location (resultant vector lengths > .37; Rayleigh test: z > 4.01, $p < .017$), their counterparts for clock- and counter-clockwise rotations began to shift towards the respective correct response positions (response period: clockwise rotation: M_{circ} = 37.3°; resultant vector length = .49; one-sample test against a mean direction of 0°: $p < .05$; counter-clockwise rotation: M_{circ} = 95.6°; resultant vector length = .31; one-sample test against a mean direction of 0°: $p < .05$). During the response period, all three distributions were characterized by a different center of mass (non-parametric multi-sample test for equal medians: *p*s < .05), located in close proximity to the expected final position. Depending on the direction of the rotation, the representation of the original target location was progressively transformed into a representation of the response position. On average, then, a mental rotation following seen targets was reflected by an active transition period, during which the stimulus code was progressively replaced by the response code. Note however that, while such a smooth transition was visible in the mean, we cannot determine here whether continuous or discrete transitions occurred on individual trials (Latimer et al., 2015).

We next considered the unseen trials. If subjects similarly performed a conscious rotation of (an estimate of) unseen locations, then one would predict the response estimator to perform comparably on seen and unseen targets. This was indeed the case (Figure 4.6C). Decoding response location on unseen trials yielded consistent above-chance performance from ~2.84 s onwards ($p_{clust} < .05$; post-cue time bins: *W*s > 410.0, $ps_{corr} < .005$, Bayes' Factors > 594.74), again beginning to rise around the same time as the model for the pre-rotation location had faded (cf. time courses in Figure 4.6B and Figure 4.6C). As would be expected if the same underlying process were responsible for the generation of responses across all experimental conditions, we observed no differences as a function of accuracy (time bins: *W*s < 314.0, $ps_{corr} > .480$, Bayes' Factors < 0.81) or visibility (time bins: *W*s < 334.0, $ps_{corr} > .600$, Bayes' Factors < 2.45). Pre-rotation and response locations could also be tracked on unseen trials, albeit, as expected, with reduced accuracy (Supplementary Figure 6). The transformation from one representation into another therefore appeared to have been comparable for seen and unseen targets, in both cases relying on decodable activity patterns rather than on activity-silent brain states.

4.4 Discussion

Recent work has challenged classical views of WM as a purely conscious process based on persistent neural firing. Instead, information may also be stored in non-conscious, activity-silent WM, without any accompanying neural activity, via slowly decaying changes in synaptic weights (Mongillo et al., 2008; Rose et al., 2016; Stokes, 2015; Trübutschek et al., 2017; Wolff et al., 2015, 2017), and in the complete absence of subjective awareness (Bergström and Eriksson, 2017; Soto et al., 2011; Trübutschek et al., 2017). So far however, only the short-term maintenance of information has been explored, while its transformation, a key feature of WM, has been ignored.

Here, we show that, whether or not information was consciously perceived, manipulating it was associated with a prior reinstatement of an active neural representation, accompanied by signatures of a conscious state. These findings question the term non-conscious *working* memory, and suggest that WM manipulation requires a conversion from activity-silent to active WM.

4.4.1 Manipulation as a limit for non-conscious, silent processes

It has proven notoriously difficult to put clear upper bounds on the depth of non-conscious processing. Non-conscious signals tend to affect a wide range of behaviors and trigger activity in many different brain

areas, including the prefrontal cortex (van Gaal et al., 2010; Naccache and Dehaene, 2001; Nakamura et al., 2018; van Vugt et al., 2018). Recent work on non-conscious WM has even called into question some of the most basic assumptions regarding the nature of non-conscious processes, suggesting that non-conscious signals may be maintained much longer than previously thought (Bergström and Eriksson, 2017; King et al., 2016; Soto et al., 2011; Trübutschek et al., 2017).

Our behavioral results, superficially, support this conclusion, as they provide evidence for a non-conscious process of mental rotation. On unseen trials, subjects reported the correct response position much better than chance after several seconds, irrespective of whether they just had to maintain the original target location or rotate its position. We replicated this long-lasting blindsight effect in two independent experiments and, as such, seemingly expanded the range of possible non-conscious WM processes to include manipulation of information (Bergström and Eriksson, 2015; Bona et al., 2013; Soto et al., 2011; Trübutschek et al., 2017).

Our neural data further indicated that subjective visibility reports were genuine. Prior to the rotation cue, we observed typical markers of conscious, active processing almost exclusively for seen targets. Brain activity was amplified during the P3b time window (Gaillard et al., 2009; Sergent et al., 2005), and participants' visibility (i.e., seen vs. unseen) was decodable with high accuracy (King et al., 2016; Salti et al., 2015; Trübutschek et al., 2017). Moreover, there was a sustained desynchronization of alpha/beta frequency, which became even more pronounced after the rotation cue, thereby coinciding with the most demanding phase of our task (Pessiglione et al., 2007; Trübutschek et al., 2017; Wyart and Tallon-Baudry, 2009). By contrast, for unseen targets, signatures of conscious processing were entirely absent or markedly reduced in comparison to the ones on seen trials early during the epoch. There was neither an ignition of brain activity during the P3b time window, nor a comparably strong alpha/beta desynchronization. These findings, in line with our previous work (Trübutschek et al., 2017), show that "unseen" trials were genuine and did not correspond to a subset of miscategorized seen trials.

Those neural signatures, however, changed drastically around the time of the mental rotation cue, suggesting that an estimate of target location was reactivated and regained consciousness. Slightly before the rotation cue, around ~1 s, alpha/beta power decreased for unseen targets, reaching similar levels as on seen trials during the post-cue maintenance period. Starting at more or less the same time (i.e., around ~500 ms), a decodable representation of the pre-rotation location emerged. Participants therefore seem to have estimated and reinstated an active representation of target location in anticipation of the upcoming rotation task. On unseen trials, the weak activity-silent representation of the target may have competed against other ongoing noise fluctuations in the brain, resulting in a mixture of trials where decision was solely based on stochastic events (Vul et al., 2009) and others biased towards the correct target location. Variability across trials and participants as well as the temporal smoothing inherent to time-frequency analyses precludes a definitive determination of the exact onset of the pronounced and sustained alpha/beta desynchronization on unseen trials, but the results indicate that this transition already occurred shortly before the presentation of the symbolic rotation cue.

In conjunction with previous work (Bergström and Eriksson, 2017; Soto et al., 2011; Trübutschek et al., 2017), these findings thus highlight the limits of non-conscious WM. While information may be temporarily stored non-consciously, manipulating items is associated with a reinstatement of an active conscious representation. Our results may thus help to circumscribe the boundaries of non-conscious processing. Consciousness has been theorized and empirically demonstrated to be a necessary prerequisite for the execution of serial tasks, such as the chaining of mental operations (Dehaene, 2001; Sackur and Dehaene, 2009). We here observed that such chaining may remain possible even if the initial input was not represented consciously, but only inasmuch as subjects willfully operate on previously non-conscious information by forcing it into an active state before routing it to a conscious processor. Future research might expand on this work and attempt to more strongly encourage the reliance on non-conscious processing by, for instance, rendering the task cues subliminal.

4.4.2 The complementarity of active and silent processes in WM

Our data speak to the current debate on the nature of WM representations in the brain. Traditional models emphasize stable, persistent neural activity as the main candidate mechanism supporting WM (Fuster and Alexander, 1971; Kamiński et al., 2017). More recent, multivariate investigations point towards a more dynamic view, with the contents of WM being maintained in dynamically changing patterns of neural activity or activity-silent brain states (Rose et al., 2016; Spaak et al., 2017; Stokes, 2015; Stokes et al., 2013; Trübutschek et al., 2017; Wolff et al., 2015, 2017).

Together with our previous work (Trübutschek et al., 2017), our current results suggest that sustained neural activity and activity-silent mechanisms may accommodate different processes. Storage of information in WM need not require neural activity. Without the manipulation requirement in our task, delay-period activity vanished entirely for unseen and was only intermittent for seen targets (Trübutschek et al., 2017). Such prolonged activity-silent periods occurred less frequently in the current experiment, probably because participants tried to more actively retain information about the target location in preparation for the required mental rotation. However, even in the present setting, target-related neural activity first decayed towards chance before being reactivated by the cue.

By contrast, after the symbolic cue, once subjects were manipulating the contents of their WM, neural activity was sustained throughout the remainder of the epoch, with the representation of the response emerging while the target representation slowly faded. Importantly, we observed a similar pattern of results for unseen targets. As decodability of target location vanished, it was replaced by the emergence of the guess (i.e., pre-rotation location), that was maintained until the rise of response-related neural activity. The slightly different post-cue time courses observed for the decoding of the pre-rotation location on seen and unseen trials may not indicate any meaningful difference in the type of operation deployed by the participants, but likely reflected the differential levels of certainty with which subjects performed the mental rotation, having a clear starting point on seen trials and a more fluctuating representation on unseen trials.

Taken together, then, we propose that active and activity-silent processes make distinct contributions to WM. WM maintenance can be achieved without any accompanying neural activity via activity-silent mechanisms, but WM manipulation appears to depend on active neural firing. Recent evidence from a computational model corroborates this conclusion by demonstrating that, while short-term synaptic plasticity may support short-term maintenance, persistent neuronal activity automatically emerges from learning during active manipulation (Masse et al., 2018). Moreover, similar divisions of labor between activity-silent and activity-based brain states have recently been observed for the active selection vs. maintenance of WM contents (Quentin et al., 2018). All of these data thus lend support to the emerging view that WM is best conceptualized as an activity-induced temporary and flexible shift in the functionality of a network (i.e., dynamic coding; Stokes, 2015).

4.4.3 Tracking intermediate representations during a mental rotation

A last aspect of our work that deserves attention concerns the act of mental rotation itself. Numerous behavioral and neuroimaging studies support the idea that mental rotation depends on analog spatial representations, with the initial representation progressively being rotated through intermediate positions or views. Reaction times have been found to increase in near-linear fashion with the size of the rotation angle (Cooper, 1975; Shepard and Cooper, 1986; Shepard and Metzler, 1971), and activity in spatially mapped brain areas, such as the posterior parietal cortex, has been reported to be modulated parametrically by angular distance (Gauthier et al., 2002; Jordan et al., 2001; Wager and Smith, 2003). Recordings of single-neuron activity from the motor cortex during a motor rotation task also suggest a gradual rotation of a neural population vector (Georgopoulos et al., 1989).

Our results indicate that such a transformation of neural representations is now decodable from human MEG recordings. On seen trials, following the rotation cue, average decoder estimates of target and response angle progressively moved away from the original target location towards the expected response position, seemingly passing through a series of intermediate locations. A similar transformation may also have been present for the pre-rotation location for unseen targets, though data were too noisy to support any definitive conclusions. These findings are compatible with the view that locations intermediate between the target/pre-rotation position and the response location were coded and represented in the brain. However, this interpretation is based on an analysis of multivariate estimates averaged across trials and participants. Isolated bursts of activity, occurring at different points in time and coding for discrete spatial positions, if averaged over many events, might also result in the apparent smooth transition we observed here (Lundqvist et al., 2016; Stokes and Spaak, 2016). Future research relying on single-trial analyses will be needed to disambiguate between these alternatives.

4.4.4 Conclusion

In the wake of recent proposals of non-conscious and/or activity-silent WM, we have identified an important boundary condition: While the storage of information in WM requires neither consciousness nor persistent activity, the manipulation of WM contents is associated with both. This conclusion is at odds with the very idea of non-conscious *working* memory. We therefore propose "activity-silent short-term memory" as an alternative term for the phenomenon of long-lasting blindsight. This observation may also help reconcile current debates on the nature of WM. WM is a generic term that refers to a conglomerate of cognitive processes including attentional selection, storage, and manipulation. Active and activity-silent brain states both contribute to produce these behaviors, and an essential goal for future research will be to further disentangle their differential contribution to WM.

4.5 Methods

4.5.1 Participants

23 healthy volunteers (4 men; M_{age} = 23 years, SD_{age} = 2.5 years) with normal or corrected-to-normal vision were included in the behavioral experiment. Another 30 participants (14 men; M_{age} = 25.4 years, SD_{age} = 3.8 years) were entered in the analyses of the MEG study. In compliance with institutional guidelines, all subjects gave written informed consent prior to enrollment and received up to 80€ as compensation.

4.5.2 WM task

We adapted our previous paradigm (Trübutschek et al., 2017) to probe participants' ability to manipulate WM representations under varying levels of subjective visibility (Figure 4.1). Following a 1 s fixation period, a small, gray target square was flashed for 17 ms in 1 of 24 circular locations and subsequently masked (233 ms). Mask contrast was calibrated separately for each subject to yield ~equal proportions of seen and unseen trials (see below). Halfway throughout a 3 s delay period, a centrally presented, symbolic cue in white ink instructed participants as to the specific task to be performed: A third of the trials, indexed by an equal sign, served as a control condition, requiring subjects to maintain and identify the position in which the target had appeared. On the remainder of the trials, participants were to mentally rotate the original target location and report this rotated position. While the uppercase letter *D* necessitated a 120° clockwise rotation (1/3 of the trials), the letter *G* indicated a 120° counter-clockwise rotation (1/3 of the trials). Subjects responded by either speaking (MEG experiment; 2.5 s) or typing on a standard AZERTY keyboard (behavioral experiment; 3 s) the letter – out of a set of 24 (excluded: *j*, *p*) randomly presented in all possible locations, – corresponding to the desired position. For example, had the cue in Figure 4.1 been an equal sign, participants would have had to report the letter *w*. Had it been a *D*, the correct answer would have been *b*. With the trial as shown, subjects should have indicated the letter

g. Importantly, a location response was required even when participants had not seen the target square; in that case, they were instructed to guess the correct final position. Subjects then rated their visibility of the target on the 4-point Perceptual Awareness Scale (Ramsøy and Overgaard, 2004), using the index, middle, ring, and little finger of their right hand to operate either the number-pad keys of the computer keyboard (behavioral experiment; 2 s) or the buttons of a non-magnetic response box (Fiber Optic Response Pad, Cambridge Research Systems Ltd; MEG experiment; 2 s). To qualify as unseen (visibility = *1*), participants were to have no visual experience whatsoever of the target stimulus as well as no hunch concerning its location. All other subjective impressions were to be categorized as seen (visibility *2*, *3*, or *4*). Inter-trial intervals (ITIs) ranged between 333 and 666 ms (MEG experiment) or between 1 and 2 s (behavioral experiment). A central fixation cross was shown throughout the entire trial, and 20% target-absent catch trials were included to allow for the computation of objective measures of subjects' perceptual sensitivity and for the isolation of brain activity specific to the target square.

4.5.3 Calibration task

Participants performed a separate calibration procedure to identify the mask contrast needed for roughly equal proportions of seen and unseen targets in the WM paradigm. Trials were identical to the first part of the main experimental task (up to, and including, the presentation of the mask), but required either an immediate target localization and visibility response (behavioral experiment) or just an instantaneous visibility rating (MEG experiment). Mask contrasts were adjusted on a trial-by-trial basis with a double-staircase technique: We first divided the color spectrum between black and white into 20 equally spaced hues. Following an unseen target (visibility = *1*), mask contrast was reduced by one step on the subsequent trial, whereas it was increased by the same amount when subjects had seen the target (visibility > *1*). Initial values for the two staircases were set to RGB values of 12.75, 12.75, 12.75 and 242.5, 242.5, 242.5, respectively, and one of the two staircases was selected randomly at the beginning of each trial. In case of target-absent trials, the previous mask contrast from a randomly chosen staircase was re-used without being updated. We computed individual mask contrasts for the WM task by taking the grand average of the last four switches (i.e., from seen to unseen or vice versa) across the two staircases.

4.5.4 Experimental protocol

Each experimental session began with written and verbal instructions for all tasks. Subjects then performed either 60 (behavioral experiment; 1 block) or 90 training trials (MEG experiment; 2 blocks) of the WM paradigm. In contrast to the main experiment, during this training session, the target stimulus was always visible (mask set to the lowest contrast possible) and visual feedback on localization and rotation performance was provided at the end of each trial (2.5 s): The target location, connected by a white arc to the correct response position (in green ink), was displayed. If the participant had answered incorrectly, this location was also shown in red ink. Following the training, participants completed the calibration and WM task. While the former was comprised of 125 trials (1 block) in the behavioral and 120 trials (1 block) in the MEG experiment, the latter consisted of 180 (2 blocks; 2 repetitions of each of the three rotation conditions/location) and 450 trials (10 blocks; 5 repetitions of each of the three rotation conditions/location), respectively.

4.5.5 Behavioral analyses

We followed our previous approach (Trübutschek et al., 2017) to evaluate working memory performance as a function of subjective visibility. Repeated-measures analysis of variance (ANOVA) was applied to three indices of objective performance: (1) Accuracy refers to that proportion of trials that falls exactly onto the correct response location and serves as a crude measure of the amount of information which can be maintained and manipulated in working memory. Chance performance corresponds to 1/24 (i.e. 4.17%). (2) The rate of correct responding also reflects the quantity of information held in working

memory, but is more refined than accuracy alone, as it allows accounting for small errors in subjects' ability to identify the correct response location. It was defined as the proportion of trials within ± 2 positions of the correct response location (i.e., ± 30°), leading to a chance-level of 5/24 (i.e., 20.83%). (3) As an estimate of the precision of working memory representations, we computed the standard deviation of that part of the distribution of participants' spatial responses that corresponded to genuine working memory (as opposed to random guessing within the region of correct responding; Trübutschek et al., 2017). Only subjects with sufficient blindsight (i.e., $p < .05$ in a χ^2-test against chance) when collapsing across all experimental conditions were included in this analysis.

4.5.6 MEG ACQUISITION, PREPROCESSING, AND DECOMPOSITION

We installed participants inside an electromagnetically shielded room and recorded their brain activity continuously during the WM paradigm with a 306-channel, whole-head magnetometer by Elekta Neuromag® (Helsinki, Finnland). MEG sensors were arranged in 102 triplets, comprised of one magnetometer and two orthogonal planar gradiometers, and MEG signals were acquired at a sampling rate of 1000 Hz with a hardware bandpass filter between 0.1 and 330 Hz. To allow for offline rejection of artifacts induced by eye movements and heartbeat, we monitored these bodily functions with vertical and horizontal electro-oculograms (EOGs) and electrocardiograms (ECGs). Subjects' head position inside the MEG helmet was inferred at the beginning of each run with an isotrack Polhemus Inc. system from the location of four coils placed over frontal and mastoïdian skull areas.

We adapted Marti and colleagues' (2015) preprocessing pipeline. First, we identified bad MEG channels visually in the raw signal and then employed MaxFilter software (ElektaNeuromag®, Helsinki, Finland) to (1) compensate for head movements between experimental blocks by realigning all data to the head position of the first run and (2) apply the signal space separation algorithm (Taulu et al., 2004) to suppress magnetic interference from outside the sensor helmet and interpolate bad channels. We then switched to Fieldtrip for further preprocessing (Oostenveld et al., 2011). Continuous data were first epoched with respect to target onset (i.e, -0.5 to 3.5 s). The resulting trials were downsampled to 250 Hz, and any artifacted epoch removed by means of a semi-automatic procedure: We visually inspected scatter plots of the trial-wise variance of the MEG signals across all sensors to identify and reject contaminated epochs. In a last step, we performed independent component analysis (ICA) separately for each channel type to remove any residual artifacts related to eye movements or cardiac activity: Topographies of the first 30 components were displayed for visual inspection, their time courses correlated with the EOG/ECG signals, and contaminated components subtracted from the MEG data.

Depending on the nature of the subsequent investigation, further preprocessing steps then diverged. For any univariate analysis based on evoked responses (i.e., ERFs), we only low-pass filtered the MEG signal at 30 Hz. However, to extract the spectral component of our data, we relied on unfiltered epochs: Power estimates between 1 and 99 Hz (in 2 Hz steps) were obtained by convolving overlapping segments of the data with a frequency-independent Hann taper (window size: 500 ms, step size: 20 ms). Multivariate analysis required additional downsampling of the signal to 125 Hz. After all necessary transformations and decompositions, we applied a baseline correction prior to any analysis between -200 and 0 ms.

4.5.7 ESTIMATING CHANCE-FREE BRAIN ACTIVITY FOR UNSEEN CORRECT TRIALS

To account for chance-responding on unseen correct trials, we employed a strategy developed by Lamy and colleagues (2009) and first calculated the proportion of unseen correct trials correctly responded to by chance separately for each subject:

(1) $P_{UC} = ((1 - r) / (19/24)) * (5/24)$,

where $P_{UC} = \%UnseenCorrect_{Chance}$ and r = rate of correct responding.

We then estimated brain activity on the unseen correct trials reflecting chance-free responding, operating under the assumption that the actual observed amplitude *A* was a linear combination of genuine blindsight and random guessing:

(2) $A(UnseenCorrect_{Observed}) = P_{UC} * A(\%UnseenCorrect_{Chance}) + (1 - P_{UC}) * A(\%UnseenCorrect_{ChanceFree})$

(3) $A(UnseenCorrect_{ChanceFree}) = [A(UnseenCorrect_{Observed}) - P_{UC} * A(\%UnseenIncorrect_{Observed})] / (1 - P_{UC})$, assuming that $A(UnseenCorrect_{Chance}) = A(\%UnseenIncorrect_{Observed})$.

Similarly, we then reverted the process, mixing activity from seen trials with that from unseen incorrect trials, to obtain an estimate of what brain activity might have looked like under the miscategorization hypothesis.

(4) $A(UnseenCorrect_{Miscategorized}) = (1 - P_{UC}) * A(Seen_{Observed}) + P_{UC} * A(\%UnseenIncorrect_{Observed})$.

4.5.8 Source reconstruction

Structural magnetic resonance (MR) scans were available for 29 of our 30 subjects, having been acquired as part of previous experiments from our lab with a 3D T1-weighted spoiled gradient recalled pulse sequence (voxel size: 1 * 1 * 1 mm; repetition time [TR]: 2,300 ms; echo time [TE]: 2.98 ms; field of view [FOV]: 256 * 240 * 176 mm; 160 slices). To identify the anatomical locations of the MEG signals in these participants, we first segmented subjects' T1 images into gray/white matter using FreeSurfer (https://surfer.nmr.mgh.harvard.edu/) and then reconstructed the cortical, scalp, and head surfaces in Brainstorm (Tadel et al., 2011). Co-registration between the anatomical scans and the MEG data was based on participants' head position in the MEG helmet, recorded and tracked throughout the entire experiment. Subject-specific forward models relied on analytical models with overlapping spheres. Separately for each condition and participant, we modeled neuronal current sources with a constrained weighted minimum-norm current estimate (wMNE; depth-weighting factor: 0.5). Noise covariance matrices were computed from ~5 min-long empty-room recordings, measured immediately after each individual subject. Prior to group analysis, single-trial source estimates were either (1) averaged within each subject and condition, transformed into *z*-scores relative to our pre-stimulus baseline (-0.2 – 0 s), rectified, and spatially smoothed over 5 mm, or (2), in the case of time-frequency decompositions, transformed into average power in the alpha (8 – 12 Hz) and low (13 – 20 Hz) as well as high beta (20 – 27 Hz) bands with complex Morlet wavelets (Brainstorm default parameters). We then computed the contrasts of interests and projected the resulting participant-specific source estimates on a generic brain model built from the standard template of the Montreal Neurological Institute (MNI). Group averages for spatial clusters of at least 50 vertices and thresholded at 50% of the maximum amplitude are shown for each time window under consideration (cortex smoothed at 60%).

4.5.9 Multivariate pattern analysis (MVPA)

In this set of analyses, we aimed at predicting the identity and/or value of a specific categorical (i.e., visibility, accuracy) or circular (i.e., target, pre-rotation, or response location) variable (*y*) from single-trial brain activity (*X*) separately for each participant and time point. Relying on the Scikit-Learn package (Pedregosa et al., 2011) for MNE 0.15 (Gramfort, 2013; Gramfort et al., 2014), we therefore adapted the pipeline developed by King and colleagues (2016) to (1) fit a linear estimator *w* to a training subset of *X* (X_{train}) to isolate the topographical patterns best differentiating our experimental conditions, (2) predict an estimate of *y* ($\hat{y}$) from a test set (X_{test}), and (3) compare the resulting predictions to the true value of *y* either for the entire set of labels (*score(y, ŷ)*) or a specific subset (*subscore(y, ŷ)*).

Here, two main classes of estimators were used: A linear support vector machine (SVM) was employed in the case of categorical, and a combination of two ridge regressions in the case of circular data. Whereas the former was set to generate a continuous output in the form of the distance between the hyperplane

(*w*) and the respective sample of *y*, the latter first separately fit the sine (*sin(y)*) and cosine (*cos(y)*) of the spatial position in question and then estimated an angle from the arctangent of the individual predictions ($\hat{y} = arctan2(\hat{y}_{sin}, \hat{y}_{cos})$). To increase the number of instances available for each circular label, we averaged neighboring spatial locations (effectively reducing the number of positions from 24 to 12). Prior to model fitting, all channel-time features (*X*) were *z*-score normalized, and, for any analysis involving SVMs, a weighting procedure applied to counteract the effects of potential class imbalances. All other model parameters were left with their Scikit-Learn default values.

To avoid overfitting, we embedded this sequence of analysis steps in a 5-fold, stratified cross-validation procedure: For non-independent training and test sets, estimators were iteratively fitted on $4/5^{th}$ of the data (X_{train}) and generated predictions for the remaining $1/5^{th}$ (X_{test}). By contrast, when generalizing from one task to the other (i.e., no-rotation to rotation condition), estimators from each training set were directly applied to the entire test set and the respective predictions averaged. Within the same cross-validation loop, we also evaluated time generalization (King and Dehaene, 2014): Each estimator was first trained at time *t* and then tested at all other time points, resulting in a square matrix of training time x testing time. As such, this temporal generalization analysis permits an interrogation of the durability and stability of patterns of brain activity.

We summarized within-participant, across-trial decoding performance of categorical data with the area under the curve (AUC), presenting an unbiased measure of the true-positive rate as a function of the false-positive rate (range: 0 – 1; chance = 0.5). Two different summary statistics were used for circular decoding: (1) For non-directional analyses, the mean absolute difference between the predicted ($\hat{y}$) and actual angle (*y*) across all trials was first computed (range: 0 – π; chance = $\frac{\pi}{2}$), and this "error metric" was then transformed into an "accuracy score" (range: $-\frac{\pi}{2}$ to $\frac{\pi}{2}$; chance = 0). (2) In contrast, the probability distribution of the signed difference between $\hat{y}$ and an actual location was retained for directional analysis (i.e., tracking the rotation itself). The resulting, continuous angular distance estimates were then assigned to 1 of 24 evenly spaced bins (discontinuous; range: [-π, : π/24 : π]) and the probability of a given estimate falling within the range of a given bin was calculated across trials.

4.5.10 Statistical analysis

All statistics reported in the text refer to group-level analyses. In the case of ERF and frequency data, we (1) performed cluster-based, non-parametric *t*-tests with 1,000 Monte Carlo permutations to identify significant spatio-temporal differences between experimental conditions, while simultaneously correcting for multiple comparisons (Maris and Oostenveld, 2007), and (2) additionally present uncorrected outcomes of non-parametric signed-rank tests for follow-up analyses of amplitude/power differences in time courses ($p_{uncorrected} < .05$). We again relied on the above cluster-based permutation analysis to assess multivariate decoding performance (i.e., categorical data: AUC > 0.5; circular data: rad > 0; 5000 permutations). Temporal averages over five a-priori time bins, corresponding to an early perceptual period (0.1 – 0.3 s), the P3b time window (0.3 – 0.6 s), the maintenance period before (0.6 – 1.76 s) and after the cue (1.76 – 3.26 s), as well as the response (3.26 – 3.5 s), are also provided. Bonferonni correction was applied to these a-priori analyses to correct for multiple comparisons ($p_{corr} < .05/5$). When appropriate, we present circular statistics and computed Bayesian statistics based on two- or one-sided *t*-tests ($r = .707$; Rouder et al., 2009).

4.6 Acknowledgements

This work was funded by INSERM, CEA, Collège de France, ERC, and Fondation Roger de Spoelberch. D.T. was funded by a graduate fellowship from the Ecole des Neurosciences de Paris (ENP) and Fondation Schneider Electric. We gratefully acknowledge Valentina Borghesani, Pedro Pinheiro Chagas, and Fosca Al

Roumi for their invaluable daily support and stimulating discussion and specifically thank Theofanis I. Panagiotaropoulos for helpful comments on a previous version of this manuscript.

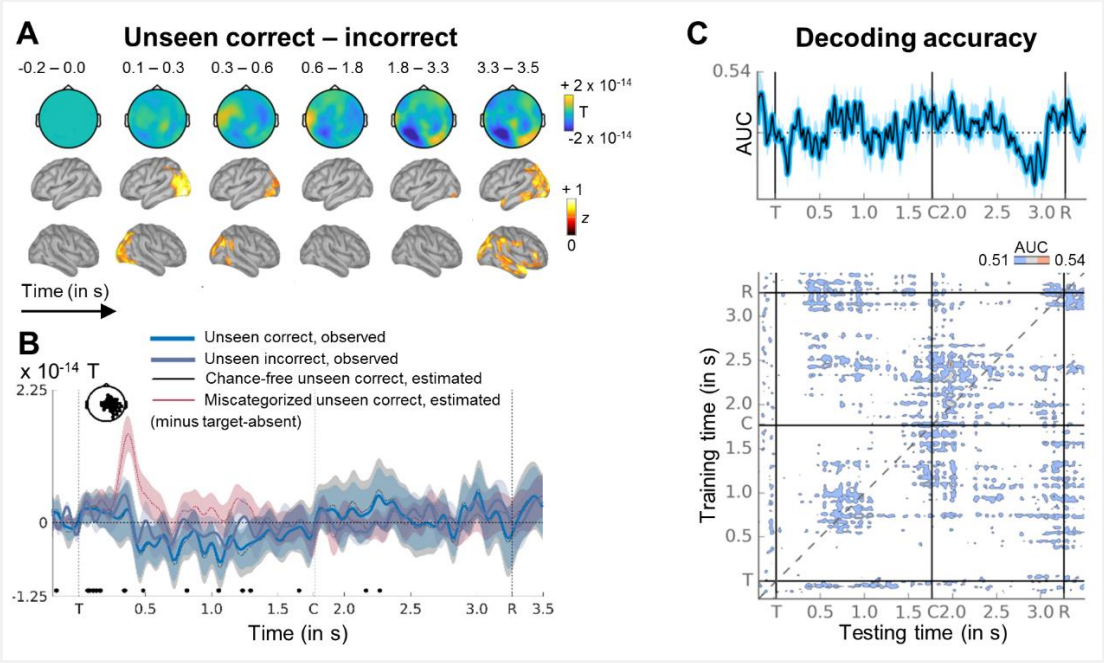
4.7 TABLES

Experiment	No rotation			Combined rotation			Clockwise rotation			Counter-clockwise rotation		
	M±*SE*	*t*(df)	*p*	*M*±*SE*	*t*(df)	*p*	*M*±*SE*	*t*(df)	*p*	*M*±SE	*t*(df)	*p*
Behavior	26.2±4.6%	4.8 (22)	**< .001**	14.2±2.1%	4.8 (22)	**< .001**	12.5±2.2%	3.7 (22)	**< .001**	16.0±2.6%	4.5 (22)	**< .001**
MEG	15.4±1.5%	7.3 (29)	**< .001**	9.2±1.1%	4.3 (29)	< **.001**	9.0±1.3%	3.8 (29)	< **.001**	9.2±1.3%	3.8 (29)	< **.001**

Table 4.1. Summary statistics for long-lasting blindsight effect.

We display the mean (*M*) and standard error (*SE*) for accuracy on the unseen trials as a function of experiment and rotation condition. *T*-statistic refers to a one-sample test against chance (i.e., 4.17%). Bold numbers indicate significant above-chance localization performance (one-tailed). df = degrees of freedom.

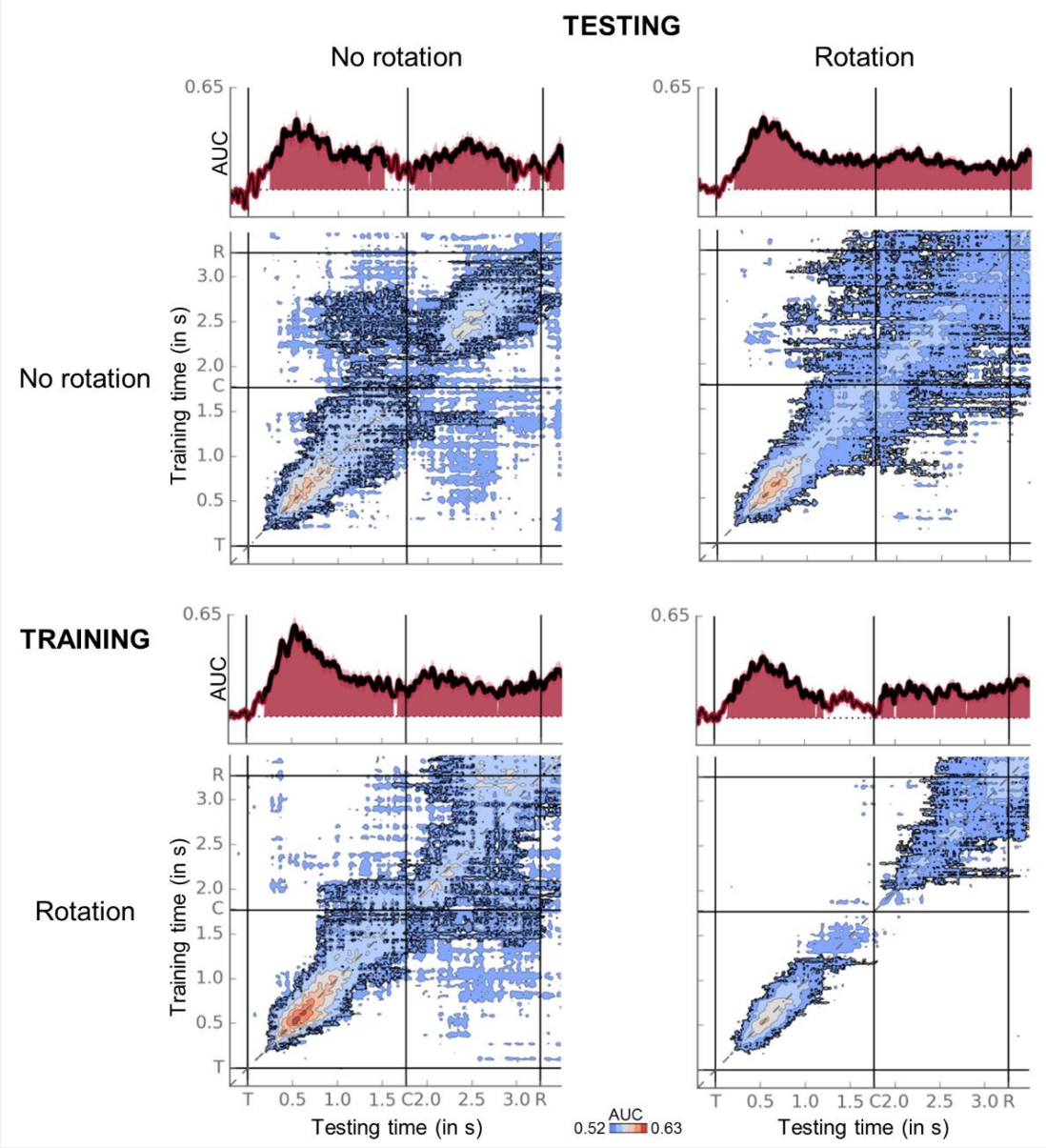
4.8 Supplementary figures



Supplementary Figure 1

No signatures of conscious processing on the unseen correct trials.

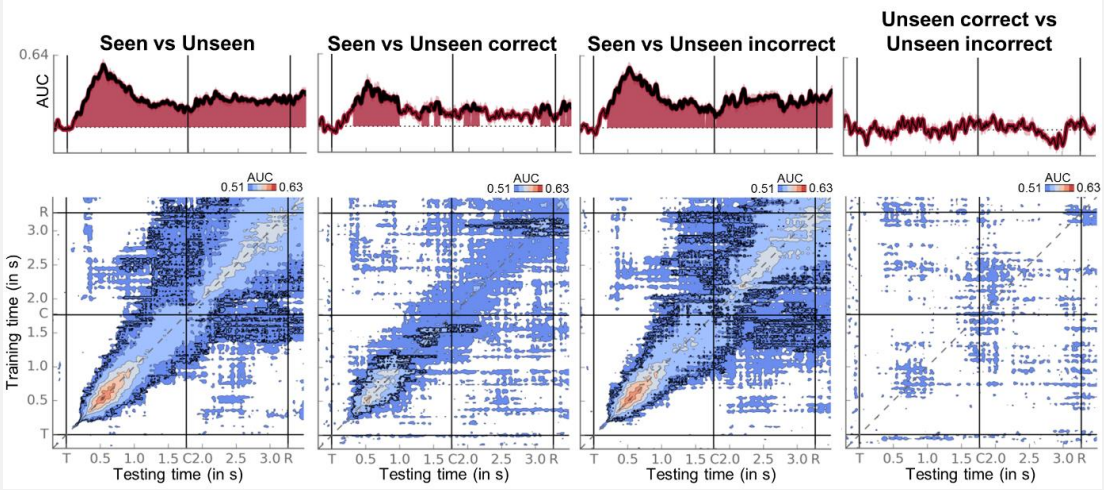
(A) Sequence of brain activations (-0.2 – 3.5 s) evoked by non-consciously perceiving the target in both tasks in sensor (top) and source space (bottom). Each topography depicts the difference in amplitude between unseen correct and unseen incorrect trials averaged over the time window shown (magnetometers only). Sources reflect *z*-scores of absolute difference with respect to a pre-stimulus baseline.
(B) Average time courses (-0.2 – 3.5 s) of unseen correct (light blue) and unseen incorrect (dark blue) trials in that subset of magnetometers having shown a significant difference in amplitude between seen and unseen targets. Black trace reflects brain activity on the unseen correct trials after having been corrected for chance-responding. Red time course illustrates what the signal on the unseen correct epochs should have looked like, had the miscategorization hypothesis been true. Shaded area denotes standard error of the mean (SEM) across subjects. Significant differences between unseen correct and incorrect epochs are depicted with the thick, black line (two-tailed Wilcoxon signed-rank test, uncorrected). Vertical dotted lines index onset of the target (T), symbolic cue (C), and response (R) screens. For display purposes only, data were lowpass-filtered at 8 Hz.
(C) (Top) Average time course of diagonal decoding of accuracy on the unseen trials (i.e., unseen correct vs. unseen incorrect). Horizontal, dotted line represents chance level at 50%. (Bottom) Temporal generalization matrix of the same accuracy decoder. Each horizontal row in the matrix corresponds to an estimator trained at time *t* and tested on all other time points *t'*. The diagonal gray line demarks classifiers trained and tested on the same time points (i.e., the diagonal estimator shown on top). In both plots, vertical lines mark onset of the target (T), symbolic cue (C), and response (R) screens. Only for display purposes, data were smoothed with a moving average of 5 samples (i.e., 40 ms). AUC = area under the curve.



SUPPLEMENTARY FIGURE 2

CONSCIOUS PERCEPTION ENTAILS SIMILAR NEURAL DYNAMICS IN BOTH TASKS.

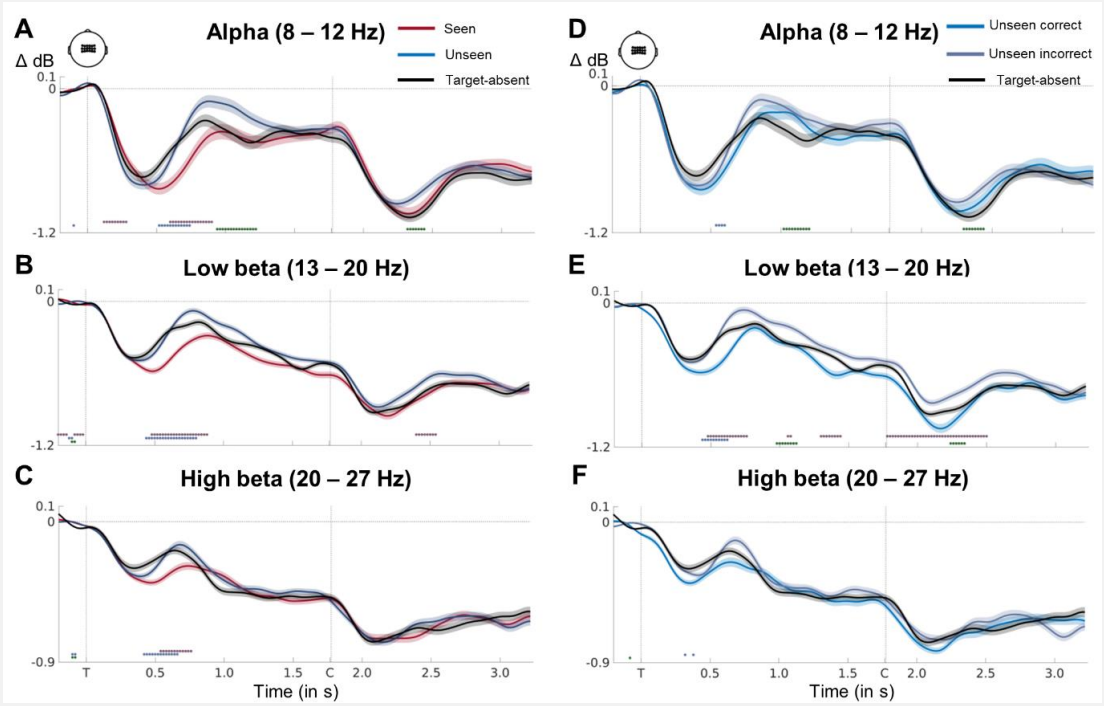
Temporal generalization matrices (bottom) for decoding of visibility category (i.e., seen vs. unseen) as a function of training and testing task (i.e., no rotation vs. rotation). In each panel, a classifier was trained at every time sample (y-axis) and tested on all other time points (x-axis). The diagonal gray line demarks classifiers trained and tested on the same time sample. Event markers (i.e., vertical/horizontal lines) denote onset of the target (T), cue (C), and response (R) screens. Time courses of diagonal decoding are shown on top. Black outlines in matrix plots and thick lines/shaded areas in time courses show periods of significant decoding (cluster-based permutation test, two-tailed except for diagonal). For display purposes, data were smoothed using a moving average with a window of 5 samples (i.e., 40 ms). AUC = area under the curve.



SUPPLEMENTARY FIGURE 3

COMPARING VISIBILITY TO ACCURACY DECODER.

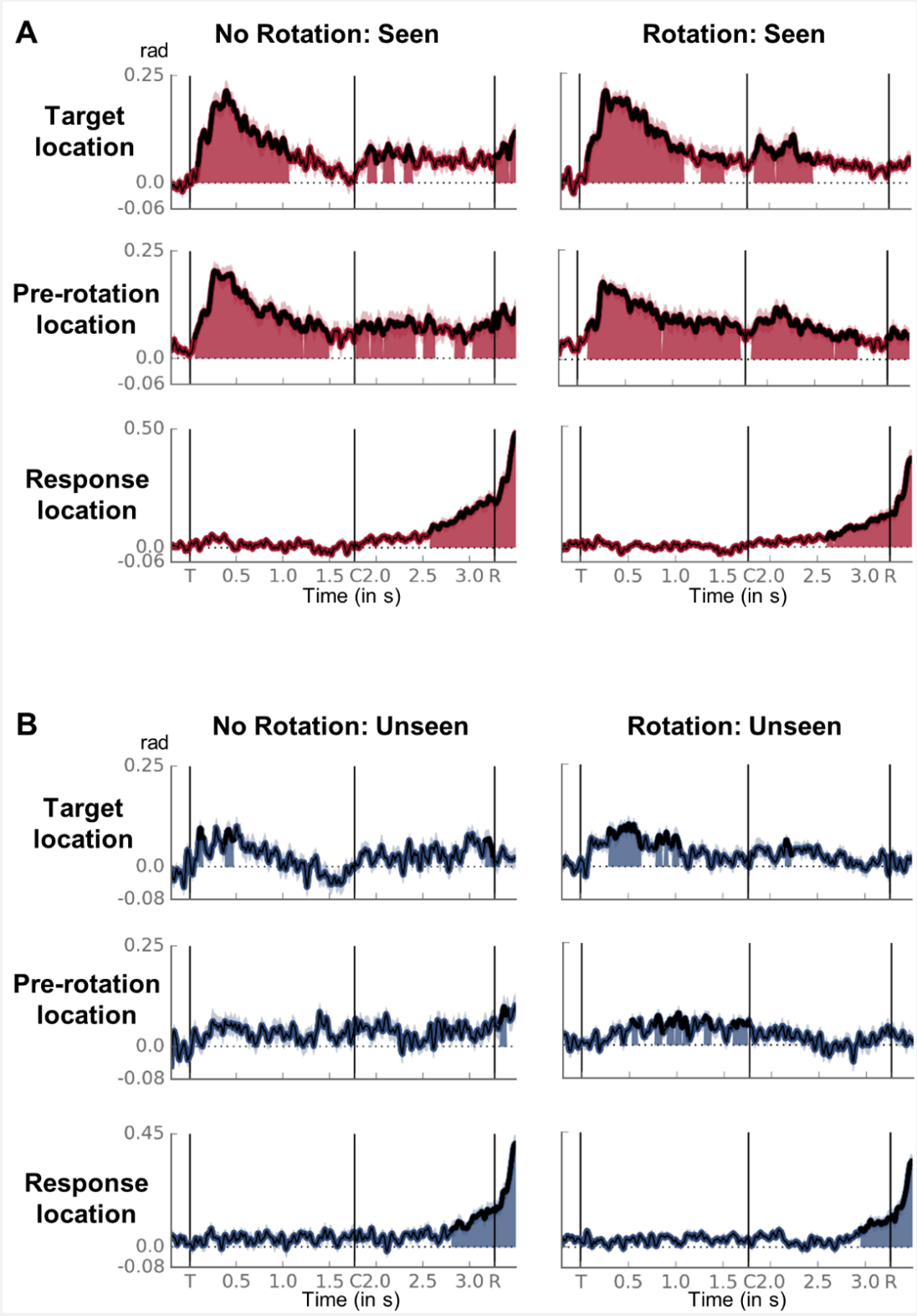
Each panel displays the generalization matrix (bottom) and time course of diagonal decoding (top) of a specific visibility or accuracy estimator. Horizontal, dotted line in time course represents chance level at 50%. Each horizontal row in the matrix corresponds to an estimator trained at time *t* and tested on all other time points *t'*. The diagonal gray line demarks classifiers trained and tested on the same time points (i.e., the diagonal estimator shown on top). In both plots, vertical lines mark onset of the target (T), symbolic cue (C), and response (R) screens. Thick lines/shaded areas as well as black outlines denote above-chance decoding as assessed by a cluster-based permutation test (two-tailed, with the exception of the diagonal). Only for display purposes, data were smoothed with a moving average of 5 samples (i.e., 40 ms). AUC = area under the curve.



Supplementary Figure 4

Average time courses of alpha, low beta, and high beta power.

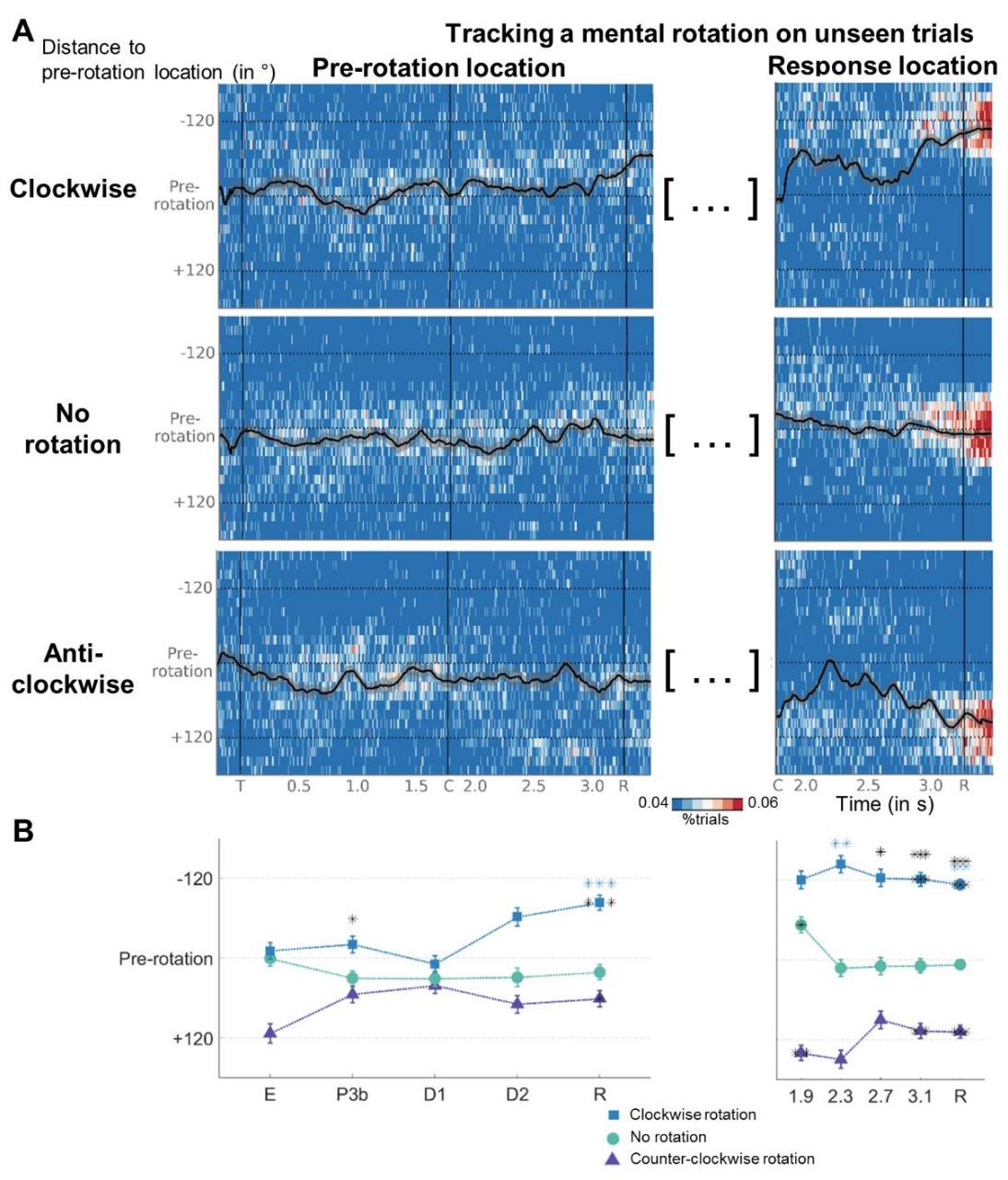
Time courses of average alpha (8 – 12 Hz; **A**), low beta (13 – 20 Hz; **B**), and high beta (20 – 27 Hz; **C**) band activity in a group of central sensors as a function of visibility and target presence. Shaded area demarks standard error of the mean (SEM) across subjects. Thick lines represents significant difference in power between conditions (red = seen vs. unseen; blue = seen vs. target-absent; green = unseen vs. target-absent; two-tailed Wilcoxon signed-rank test across subjects, uncorrected). Vertical line demarks onset of target (T) and cue (C) screens. **(D-F)** Same as in (A-C), except for unseen correct and unseen incorrect trials. Color code for significant differences is as follows: red = unseen correct vs. unseen incorrect, blue = unseen correct vs. target-absent, green = unseen incorrect vs. target-absent.



Supplementary Figure 5

Tracking a mental rotation on seen and unseen trials in the rotation and no-rotation task.

(A) Time courses of average decoding of target location (top), pre-rotation location (middle) and response location (bottom) on seen trials as a function of task (i.e., no rotation vs. rotation). Thick lines and shaded areas represent above-chance performance as assessed by a one-tailed cluster-based permutation test. Horizontal dotted lines index chance. Event markers denote the onset of the target (T), cue (C), and response (R) screens. For illustration purposes, data were smoothed with a moving average of 5 samples (i.e., 40 ms).
(B) Same as in (A), but for unseen trials.



SUPPLEMENTARY FIGURE 6

TRACKING A MENTAL ROTATION ON UNSEEN TRIALS.

(A) (Left) Time courses of probability density distributions of the angular distance between the estimates of a decoder trained with pre-rotation angle and actual pre-rotation location are shown as a function of rotation condition. For display purposes only, data were smoothed with a moving average of 12 samples (i.e., 96 ms). Overlaid black line illustrates the evolution of the circular mean of the individual distributions (also smoothed). Shaded area reflects circular standard variation across subjects. Vertical event markers denote the onset of the target (T), cue (C), and response (R) screens, horizontal markers index correct response positions after rotation. (Right) Same as in the left panels, except for angular distance between the estimates of a decoder trained with response angle and actual pre-rotation location.
(B) Circular means of the above distributions as a function of rotation condition and time bin (i.e., E = 100 – 300 ms, P3b = 300 – 600 ms, D1 = 0.6 – 1.76 s, D2 = 1.76 – 3.26 s, R = 3.26 – 3.5 s). Error bars reflect circular standard deviation. Asterisks inside markers denote significant deviation from mean direction of 0 (as assessed by a circular equivalent of a one-sample *t*-test), asterisks on top significant differences in median direction between conditions (as assessed by a circular equivalent to the Kruskal-Wallis test; black = clockwise vs. counter-clockwise; blue = clockwise vs. no rotation; violet = counter-clockwise vs. no rotation). * p < .05, ** p < .01, *** p < .001.

CHAPTER 5 –
GENERAL DISCUSSION AND PERSPECTIVES

Science is a struggle for truth
aainst methodological, psychological, and sociological obstacles.
- FANELLI AND IOANNIDIS (2013)

5.1 SUMMARY OF THE MAIN FINDINGS

An assumption that permeates nearly all aspects of our work and thinking as psychologists and cognitive neuroscientists is the idea that conscious perception and working memory are virtually indistinguishable. Many of us may not even be explicitly aware of this belief any more, clearly attesting to just how deeply engrained this premise has become in our culture. In Chapter 1, we started our journey with my attempt to convey to you as to why exactly this particular view may be so appealing. We have seen how both cognitive and neurobiological models of consciousness and working memory are deeply intertwined, featuring the short-term maintenance of information in a highly distributed network of brain areas in the service of complex behavior, and how both cognitive functions share many other central characteristics and properties. Yet I also introduced you to some very recent behavioral and, to a lesser extent, neuroimaging evidence that challenges these prevailing assumptions. In some circumstances, working memory appears to operate implicitly, outside the realms of conscious awareness, and the contents of working memory may also be dissociated from the ones of current, conscious experience.

What should we make of such findings? How can they be integrated into our current framework? And what do they tell us about the nature of visual awareness and working memory? My goal for this thesis was to shed some light on these questions and begin to unravel the nature and neuro-cognitive architecture underpinning such a long-lasting blindsight effect. I began this endeavor with a thorough evaluation of alternative explanations for the phenomenon at hand. In Chapter 2, I used MEG, time-resolved multivariate decoding, and computational modeling to assess the long-lasting blindsight in light of two particularly important alternate hypotheses: Could the observed, long-lasting blindsight effect just have resulted from subjects' accidental miscategorization of a small number of seen trials as unseen (i.e., miscategorization hypothesis)? Or could it potentially reflect the conscious maintenance of an early guess, with participants committing to a response right after the presentation of the non-conscious target and then holding onto this guess consciously (i.e., conscious maintenance hypothesis)?

The data suggested that this was unlikely. Brain responses to seen targets were characterized by a sustained desynchronization in the alpha and beta frequency bands, and target location itself could be tracked throughout the entire epoch. By contrast, all of these markers of conscious processing vanished for unseen targets. Even on the blindsight (i.e., unseen correct) trials, there was no evidence for any power suppression and, after an initial encoding period, content-specific delay-period activity disappeared altogether. As such, these findings show that different neural processes subtend the storage of seen and unseen correct/incorrect targets, implying that these two groups of trials indeed stem from qualitatively different conditions. They do, however, also raise a further question: How, if not through sustained brain activity, could information have been maintained at all? Modeling the behavioral paradigm within the recent framework of dynamic, activity-silent working memory revealed that target location may have been stored with a mixture of activity-based and activity-silent brain states. While unseen targets seem to have been maintained exclusively via activity-silent mechanisms, seen targets primarily relied on neural firing interspersed with activity-silent periods.

Taken together, these findings suggest that the long-lasting blindsight effect results from a genuinely non-conscious process that cannot simply be explained by any of the alternative hypotheses. But, what exactly, then, is its relation to working memory? In order to tackle this question, I next set out to systematically probe key features of conscious working memory in the context of long-lasting blindsight. In a first behavioral experiment, presented in Chapter 3, I evaluated the possibility of storing multiple items in addition to temporal-order information non-consciously. Prior research on non-conscious working memory had primarily examined the storage of single, visual items (e.g., orientation, spatial location, alphanumeric characters, etc.) or features that, although supposedly independent, likely were automatically bound into a unified representation (i.e., object in a given spatial position). Especially in light of the known capacity limits for conscious working memory, it was therefore important to establish whether non-conscious storage is limited to a single, sensory representation, or whether it may also accommodate multiple items and order information.

The data supported the latter possibility. I, once again, observed a robust, long-lasting blindsight effect, whose properties were unaffected by the number of targets to be retained (i.e., one vs. two) as well as by the specific combination of subjective visibility (i.e., both unseen, first target seen/second target unseen, first target unseen/second target seen). Moreover, there was no evidence for swapping errors, suggesting that, even when none of the targets had been seen, their serial order could nevertheless be stored. Crucially, all of these results are compatible with the proposal of non-conscious maintenance based on activity-silent mechanisms. Non-consciously storing temporal order therefore appears to be within the realm of possibilities.

What does seem to co-occur with prior access to consciousness, in contrast, is the manipulation of maintained representations. In Chapter 4, I presented the outcomes of the last two studies I ran as part of this thesis. Here, I again employed behavioral techniques in conjunction with time-resolved MEG recordings and multivariate pattern analysis in order to confront the long-lasting blindsight effect with an additional component: a mental rotation. As before, subjects were asked to retain the location of a masked target square over a long delay. However, halfway throughout the maintenance period, a visible cue instructed them as to whether (1) simply retain the original location, or (2) rotate it 120° clockwise or counter-clockwise. Strikingly, in two independent experiments, the long-lasting blindsight effect persisted even in the face of the rotation task. The MEG recordings, however, revealed that, in stark contrast to the first experiment, this time, the blindsight effect appeared to be the result of an effortful, conscious rotation. Around the time of the presentation of the rotation cue, any pre-existing differences in oscillatory brain response between seen and unseen targets vanished. Alpha and beta power was similarly suppressed during the rotation period for both trial types. Moreover, target, pre-rotation, as well as response location, and possibly even intermediate representations, could be decoded during similar time periods for both seen and unseen targets. As such, similar processes seem to have been at work on the seen and unseen trials, suggesting that, following an unseen target, subjects guessed a position and then consciously rotated this guess. At least with the experimental setup we had chosen, manipulating information held in working memory therefore appeared to have required both conscious access and neural activity.

5.2 Non-conscious … wait, what?

I think the time may finally have come. Let us start talking about that elephant in the room. For the past 100 pages or so, we have discussed a phenomenon that I have dubbed long-lasting blindsight. You have seen that this effect is robust and replicable: It may last for at least 15 s (Bergström and Eriksson, 2014; Bergström and Eriksson, 2015, 2017), is able to withstand visible distractors (Bergström and Eriksson, 2014; Soto et al., 2011; Trübutschek et al., 2017), appears to be modulated, though not abolished, by a large conscious working memory load (Trübutschek et al., 2017), yet is, at the same time, resistant to a minimal non-conscious load (Chapter 3). Crucially, it seems to result from genuine, non-conscious

maintenance of information (Trübutschek et al., 2017), potentially even recruiting prefrontal brain areas in some capacity (Bergström and Eriksson, 2014; Bergström and Eriksson, 2017; Dutta et al., 2014). If this does not sound like working memory to you, then I do not know what does. On the other hand, I have also shown you that it does not depend on sustained delay-period activity (Trübutschek et al., 2017) and does not appear to be able to accommodate a genuinely non-conscious manipulation of information (though it may serve as an input for a conscious transformation; Chapter 4). The question that must have been bugging you for a while now therefore surely is: What is this long-lasting blindsight effect? Is it fair to call it non-conscious working memory? Or might it better be called something else?

Based on the traditional definition of working memory as a memory system "for the execution of our Plans" (Miller et al., 1960) and based on the evidence presented in this dissertation and elsewhere, long-lasting blindsight appears to be difficult to reconcile with a veritable *working* memory. Even if, in some instances, non-conscious maintenance may be tied to prospective use (e.g., Pan et al., 2014), it does not consistently (as genuine working memory should) permit the flexible use, transformation, and manipulation of currently stored representations. When asked to mentally rotate a non-conscious target location, participants may have performed above chance, but this seems to have been the result of a conscious rotation of a guess (Chapter 4). This inability to perform a non-conscious transformation of information is, in fact, compatible with prior research (van Gaal et al., 2014; Mudrik et al., 2014). Sackur and Dehaene (2009), for example, demonstrated that non-conscious processing failed in the face of a composite task. Here, participants first had to perform a simple arithmetic operation on a masked digit and then compare the result to five. While each of the individual operations could proceed in the absence of subjective, conscious experience, performance on this serial, chained task was at chance. Conscious access therefore appears to be required specifically for those types of operations that depend on working memory: multi-step rule-based algorithms and combinatorial processes.

If I have not been looking at working memory in its traditional sense, then what have I been studying for all of these years? Let us consider some of the other types of short-term memory we have talked about before. Iconic memory might perhaps constitute the most obvious alternative. A high-capacity, rapidly decaying store, it is thought to retain visual representations in a high-fidelity, literal format in the order of several hundred milliseconds (i.e., < ~1 s Neisser, 1967; Sperling, 1960). Even if we completely set aside the differences in durability of iconic memory representations and the ones associated with the long-lasting blindsight effect (i.e., up to at least 15 s), there are reasons arguing against it being purely iconic memory. On one hand, I obtained no evidence in favor of any temporal decay of the blindsight effect (Trübutschek et al., 2017). The duration of the delay period in the first MEG and the first behavioral experiment presented in Chapter 2 varied between 2.5 and 4, and between 0 and 4 s, respectively. If the long-lasting blindsight effect constituted some sort of prolonged iconic memory, then, especially in the behavioral study, I should have observed a decrease in non-conscious performance after a long delay. Yet both the amount of information as well as the precision with which it could be maintained were unaffected by delay. On the other hand, iconic memory representations are also very fickly, being easily erased and overwritten by subsequent stimulation. Sperling (1960) himself, for instance, showed that, when the stimulus array in his partial-report paradigm was followed by a uniform flash of light, subjects' performance was reduced by half. More recent studies replicated these early findings of a disruptive effect of backward masking on iconic memory representations (Tijus and Reeves, 2004). As such, the simple presence of the mask in all of my experiments should already have sufficed to interfere with iconic memory processes.

A second alternative to working memory might be fragile visual short-term memory. Sligte and colleagues (2008) combined a change detection paradigm with a retro-cue, such that they first presented their participants with a quickly flashed array of oriented bars and, then, after a variable delay, cued the spatial location at which the change between the stimulus and probe array might happen. When the delay between this retro-cue and the stimulus array was short (i.e., 10 ms), they observed features typical of iconic memory: a very high-capacity store that was, however, quickly decaying and overwritten by a light

mask. Similarly, when the cue coincided with the presentation of the probe array (i.e., 100 ms after onset of the probe array), they reported the typical findings for a capacity-limited, yet durable working memory store. What these authors also argue, however, is that there exists yet another type of memory, fragile visual short-term memory, intermediate in durability and capacity between iconic and working memory. When the retro-cue was presented during the delay period (i.e., between 1 and 4 s after the stimulus array), participants' capacity appeared to be twice as high as their "normal" working memory capacity, and the stored representations seemed to be erased only by a pattern mask, but not by a light mask. The same group of authors later reported differential activation in V4 as a function of memory status, with items supposedly held in fragile visual short-term memory associated with lower activity than representations in working memory (Sligte et al., 2009). Most importantly for the intents and purposes of our current discussion, however, they also demonstrated that, these fragile visual short-term memories were indeed effectively erased by intervening stimuli in the same spatial location (very broadly defined) and of the same category as the memoranda (Pinto et al., 2013). Even when leaving aside the discussion of whether or not this fragile visual short-term memory genuinely exists as a separate system or is perhaps best described as the contents of working memory pre attentional selection, it seems unlikely to have accounted for the long-lasting blindsight effect I and others have observed. A recurring feature in almost all of the experiments conducted on this phenomenon so far is that it appears resistant to distraction and intervening visual stimulation (Bergström and Eriksson, 2014; Bergström and Eriksson, 2017; Chapter 3), even when this overlaps in space and category with the to-be-remembered information (Soto et al., 2011; Trübutschek et al., 2017).

In summary, then, I have just argued that the long-lasting blindsight effect does not appear to have resulted from any of the traditional varieties of short-term memory. It neither appears to be a fully developed non-conscious working memory, nor a prolonged iconic or fragile visual short-term memory. This leaves us in a bit of a pickle. What exactly is this long-lasting blindsight effect? Do we need to introduce yet another type of short-term memory? Before going down that road, how about we first take a look at some of the neural correlates that have been proposed for it over the years. Prior to my own work, two early fMRI studies reported delay-period activity in dorsolateral prefrontal cortex, even on unseen trials (Bergström and Eriksson, 2014; Dutta et al., 2014), suggesting that, perhaps, networks typically involved in working memory may also be recruited during the long-lasting blindsight effect. We have already discussed in depth as to why, in my opinion, the conclusions that may be drawn from these experiments are limited (Chapter 1). More recent work by Bergström and Eriksson (2017) remedies some of these initial concerns and, apparently, still calls for an involvement of prefrontal areas during the non-conscious maintenance of information. Using continuous flash suppression to render their stimuli invisible for relatively long periods of time, Bergström and Eriksson (2017) demonstrated that, during the delay-period on unseen trials, presence vs. absence of the non-conscious stimulus could be decoded from a prefrontal region of interest (ROI), while location (but not identity) of the target (i.e., left vs. right) could be decoded in an occipital ROI. Here, too, however, the results seem difficult to interpret. The long-lasting blindsight effect itself was only present during the pre-fMRI session and could not be replicated during the actual fMRI experiment. As such, it is unclear how exactly the delay-period activity observed by this group relates to the phenomenon under consideration here. Moreover, it seems odd that information about spatial location should be retained in occipital cortex, when the much simpler classification of target presence vs. absence failed. While all of these studies thus point towards some role of prefrontal cortex for long-lasting blindsight, the precise nature of this involvement is not yet clear and will have to be determined in future research.

To make matters even more complicated, my own work, in conjunction with complementary findings from our group (King et al., 2016), suggests that non-conscious maintenance may not even require persistent, content-specific delay-period activity at all. When participants simply had to keep in mind a masked spatial location, content-specific delay-period activity for unseen targets vanished entirely after ~ 1 s, while being intermittent on seen trials (Chapter 2). By contrast, when subjects had to rotate the spatial

location in addition to maintaining it, participants' initial guess and response position could be tracked throughout the epoch (Chapter 4). You have already seen how these data are compatible with recent theoretical developments that propose that information may also be stored in activity-silent brain states via short-term changes in synaptic weights (Mongillo et al., 2008; Stokes, 2015). But where does this leave us with regards to our mission of unraveling the nature of the long-lasting blindsight effect and establishing an appropriate terminology?

Insofar as a common language facilitates (scientific) communication, *activity-silent short-term memory* might be a fitting descriptor for long-lasting blindsight. It would, all at once, highlight the type of cognitive function under consideration (i.e., a type of memory) and point to the proposed neural mechanism (i.e., transient changes in patterns of functional connectivity). However, it might also create artificial barriers and boundaries with regard to other mental processes (and, by consequence, scientific communities) that might not necessarily exist. Let me explain this last part in a bit more detail. Throughout this entire dissertation, you have seen that, while dissociable from working memory in terms of subjective experience long-lasting blindsight also shares many of the key characteristics of working memory. It allows for the short-term maintenance of sensory and temporal-order information (Chapters 2, 3, and 4), interacts with other items currently held in working memory (Chapter 2) and is resistant to distraction (Chapter 2). Even at the neural level, there seems to be a certain overlap between the long-lasting blindsight effect and working memory. Similar brain areas in sensory regions (Chapters 2 and 4) and prefrontal cortex (Bergström and Eriksson, 2017; Dutta et al., 2014) appear to be recruited for both phenomena and the hypothesized activity-silent mechanism is not exclusively reserved for non-conscious representations either. In Chapter 2, I have shown that even seen targets might have been stored in a mixture of activity-based and activity-silent brain states, and work from other groups suggests that, while attended items are maintained with persistent neural firing, currently unattended items might rely on activity-silent storage (Rose et al., 2016; Wolff et al., 2017). Long-lasting blindsight is thus not completely orthogonal to working memory.

If, for a moment, you consider the theoretical conceptualization of working memory, this need not be mutually exclusive or incompatible. Working memory is a generic term, referring to all those brain systems involved in the online storage of information for rapid access, transformation, and flexible use (e.g., Atkinson and Shiffrin, 1968; Baddeley and Hitch, 1974; Cowan, 1997; Miller et al., 1960). As such, many different mental processes and cognitive functions actually contribute to the successful completion of typical "working memory" tasks. Take mental arithmetic as an example. What kinds of operations and computations have to occur for you to solve this (fairly) simple addition (without a pen and paper, of course): 326 + 45 + 289? First, this information has to be encoded consciously, so you have to pay a sufficient amount of attention to it, and then you have to retrieve the corresponding long-term memory representations in order to access the meaning of all of the individual components. Next, you actually have to carry out the addition, thus sustaining your attention to the task at hand, storing and continuously updating intermediate results and inhibiting and correcting incorrect ones.

Adopting such a component-process view of working memory (Eriksson et al., 2015; Goldhill, 2018) and firmly grounding it in neurobiology may help reconcile the notion of long-lasting blindsight with working memory. Put simply, not all of the (myriad) processes that might be recruited in the name of working memory necessarily need to require access to consciousness and/or accompanying neural activity. Contemporary state-based models of working memory already acknowledge the existence of several representational states for the contents of working memory (e.g., Cowan, 1997; McElree, 2001; Oberauer, 2002, 2005). Items may either be attended, thus populating your mind, or they may be held in the activated portion of long-term memory, thereby, although currently unattended, easily shifted into the focus of attention if need be. As you have just seen, different brain states (i.e., activity-based vs. activity-silent) might correspond to these different attentional states (Rose et al., 2016; Wolff et al., 2017). From here, it does not seem like too far a jump to also integrate the notion of genuinely non-conscious representations (e.g., Bergström and Eriksson, 2017; Soto et al., 2011; Trübutschek et al., 2017). In fact, based only on the

current evidence, it is not clear whether these two distinct literatures have even focused on different phenomena at all, or whether the driving factor determining the neural fate of a stored representation relates to the amount of attention or conscious processing that information has received. Currently unattended stimuli in the above retro-cue paradigms (Rose et al., 2016; Wolff et al., 2017) are likely also non-conscious, and the attentional status of the non-conscious target in my own work is not evident either (though, the fact, that it could resist distraction seems to imply that it may have been attended).

The key question that, in my view, remains for future research is to understand exactly how these different types of representational states relate to each other, what types of brain mechanisms they may recruit in different contexts and circumstances, and what sorts of computations and operations they might support. For instance, short-term storage of information per se, irrespective of representational state, may not require accompanying content-specific neural activity (Trübutschek et al., 2017), but attentionally selecting it (Quentin et al., 2018; Rose et al., 2016; Wolff et al., 2017) or transforming it (Chapter 4, Masse et al., 2018) might. According to this perspective, then, working memory is nothing but a specific configuration of a variety of sub-processes that may be combined in different ways to solve the task at hand. Long-lasting blindsight is but one of these and may only be amenable for a certain subset of (working memory) tasks. The challenge will be to identify specifically which ones.

5.3 Should we equate consciousness with maintenance of information?

We have just spent a lot of time discussing how the notion of non-conscious short-term maintenance may be integrated with contemporary conceptualizations of working memory in particular and (short-term) memory more generally. I have argued that, as long as we employ a component-process approach, the existence of a long-lasting blindsight effect does not pose an insurmountable obstacle. Storing a representation (for a certain period of time) and experiencing it consciously may simply depend on dissociable neural mechanisms that may be used in different combinations. But how do theories of consciousness fare in light of these novel data? Is the finding of non-conscious maintenance of information compatible with these current views? I, again, believe that while updating may be needed, there exist no major incompatibilities.

You may still recall from the introductory chapter that maintenance of information plays an important role in many theories of consciousness. Representations may either be amplified and broadcast globally via sustained activity in a fronto-parietal network (Dehaene and Changeux, 2011; Dehaene et al., 1998, 2014), be maintained via recurrent feedback loops (Lamme and Roelfsema, 2000), or be retained via thalamo-cortical interactions (Tononi and Koch, 2008). Initial empirical evidence supported these models, showing that, behaviorally, non-conscious stimuli tend to stop affecting subsequent processing after just a couple hundred milliseconds (Dupoux et al., 2008; Greenwald et al., 1996), and that the brain responses following conscious stimuli are typically later, more robust, and more sustained than the ones associated with non-conscious input (e.g., Del Cul et al., 2007; Lamy et al., 2009; Polich, 2007). In stark contrast to this early work, recent reports challenged the short-lived nature of non-conscious representations, proposing that information instead may also be maintained in non-conscious working memory (Bergström and Eriksson, 2017; Soto et al., 2011).

I hope that, throughout all of this work (and especially the preceding part of the discussion), you have already seen how these divergent views may be reconciled. On one hand, my findings support the idea that non-conscious information may be maintained for much longer periods of time than previously thought (Chapters 2 and 3). While perhaps not yet integrated into contemporary theories of consciousness, evidence for such durable non-conscious representations actually also exists outside the domain of non-conscious working memory. Sergent and colleagues (Sergent et al., 2013; Thibault et al., 2016), for example, retrospectively cued their subjects' attention to a spatial location up to 400 ms after the presentation of a barely visible target stimulus and showed that, when this retro-cue coincided with the position of the target, participants' detection of the target stimulus itself improved. Similarly, Salti and

collaborators (2015) very recently tracked the fate of neural representations of consciously and non-consciously perceived stimuli. In line with current models of conscious access, these authors reported that, starting from ~ 270 ms onwards, brain responses between seen and unseen correct/incorrect trials began to diverge, with information on the former being selectively amplified and maintained for a slightly longer duration than its non-conscious counterpart. You may remember this part of the story from Chapter 1. What I did not mention at this stage was that even the unseen incorrect targets could be decoded for the entirety of the 800 ms long epoch. Recall that, in my initial study, too, I could track the location of unseen targets for ~ 1 s before activity-silent mechanisms seem to have taken over (Chapter 2).

Clearly, non-conscious information is much more durable and appears to decay much less rapidly than most theories of consciousness have acknowledged so far. At least for the range of time I and others in my community have considered, maintenance, by itself, does not seem to lead to the conscious experience of information. But, even if often depicted in such a manner, hardly any theory of consciousness focuses only on the maintenance of information. In fact, in many such models, it almost appears as a byproduct of another, overarching goal. Take the global neuronal workspace as an example. Here, consciously represented information is thought to have gained access to a mental arena, centered on a distributed fronto-parietal network, that allows it to be shared with a variety of independent processors (Dehaene et al., 2017). Part of this global broadcasting may certainly allow for the information to be retained for longer periods of time (e.g., through overt or covert rehearsal), but it seems to be the global availability of information that takes center stage here. Similarly, the integrated information theory of consciousness, though achieved by prolonged thalamo-cortical interactions, focuses on the degree of integration as its major determinant of consciousness (Tononi and Edelman, 1998; Tononi and Koch, 2008; Tononi et al., 1998). Current theories of consciousness may thus need to be partially revised in order to account for the more durable nature of non-conscious representations. One possibility here might be to decouple conscious access from sensory processing (thereby allowing for information to be stored non-consciously for a certain period of time before, potentially, crossing the threshold for conscious perception; Sergent, 2018), or to revert to a dynamic processing hierarchy, in which, in particular the durability and amplitude of late stages determine access to consciousness (King et al., 2016).

While models of conscious perception may thus fairly easily accommodate non-conscious maintenance of information, a much more problematic situation for these theories were to arise if there indeed existed a genuine non-conscious working memory in the traditional sense, with the ability to store, manipulate, transform and integrate information. We have already discussed why, in light of the evidence I have presented in this thesis (Chapter 4), this seems unlikely. Storing a masked target location could be dissociated from conscious perception (Chapters 2 and 3), but mentally rotating it appeared to co-occur with conscious access. Given the neural mechanism I proposed for the short-term maintenance of information, this divergence seems plausible. According to the activity-silent, dynamic coding framework for working memory, any input to a network will lead to a transient shift in the functional connectivity of this network, such that, when being reactivated by subsequent stimulation, the network response will be patterned according to the previous input (Stokes, 2015). For the stored information to change (i.e., to be manipulated), stimulation in the form of neural activity is therefore a necessity. Maintenance of non-conscious stimuli may therefore occur if the initial signal is able to sufficiently modify the synaptic weights of a given network, yet manipulating it requires continuous neural input and concomitant conscious access.

5.4 Limitations and outstanding questions

As is so often the case in science, I have presented a body of work that, in the end, may have led to more questions than answers. In this last section of this thesis, I want to draw your attention to some of the general limitations of the experiments I conducted as well as some of the many outstanding questions

in the hopes of inspiring future research in this domain of science. Let us start by reviewing some of the skepticism and critique that has been voiced with respect to the notion of non-conscious working memory.

Initial critiques against non-conscious working memory were primarily concerned with the alternative hypotheses I evaluated in Chapter 2. Essentially, both of these boil down to the effects observed in studies on non-conscious working memory not being the result of a genuinely non-conscious process (Stein et al., 2016). Either participants' visibility judgments were not an accurate reflection of their actual perceptual experience, due to, for instance, erroneous miscategorization of some seen trials as unseen or a general response bias towards underreporting seen trials, or they may simply have maintained a conscious guess. I have already presented extensive evidence, grounded in an analysis of the brain responses to seen and unseen correct targets, that argues against these alternatives (Chapter 2). However, in this work, too, I have relied on subjective measures to assess conscious perception so it, nevertheless, remains a possibility that some conscious processes may have contributed to the long-lasting blindsight effect. One need only assume that some of the signatures of conscious processing (e.g., P3b, alpha/beta desynchronization) may have been modified on trials, in which the subjects perceived the target, yet incorrectly identified their subjective visibility. Then, we might still have observed the apparent distinction in brain responses between seen and unseen correct trials, yet this might not only have been driven by conscious perception. It will therefore be important for future research to replicate these findings with an objective measure of visual awareness, with the goal being to obtain null sensitivity for the detection of the memorandum while maintaining above-chance performance on a forced-choice discrimination task.

A second major contention pertains to the specific nature of the long-lasting blindsight effect (Persuh et al., 2018). Is it fair to refer to it as non-conscious working memory if one of the major hallmarks of conscious working memory, the ability to manipulate information, has not yet been investigated in the context of this long-lasting blindsight? Here, too, I hope that my work may already offer first insights. As it stands today, the findings I presented in this thesis do not support the notion of non-conscious manipulation (Chapter 4) and, as such, argue against the long-lasting blindsight effect reflecting non-conscious working memory in its traditional sense. However, I have also tried to emphasize that, in general, the term "working memory" might be a bit of a misnomer and that, perhaps, it might be advantageous to focus on the underlying brain mechanisms (and their interplay) as opposed to more folk-psychological concepts when describing cognitive functions and phenomena. This perspective is, of course, just based on a limited amount of data from a very specific type of mental rotation paradigm. Perhaps, you may think, the inability to perform a non-conscious mental rotation is a reflection of my experimental design, rather than a true limitation of the system under investigation. The rotation cue being clearly visible may have encouraged subjects to adopt the strategy of consciously rotating a guess. Though certainly challenging, a possibility for future investigations might therefore be to include a subliminal rotation cue, such that participants no longer have access to a conscious representation of the task at hand. If, under these circumstances, objective performance were at chance, we would have strong evidence against an entirely non-conscious execution of a complex manipulation task.

The last limitation I would like to acknowledge in this section is not strictly related to the question of non-conscious working memory per se, but rather concerns the nature of the proposed neural mechanism. While content-specific neural activity is still considered to be the prime candidate for the neural correlate of the working memory engram (e.g., Fuster and Alexander, 1971; Kornblith et al., 2017), in conjunction with other very recent experiments (Mongillo et al., 2008; Rose et al., 2016; Sprague et al., 2016; Wolff et al., 2015, 2017), my own work suggests that information may also be stored in activity-silent brain states mediated by short-term changes in synaptic weights. At the moment, evidence for such activity-silent mechanisms is still primarily indirect. I, for instance, have shown that our behavioral results align themselves beautifully with the ones obtained from simulations under the hypothesis of such activity-silent maintenance (Chapter 2), and others have been able to reactivate previously silent representations with a non-specific impulse stimulus (Rose et al., 2016; Wolff et al., 2017). All of these attempts thus hinge on the assumption that, when decodability of a certain variable does not exceed chance-level, the

underlying representation must no longer have been coded in neural activity. Yet this obviously does not necessarily have to be the case. One may instead also imagine that our current technology simply is not sensitive to very weak and noisy neural signals, thus not being able to capture very subtle, yet still present information. In other words, the absence of evidence does not imply evidence for absence. Ultimately, we will have to show that, in our case, the contents of working memory may be stored in synaptic variables, such as neurotransmitter concentration. This is clearly a challenging endeavor, so, perhaps a more tangible goal for the near-future might be to rely on technologies with a higher signal-to-noise ratio, such as intracranial recordings in epileptic patients or monkeys.

Apart from these considerations based primarily on some of the limitations inherent to the work I conducted here, I also think that this thesis may serve as a corner stone for future investigations into the phenomena described. On one hand, we still know very little about even some of the most basic features and characteristics of the long-lasting blindsight effect: What are its limits in terms of durability and capacity? Is this phenomenon restricted to visual input, or may it also be observed for other sensory modalities, or for even more abstract concepts (other than temporal order), such as semantic representations? What role do the different brain regions recruited during such non-conscious maintenance play? In my own work, I have primarily focused on occipital regions (Chapters 2 and 4), but other studies have also reported activations in prefrontal cortex (Bergström and Eriksson, 2017; Dutta et al., 2014). Are the latter really causally involved in the task at hand, or do they reflect "peripheral" phenomena, such as top-down attention? Similarly, how do the conscious and non-conscious maintenance of information interact? May they interfere with each other? Could one, for instance, specifically instruct participants to "forget" all but a specific non-conscious representation? The possibilities here are really endless. However, I think that the ultimate challenge for future research will be to integrate all of these findings into a coherent, cohesive, updated framework of "working memory" (for lack of better terminology). What role do activity-silent and activity-based brain states play in the service of flexible, goal-directed behavior? Are representations not currently coded with activity-silent mechanisms automatically conscious? Are there situations, in which non-conscious information may also be stored in activity-based brain states? I hope that, by adopting a component-process approach, we will be able to unravel some of these mysteries in the not too distant future.

5.5 Conclusion

Our daily and intellectual lives depend on our ability to hold information in mind for immediate use. Despite a rich history of research, cracking the neuro-cognitive code of working memory remains one of the most important challenges of neuroscience to date. According to prevailing views, maintaining information in working memory requires conscious, effortful activity sustained over the entire delay period. However, this might only reflect the tip of the iceberg. The work I presented throughout this thesis challenges these notions, showing that information may also be stored non-consciously through activity-silent brain states. By contrast, manipulating such representations recruits both sustained neural activity and prior access to consciousness. I hope that this work may inspire future research to unravel the common functional architecture supporting short-term maintenance and manipulation in the brain.

REFERENCES

Aben, B., Stapert, S., and Blokland, A. (2012). About the Distinction between Working Memory and Short-Term Memory. Frontiers in Psychology *3*.

Adams, J.H., Graham, D.I., and Jennett, B. (2000). The neuropathology of the vegetative state after an acute brain insult. Brain *123*, 1327–1338.

Adolphs, R. (2015). The unsolved problems of neuroscience. Trends in Cognitive Sciences *19*, 173–175.

Alkire, M.T., and Miller, J. (2005). General anesthesia and the neural correlates of consciousness. Progress in Brain Research, 229–597.

Alkire, M.T., Haier, R.J., and Fallon, J.H. (2000). Toward a Unified Theory of Narcosis: Brain Imaging Evidence for a Thalamocortical Switch as the Neurophysiologic Basis of Anesthetic-Induced Unconsciousness. Consciousness and Cognition *9*, 370–386.

Aru, J., Bachmann, T., Singer, W., and Melloni, L. (2012). Distilling the neural correlates of consciousness. Neuroscience & Biobehavioral Reviews *36*, 737–746.

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

Atkinson, R.C., and Shiffrin, R.M. (1971). The Control Processes of Short-term Memory (Institute for Mathematical Studies in the Social Sciences, Stanford University).

Averbach, E., and Coriell, A.S. (1961). Short-Term Memory in Vision. Bell System Technical Journal *40*, 309–328.

Baars, B.J. (1988). A cognitive theory of consciousness (New York: Cambridge University Press).

Baars, B.J. (1994). A Thoroughly Empirical Approach To Consciousness.

Baars, B.J. (1997). In the Theater of Consciousness (Oxford University Press).

Baars, B.J., and Franklin, S. (2003). How conscious experience and working memory interact. Trends in Cognitive Sciences *7*, 166–172.

Backer, K.C., Binns, M.A., and Alain, C. (2015). Neural Dynamics Underlying Attentional Orienting to Auditory Representations in Short-Term Memory. Journal of Neuroscience *35*, 1307–1318.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? Trends in Cognitive Sciences *4*, 417–423.

Baddeley, A. (1992a). Working memory. Science *255*, 556–559.

Baddeley, A. (1992b). Consciousness and working memory. Consciousness and Cognition *1*, 3–6.

Baddeley, A. (2003). Working memory: looking back and looking forward. Nature Reviews Neuroscience *4*, 829–839.

Baddeley, A. (2012). Working Memory: Theories, Models, and Controversies. Annual Review of Psychology *63*, 1–29.

Baddeley, A.D. (1983). Working Memory. Philosophical Transactions of the Royal Society B: Biological Sciences *302*, 311–324.

Baddeley, A.D. (1993). Verbal and visual subsystems of working memory. Current Biology *3*, 563–565.

Baddeley, A.D., and Hitch, G. (1974). Working Memory. In Psychology of Learning and Motivation, (Elsevier), pp. 47–89.

Bapat, A.N., Shafer-Skelton, A., Kupitz, C.N., and Golomb, J.D. (2017). Binding object features to locations: Does the “spatial congruency bias” update with object movement? Attention, Perception, & Psychophysics *79*, 1682–1694.

Barak, O., and Tsodyks, M. (2007). Persistent Activity in Neural Networks with Dynamic Synapses. PLoS Computational Biology *3*, e35.

Barbey, A.K., Koenigs, M., and Grafman, J. (2013). Dorsolateral prefrontal contributions to human working memory. Cortex *49*, 1195–1205.

Baria, A.T., Maniscalco, B., and He, B.J. (2017). Initial-state-dependent, robust, transient neural dynamics encode conscious visual perception. PLOS Computational Biology *13*, e1005806.

Barrouillet, P., and Camos, V. (2012). As Time Goes By: Temporal Constraints in Working Memory. Current Directions in Psychological Science *21*, 413–419.

Barrouillet, P., Bernardin, S., and Camos, V. (2004). Time Constraints and Resource Sharing in Adults' Working Memory Spans. Journal of Experimental Psychology: General *133*, 83–100.

Barrouillet, P., Portrat, S., Vergauwe, E., Diependaele, K., and Camos, V. (2011). Further evidence for temporal decay in working memory: Reply to Lewandowsky and Oberauer (2009). Journal of Experimental Psychology: Learning, Memory, and Cognition *37*, 1302–1317.

Bartolomeo, P., Thiebaut de Schotten, M., and Doricchi, F. (2007). Left Unilateral Neglect as a Disconnection Syndrome. Cerebral Cortex *17*, 2479–2490.

Barttfeld, P., Uhrig, L., Sitt, J.D., Sigman, M., Jarraya, B., and Dehaene, S. (2015). Signature of consciousness in the dynamics of resting-state brain activity. Proc. Natl. Acad. Sci. USA *112*, 887–892.

Bays, P.M., and Husain, M. (2008). Dynamic Shifts of Limited Working Memory Resources in Human Vision. Science *321*, 851–854.

Beck, D.M., Rees, G., Frith, C.D., and Lavie, N. (2001). Neural correlates of change detection and change blindness. Nature Neuroscience *4*, 645–650.

Bergström, F., and Eriksson, J. (2014). Maintenance of non-consciously presented information engages the prefrontal cortex. Frontiers in Human Neuroscience *8*.

Bergström, F., and Eriksson, J. (2015). The conjunction of non-consciously perceived object identity and spatial position can be retained during a visual short-term memory task. Frontiers in Psychology *6*.

Bergström, F., and Eriksson, J. (2017). Neural Evidence for Non-conscious Working Memory. Cerebral Cortex 1–12.

Bianchi, L. (1895). THE FUNCTIONS OF THE FRONTAL LOBES. Brain *18*, 497–522.

Block, N. (2011). Perceptual consciousness overflows cognitive access. Trends in Cognitive Sciences *15*, 567–575.

Boehler, C.N., Schoenfeld, M.A., Heinze, H.-J., and Hopf, J.-M. (2008). Rapid recurrent processing gates awareness in primary visual cortex. Proc. Natl. Acad. Sci. U.S.A. *105*, 8742–8747.

Bona, S., and Silvanto, J. (2014). Accuracy and Confidence of Visual Short-Term Memory Do Not Go Hand-In-Hand: Behavioral and Neural Dissociations. PLoS ONE *9*, e90808.

Bona, S., Cattaneo, Z., Vecchi, T., Soto, D., and Silvanto, J. (2013). Metacognition of Visual Short-Term Memory: Dissociation between Objective and Subjective Components of VSTM. Frontiers in Psychology *4*.

Booth, M.C., and Rolls, E.T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. Cereb. Cortex *8*, 510–523.

Bor, D., Schwartzman, D.J., Barrett, A.B., and Seth, A.K. (2017). Theta-burst transcranial magnetic stimulation to the prefrontal or parietal cortex does not impair metacognitive visual awareness. PLOS ONE *12*, e0171793.

Botvinick, M., and Watanabe, T. (2007). From Numerosity to Ordinal Rank: A Gain-Field Model of Serial Order Representation in Cortical Working Memory. Journal of Neuroscience *27*, 8636–8642.

Boy, F., Husain, M., Singh, K.D., and Sumner, P. (2010). Supplementary motor area activations in unconscious inhibition of voluntary action. Experimental Brain Research *206*, 441–448.

Brady, T.F., Konkle, T., Alvarez, G.A., and Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. Proc. Natl. Acad. Sci. USA *105*, 14325–14329.

Breitmeyer, B.G., and Öğmen, H. (2006). Visual masking: time slices through conscious and unconscious vision (Oxford ; New York: Oxford University Press).

Brent, P.J., Kennard, C., and Ruddock, K.H. (1994). Residual Colour Vision in a Human Hemianope: Spectral Responses and Colour Discrimination. Proceedings of the Royal Society B: Biological Sciences *256*, 219–225.

Broadbent, D.E. (1957). A mechanical model for human attention and immediate memory. Psychol Rev *64*, 205–215.

Broadbent, D.E. (1958). Perception and communication.

Brown, G.D.A., Preece, T., and Hulme, C. (2000). Oscillator-based memory for serial order. Psychological Review *107*, 127–181.

Buonomano, D.V., and Maass, W. (2009). State-dependent computations: spatiotemporal processing in cortical networks. Nature Reviews Neuroscience *10*, 113–125.

Burgess, N., and Hitch, G.J. (1999). Memory for serial order: A network model of the phonological loop and its timing. Psychological Review *106*, 551–581.

Buschman, T.J., and Miller, E.K. (2007). Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices. Science *315*, 1860–1862.

Buschman, T.J., Siegel, M., Roy, J.E., and Miller, E.K. (2011). Neural substrates of cognitive capacity limitations. Proc. Natl. Acad. Sci. USA *108*, 11252–11255.

Butters, N., and Pandya, D. (1969). Retention of Delayed-Alternation: Effect of Selective Lesions of Sulcus Principalis. Science *165*, 1271–1273.

Cavanagh, S.E., Towers, J.P., Wallis, J.D., Hunt, L.T., and Kennerley, S.W. (2017). Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex.

Chalmers, D. (1995). Facing up the problem of consciousness. Journal of Consciousness Studies.

Chalmers, D.J. (1997). The conscious mind: in search of a fundamental theory (New York: Oxford Univ. Press).

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, Article 27.

Charles, L., Van Opstal, F., Marti, S., and Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. NeuroImage *73*, 80–94.

Charles, L., King, J.-R., and Dehaene, S. (2014). Decoding the Dynamics of Action, Intention, and Error Detection for Conscious and Subliminal Stimuli. Journal of Neuroscience *34*, 1158–1170.

Charles, L., Gaillard, R., Amado, I., Krebs, M.-O., Bendjemaa, N., and Dehaene, S. (2017). Conscious and unconscious performance monitoring: Evidence from patients with schizophrenia. NeuroImage *144*, 153–163.

Chaudhuri, R., Bernacchia, A., and Wang, X.-J. (2014). A diversity of localized timescales in network activity. ELife *3*.

Cheesman, J., and Merikle, P.M. (1986). Distinguishing conscious from unconscious perceptual processes. Can J Psychol *40*, 343–367.

Cherry, E.C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. The Journal of the Acoustical Society of America *25*, 975–979.

Chong, T.T.-J., Husain, M., and Rosenthal, C.R. (2014). Recognizing the unconscious. Current Biology *24*, R1033–R1035.

Christophel, T.B., Hebart, M.N., and Haynes, J.-D. (2012). Decoding the Contents of Visual Short-Term Memory from Human Visual and Parietal Cortex. Journal of Neuroscience *32*, 12983–12989.

Christophel, T.B., Iamshchinina, P., Yan, C., Allefeld, C., and Haynes, J.-D. (2018). Cortical specialization for attended versus unattended working memory. Nature Neuroscience *21*, 494–496.

Christophel, T.B., Klink, P.C., Spitzer, B., Roelfsema, P.R., and Haynes, J.-D. (2017). The distributed nature of working memory. Trends in Cognitive Sciences *21*, 111-124.

Cohen, M.A., and Dennett, D.C. (2011). Consciousness cannot be separated from function. Trends in Cognitive Sciences *15*, 358–364.

Cohen, M.A., Dennett, D.C., and Kanwisher, N. (2016). What is the Bandwidth of Perceptual Experience? Trends in Cognitive Sciences *20*, 324–335.

Constantinidis, C., and Klingberg, T. (2016). The neuroscience of working memory capacity and training. Nature Reviews Neuroscience *17*, 438–449.

Cooper, L.A. (1975). Mental rotation of random two-dimensional shapes. Cognitive Psychology *7*, 20–43.

Cooper, N.R., Croft, R.J., Dominey, S.J.., Burgess, A.P., and Gruzelier, J.H. (2003). Paradox lost? Exploring the role of alpha oscillations during externally vs. internally directed attention and the implications for idling and inhibition hypotheses. International Journal of Psychophysiology *47*, 65–74.

Corbetta, M., Kincade, M.J., Lewis, C., Snyder, A.Z., and Sapir, A. (2005). Neural basis and recovery of spatial attention deficits in spatial neglect. Nature Neuroscience *8*, 1603.

Courtney, S.M., Ungerleider, L.G., Keil, K., and Haxby, J.V. (1997). Transient and sustained activity in a distributed neural system for human working memory. Nature *386*, 608–611.

Courtney, S.M., Petit, L., Maisog, J.M., Ungerleider, L.G., and Haxby, J.V. (1998a). An area specialized for spatial working memory in human frontal cortex. Science *279*, 1347–1351.

Courtney, S.M., Petit, L., Haxby, J.V., and Ungerleider, L.G. (1998b). The role of prefrontal cortex in working memory: examining the contents of consciousness. Philosophical Transactions of the Royal Society B: Biological Sciences *353*, 1819–1828.

Cowan, N. (1997). Attention and memory: an integrated framework (New York: Oxford Univ. Press [u.a.]).

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behav Brain Sci *24*, 87–114; discussion 114-185.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? In Progress in Brain Research, (Elsevier), pp. 323–338.

Cowey, A., and Stoerig, P. (1995). Blindsight in monkeys. Nature *373*, 247.

Crannell, C.W., and Parrish, J.M. (1957). A Comparison of Immediate Memory Span for Digits, Letters, and Words. The Journal of Psychology *44*, 319–327.

Crick, F., and Koch, C. (2003). A framework for consciousness. Nature Neuroscience *6*, 119.

Curtis, C.E., and D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. Trends in Cognitive Sciences *7*, 415–423.

Debner, J.A., and Jacoby, L.L. (1994). Unconscious perception: Attention, awareness, and control. Journal of Experimental Psychology: Learning, Memory, and Cognition *20*, 304–317.

Dehaene, S., and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. Cognition *79*, 1–37.

Dehaene, S., and Changeux, J.-P. (2011). Experimental and Theoretical Approaches to Conscious Processing. Neuron *70*, 200–227.

Dehaene, S., Kerszberg, M., and Changeux, J.-P. (1998a). A neuronal model of a global workspace in effortful cognitive tasks. Proc. Natl. Acad. Sci. USA *95*, 14529–14534.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P.F., and Le Bihan, D. (1998b). Imaging unconscious semantic priming. Nature *395*, 597–600.

Dehaene, S., Naccache, L., Cohen, L., Bihan, D.L., Mangin, J.-F., Poline, J.-B., and Rivière, D. (2001). Cerebral mechanisms of word masking and unconscious repetition priming. Nature Neuroscience *4*, 752–758.

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., and Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends in Cognitive Sciences *10*, 204–211.

Dehaene, S., Charles, L., King, J.-R., and Marti, S. (2014). Toward a computational theory of conscious processing. Current Opinion in Neurobiology *25*, 76–84.

Dehaene, S., Lau, H., and Kouider, S. (2017). What is consciousness, and could machines have it? Science *358*, 486–492.

Del Cul, A., Baillet, S., and Dehaene, S. (2007). Brain Dynamics Underlying the Nonlinear Threshold for Access to Consciousness. PLoS Biology *5*, e260.

Descartes, R., and Renault, L. (1637). Discours de la méthode (Paris: Flammarion).

D'Esposito, M., and Postle, B.R. (1999). The dependence of span and delayed-response performance on prefrontal cortex. Neuropsychologia *37*, 1303–1315.

D'Esposito, M., Cooney, J.W., Gazzaley, A., Gibbs, S.E.B., and Postle, B.R. (2006). Is the Prefrontal Cortex Necessary for Delay Task Performance? Evidence from Lesion and fMRI Data. Journal of the International Neuropsychological Society *12*.

Dick, A.O. (1974). Iconic memory and its relation to perceptual processing and other memory mechanisms. Perception & Psychophysics *16*, 575–596.

Dienes, Z., Altmann, G.T.M., Kwan, L., and Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. Journal of Experimental Psychology: Learning, Memory, and Cognition *21*, 1322–1338.

Dosher, B., Liu, S.H., and Lu, Z.L. (2005). The decay of perceptual representations in iconic memory. Journal of Vision *5*, 912–912.

Dupoux, E., Gardelle, V. de, and Kouider, S. (2008). Subliminal speech perception and auditory streaming. Cognition *109*, 267–273.

Dutta, A., Shah, K., Silvanto, J., and Soto, D. (2014). Neural basis of non-conscious visual working memory. NeuroImage *91*, 336–343.

Ebbinghaus, H. (1885). Über das Gedächtnis: Untersuchungen zur Experimentellen Psychologie (Leipzig).

Emrich, S.M., Riggall, A.C., LaRocque, J.J., and 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*, 6516–6523.

Engel, A.K., and Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. Trends in Cognitive Sciences *5*, 16–25.

Enns, J.T., and Di Lollo, V. (2000). What's new in visual masking? Trends in Cognitive Sciences *4*, 345–352.

Eriksson, J., Vogel, E.K., Lansner, A., Bergström, F., and Nyberg, L. (2015). Neurocognitive Architecture of Working Memory. Neuron *88*, 33–46.

Ester, E.F., Serences, J.T., and Awh, E. (2009). Spatially Global Representations in Human Primary Visual Cortex during Working Memory Maintenance. Journal of Neuroscience *29*, 15258–15265.

Ester, E.F., Sprague, T.C., and Serences, J.T. (2015). Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. Neuron *87*, 893–905.

Fagot, J., and Cook, R.G. (2006). Evidence for large long-term memory capacities in baboons and pigeons and its implications for learning and the evolution of cognition. Proc. Natl. Acad. Sci. USA *103*, 17564–17567.

Felleman, D.J., and Van Essen, D.C. (1991). Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex *1*, 1–47.

Ferrier, D. (1876). The functions of the brain (London: Smith, Elder, & Co.).

Fiebig, F., and Lansner, A. (2017). A Spiking Working Memory Model Based on Hebbian Short-Term Potentiation. The Journal of Neuroscience *37*, 83–96.

Fonov, V., Evans, A., McKinstry, R., Almli, C., and Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage *47*, S102.

Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., and Collins, D.L. (2011). Unbiased average age-appropriate atlases for pediatric studies. NeuroImage *54*, 313–327.

Fuentemilla, L., Penny, W.D., Cashdollar, N., Bunzeck, N., and Düzel, E. (2010). Theta-Coupled Periodic Replay in Working Memory. Current Biology *20*, 606–612.

Fujisawa, S., Amarasingham, A., Harrison, M.T., and Buzsáki, G. (2008). Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. Nature Neuroscience *11*, 823–833.

Funahashi, S., Bruce, C.J., and Goldman-Rakic, P.S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. J. Neurophysiol. *61*, 331–349.

Fuster, J.M. (2015). The prefrontal cortex (Amsterdam ; Boston: Elsevier/AP, Academic Press is an imprint of Elsevier).

Fuster, J.M., and Alexander, G.E. (1971). Neuron activity related to short-term memory. Science *173*, 652–654.

Fuster, J.M., and Jervey, J.P. (1981). Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. Science *212*, 952–955.

van Gaal, S., and Lamme, V.A.F. (2012). Unconscious High-Level Information Processing: Implication for Neurobiological Theories of Consciousness. The Neuroscientist *18*, 287–301.

van Gaal, S., Ridderinkhof, K.R., Scholte, H.S., and Lamme, V.A.F. (2010). Unconscious Activation of the Prefrontal No-Go Network. Journal of Neuroscience *30*, 4143–4150.

van Gaal, S., Naccache, L., Meuwese, J.D.I., van Loon, A.M., Leighton, A.H., Cohen, L., and Dehaene, S. (2014). Can the meaning of multiple words be integrated unconsciously? Philosophical Transactions of the Royal Society B: Biological Sciences *369*, 20130212–20130212.

Gaillard, R., Del Cul, A., Naccache, L., Vinckier, F., Cohen, L., and Dehaene, S. (2006). Nonconscious semantic processing of emotional words modulates conscious access. Proc. Natl. Acad. Sci. USA *103*, 7524–7529.

Gaillard, R., Dehaene, S., Adam, C., Clémenceau, S., Hasboun, D., Baulac, M., Cohen, L., and Naccache, L. (2009). Converging Intracranial Markers of Conscious Access. PLoS Biology *7*, e1000061.

Galeano Weber, E.M., Hahn, T., Hilger, K., and Fiebach, C.J. (2017). Distributed patterns of occipito-parietal functional connectivity predict the precision of visual working memory. NeuroImage *146*, 404–418.

de Gardelle, V., and Kouider, S. (2009). Cognitive Theories of Consciousness. In Encyclopedia of Consciousness, (Elsevier), pp. 135–146.

Gauthier, I., Hayward, W.G., Tarr, M.J., Anderson, A.W., Skudlarski, P., and Gore, J.C. (2002). BOLD Activity during Mental Rotation and Viewpoint-Dependent Object Recognition. Neuron *34*, 161–171.

Gazzaley, A., Rissman, J., and D'Esposito, M. (2004). Functional connectivity during working memory maintenance. Cogn Affect Behav Neurosci *4*, 580–599.

Georgopoulos, A., Lurito, J., Petrides, M., Schwartz, A., and Massey, J. (1989). Mental rotation of the neuronal population vector. Science *243*, 234–236.

Goldhill, O. (2018). Psychology will fail if it keeps using ancient words like "attention" and "memory."

Goldman, P.S., and Rosvold, H.E. (1970). Localization of function within the dorsolateral prefrontal cortex of the rhesus monkey. Experimental Neurology *27*, 291–304.

Goldman-Rakic, P.S. (1995). Cellular basis of working memory. Neuron *14*, 477–485.

Gramfort, A. (2013). MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience *7*.

Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Parkkonen, L., and Hämäläinen, M.S. (2014). MNE software for processing MEG and EEG data. Neuroimage *86*, 446–460.

Green, D.M., and Swets, J.A. (2000). Signal detection theory and psychophysics (Los Altos Hills, Calif: Peninsula Publ).

Greenwald, A.G., Draine, S.C., and Abrams, R.L. (1996). Three Cognitive Markers of Unconscious Semantic Activation. Science *273*, 1699–1702.

Grill-Spector, K., Kushnir, T., Hendler, T., and Malach, R. (2000). The dynamics of object-selective activation correlate with recognitionperformance in humans. Nature Neuroscience *3*, 837–843.

Gross, J., Schmitz, F., Schnitzler, I., Kessler, K., Shapiro, K., Hommel, B., and Schnitzler, A. (2004). Modulation of long-range neural synchrony reflects temporal limitations of visual attention in humans. Proc. Natl. Acad. Sci. USA *101*, 13050–13055.

Gross, J., Schnitzler, A., Timmermann, L., and Ploner, M. (2007). Gamma Oscillations in Human Primary Somatosensory Cortex Reflect Pain Perception. PLoS Biology *5*, e133.

Halford, G.S., Cowan, N., and Andrews, G. (2007). Separating cognitive capacity from knowledge: a new hypothesis. Trends in Cognitive Sciences *11*, 236–242.

Haller, M., Case, J., Crone, N.E., Chang, E.F., King-Stephens, D., Laxer, K.D., Weber, P.B., Parvizi, J., Knight, R.T., and Shestyuk, A.Y. (2018). Persistent neuronal activity in human prefrontal cortex links perception and action. Nature Human Behaviour *2*, 80–91.

Hannula, D.E., Simons, D.J., and Cohen, N.J. (2005). Imaging implicit perception: promise and pitfalls. Nature Reviews Neuroscience *6*, 247.

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

Hassin, R.R., Bargh, J.A., Engell, A.D., and McCulloch, K.C. (2009). Implicit working memory. Conscious Cogn *18*, 665–678.

Haynes, J.-D., and Rees, G. (2006). Decoding mental states from brain activity in humans. Nature Reviews Neuroscience *7*, 523–534.

Haynes, J.-D., Deichmann, R., and Rees, G. (2005). Eye-specific effects of binocular rivalry in the human lateral geniculate nucleus. Nature *438*, 496–499.

He, B.J. (2018). Robust, Transient Neural Dynamics during Conscious Perception. Trends in Cognitive Sciences *22*, 563–565.

Hebb, D.O. (1949). The organization of behavior: a neuropsychological theory (Mahwah, N.J: L. Erlbaum Associates).

Helmholtz, H. (1866). Helmholtz's treatise on physiological optics (New York, NY: Dover Publications).

Henson, R.N.A. (1999). Positional information in short-term memory: Relative or absolute? Memory & Cognition *27*, 915–927.

Huang, Y., Matysiak, A., Heil, P., König, R., and Brosch, M. (2016). Persistent neural activity in auditory cortex is related to auditory working memory in humans and nonhuman primates. ELife *5*.

Hubel, D.H., and Wiesel, T.N. (1968). Receptive fields and functional architecture of monkey striate cortex. The Journal of Physiology *195*, 215–243.

Humphrey, N.K. (1974). Vision in a Monkey without Striate Cortex: A Case Study. Perception *3*, 241–255.

Jackendoff, R. (1994). Consciousness and the computational mind (Cambridge, Mass.: MIT Press).

Jacobs, C., and Silvanto, J. (2015). How is working memory content consciously experienced? The 'conscious copy' model of WM introspection. Neuroscience & Biobehavioral Reviews *55*, 510–519.

Jacobsen, C.F. (1935). An experimental analysis of the functions of the frontal assocation areas in primates. The Journal of Nervous and Mental Disease 1–14.

James, W. (1890). The Principles of Psychology (New York: Henry Holt and Company).

Jansma, J.M., Ramsey, N.F., Coppola, R., and Kahn, R.S. (2000). Specific versus Nonspecific Brain Activity in a Parametric N-Back Task. NeuroImage *12*, 688–697.

Jensen, O., and Mazaheri, A. (2010). Shaping Functional Architecture by Oscillatory Alpha Activity: Gating by Inhibition. Frontiers in Human Neuroscience *4*.

Jerde, T.A., Merriam, E.P., Riggall, A.C., Hedges, J.H., and Curtis, C.E. (2012). Prioritized Maps of Space in Human Frontoparietal Cortex. Journal of Neuroscience *32*, 17382–17390.

Jiang, Y., Costello, P., and He, S. (2007). Processing of Invisible Stimuli: Advantage of Upright Faces and Recognizable Words in Overcoming Interocular Suppression. Psychological Science *18*, 349–355.

Jones, E.G. (2002). Thalamic circuitry and thalamocortical synchrony. Philosophical Transactions of the Royal Society B: Biological Sciences *357*, 1659–1673.

Jonides, J., Lewis, R.L., Nee, D.E., Lustig, C.A., Berman, M.G., and Moore, K.S. (2008). The mind and brain of short-term memory. Annu Rev Psychol *59*, 193–224.

Jordan, K., Heinze, H.-J., Lutz, K., Kanowski, M., and Jäncke, L. (2001). Cortical Activations during the Mental Rotation of Different Visual Objects. NeuroImage *13*, 143–152.

Kamiński, J., Sullivan, S., Chung, J.M., Ross, I.B., Mamelak, A.N., and Rutishauser, U. (2017). Persistently active neurons in human medial frontal and medial temporal lobe support working memory. Nature Neuroscience *20*, 590–601.

Karpinski, A., Briggs, J.C., and Yale, M. (2018). A direct replication: Unconscious arithmetic processing. European Journal of Social Psychology.

van Kerkoerle, T., Self, M.W., and Roelfsema, P.R. (2017). Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. Nature Communications *8*, 13804.

Kim, C.-Y., and Blake, R. (2005). Psychophysical magic: rendering the visible 'invisible.' Trends in Cognitive Sciences *9*, 381–388.

King, J.-R. (2014). Characterizing the electro-magnetic signatures of conscious processing in healthy and impaired human brains. PhD thesis.

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

King, J.-R., Sitt, J.D., Faugeras, F., Rohaut, B., El Karoui, I., Cohen, L., Naccache, L., and Dehaene, S. (2013). Information Sharing in the Brain Indexes Consciousness in Noncommunicative Patients. Current Biology *23*, 1914–1919.

King, J.-R., Gramfort, A., Schurger, A., Naccache, L., and Dehaene, S. (2014). Two Distinct Dynamic Modes Subtend the Detection of Unexpected Sounds. PLoS ONE *9*, e85791.

King, J.-R., Pescetelli, N., and Dehaene, S. (2016). Brain Mechanisms Underlying the Brief Maintenance of Seen and Unseen Sensory Information. Neuron *92*, 1122–1134.

Kinney, H.C., Korein, J., Panigrahy, A., Dikkes, P., and Goode, R. (1994). Neuropathological Findings in the Brain of Karen Ann Quinlan -- The Role of the Thalamus in the Persistent Vegetative State. New England Journal of Medicine *330*, 1469–1475.

Kleinschmidt, A., Buchel, C., Zeki, S., and Frackowiak, R.S.J. (1998). Human brain activity during spontaneously reversing perception of ambiguous figures. Proceedings of the Royal Society B: Biological Sciences *265*, 2427–2433.

Koch, C. (2004). The quest for consciousness: a neurobiological approach (Denver, Colo: Roberts and Co).

Koch, C., Massimini, M., Boly, M., and Tononi, G. (2016). Neural correlates of consciousness: progress and problems. Nature Reviews Neuroscience *17*, 307–321.

Kok, P., Rahnev, D., Jehee, J.F.M., Lau, H.C., and de Lange, F.P. (2012). Attention Reverses the Effect of Prediction in Silencing Sensory Signals. Cerebral Cortex *22*, 2197–2206.

Kornblith, S., Quian Quiroga, R., Koch, C., Fried, I., and Mormann, F. (2017). Persistent Single-Neuron Activity during Working Memory in the Human Medial Temporal Lobe. Current Biology *27*, 1026–1032.

Kouider, S., and Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. Philosophical Transactions of the Royal Society B: Biological Sciences *362*, 857–875.

Kouider, S., de Gardelle, V., Sackur, J., and Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. Trends in Cognitive Sciences *14*, 301–307.

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

Kumar, S., Joseph, S., Gander, P.E., Barascud, N., Halpern, A.R., and Griffiths, T.D. (2016). A Brain System for Auditory Working Memory. Journal of Neuroscience *36*, 4492–4505.

Kuo, B.-C., Lin, S.-H., and Yeh, Y.-Y. (2018). Functional interplay of top-down attention with affective codes during visual short-term memory maintenance. Cortex *103*, 55–70.

Lamme, V.A.F. (2006). Towards a true neural stance on consciousness. Trends in Cognitive Sciences *10*, 494–501.

Lamme, V.A.F. (2010). How neuroscience will change our view on consciousness. Cognitive Neuroscience *1*, 204–220.

Lamme, V.A., and Roelfsema, P.R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. Trends in Neurosciences. *23*, 571–579.

Lamme, V.A., Supèr, H., Landman, R., Roelfsema, P.R., and Spekreijse, H. (2000). The role of primary visual cortex (V1) in visual awareness. Vision Res. *40*, 1507–1521.

Lamme, V.A.F., Zipser, K., and Spekreijse, H. (1998). Figure-ground activity in primary visual cortex is suppressed by anesthesia. Proc. Natl. Acad. Sci. USA *95*, 3263–3268.

Lamme, V.A.F., Zipser, K., and Spekreijse, H. (2002). Masking Interrupts Figure-Ground Signals in V1. J Cogn Neurosci *14*, 1044–1053.

Lamy, D., Salti, M., and Bar-Haim, Y. (2009). Neural Correlates of Subjective Awareness and Unconscious Processing: An ERP Study. J Cogn Neurosci *21*, 1435–1446.

LaRocque, J.J., Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K., and Postle, B.R. (2013). Decoding attended information in short-term memory: an EEG study. J Cogn Neurosci *25*, 127–142.

Latimer, K.W., Yates, J.L., Meister, M.L.R., Huk, A.C., and Pillow, J.W. (2015). Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. Science *349*, 184–187.

Lau, H.C. (2007). A higher order Bayesian decision theory of consciousness. In Progress in Brain Research, (Elsevier), pp. 35–48.

Lau, H., and Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. Trends in Cognitive Sciences *15*, 365–373.

Lau, H.C., and Passingham, R.E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. Proc. Natl. Acad. Sci. USA *103*, 18763–18768.

Lau, H.C., and Passingham, R.E. (2007). Unconscious Activation of the Cognitive Control System in the Human Prefrontal Cortex. Journal of Neuroscience *27*, 5805–5811.

Laureys, S. (2005). The neural correlate of (un)awareness: lessons from the vegetative state. Trends in Cognitive Sciences *9*, 556–559.

Lee, S.-H., Kravitz, D.J., and Baker, C.I. (2013). Goal-dependent dissociation of visual and prefrontal cortices during working memory. Nature Neuroscience *16*, 997–999.

Lemus, L., Hernandez, A., and Romo, R. (2009). Neural encoding of auditory discrimination in ventral premotor cortex. Proc. Natl. Acad. Sci. USA *106*, 14640–14645.

Leopold, D.A. (2012). Primary Visual Cortex: Awareness and Blindsight. Annual Review of Neuroscience *35*, 91–109.

Leopold, D.A., and Logothetis, N.K. (1996). Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. Nature *379*, 549.

Levy, R., and Goldman-Rakic, P.S. (1999). Association of Storage and Processing Functions in the Dorsolateral Prefrontal Cortex of the Nonhuman Primate. The Journal of Neuroscience *19*, 5149–5158.

Lewandowsky, S., and Oberauer, K. (2009). No evidence for temporal decay in working memory. Journal of Experimental Psychology: Learning, Memory, and Cognition *35*, 1545–1551.

Lewandowsky, S., Oberauer, K., and Brown, G.D.A. (2009). No temporal decay in verbal short-term memory. Trends in Cognitive Sciences *13*, 120–126.

Lewis, L.D., Weiner, V.S., Mukamel, E.A., Donoghue, J.A., Eskandar, E.N., Madsen, J.R., Anderson, W.S., Hochberg, L.R., Cash, S.S., Brown, E.N., et al. (2012). Rapid fragmentation of neuronal networks at the onset of propofol-induced unconsciousness. Proc. Natl. Acad. Sci. USA *109*, E3377–E3386.

Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K., and Postle, B.R. (2012). Neural Evidence for a Distinction between Short-term Memory and the Focus of Attention. J Cogn Neurosci *24*, 61–79.

Liebe, S., Hoerzer, G.M., Logothetis, N.K., and Rainer, G. (2012). Theta coupling between V4 and prefrontal cortex predicts visual short-term memory performance. Nature Neuroscience *15*, 456.

Logothetis, N.K., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. Current Biology *5*, 552–563.

Logothetis, N.K., Leopold, D.A., and Sheinberg, D.L. (1996). What is rivalling during binocular rivalry? Nature *380*, 621.

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

Luck, S.J., and Vogel, E.K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. Trends in Cognitive Sciences *17*, 391–400.

Lumer, E.D. (1998). Neural Correlates of Perceptual Rivalry in the Human Brain. Science *280*, 1930–1934.

Lumer, E.D., and Rees, G. (1999). Covariation of activity in visual and prefrontal cortex associated with subjective visual perception. Proc. Natl. Acad. Sci. U.S.A. *96*, 1669–1673.

Lundqvist, M., Herman, P., and Lansner, A. (2011). Theta and gamma power increases and alpha/beta power decreases with memory load in an attractor network model. J Cogn Neurosci *23*, 3008–3020.

Lundqvist, M., Rose, J., Herman, P., Brincat, S.L., Buschman, T.J., and Miller, E.K. (2016). Gamma and Beta Bursts Underlie Working Memory. Neuron *90*, 152–164.

Lundqvist, M., Herman, P., Warden, M.R., Brincat, S.L., and Miller, E.K. (2018). Gamma and beta bursts during working memory readout suggest roles in its volitional control. Nature Communications *9*.

Lutz, A., Lachaux, J.-P., Martinerie, J., and Varela, F.J. (2002). Guiding the study of brain dynamics by using first-person data: Synchrony patterns correlate with ongoing conscious states during a simple visual task. Proc. Natl. Acad. Sci. USA *99*, 1586–1591.

Malmo, R.B. (1942). INTERFERENCE FACTORS IN DELAYED RESPONSE IN MONKEYS AFTER REMOVAL OF FRONTAL LOBES. Journal of Neurophysiology *5*, 295–308.

Manes, F., Sahakian, B., Clark, L., Rogers, R., Antoun, N., Aitken, M., and Robbins, T. (2002). Decision-making processes following damage to the prefrontal cortex. Brain *125*, 624–639.

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods *164*, 177–190.

Marois, R., and Ivanoff, J. (2005). Capacity limits of information processing in the brain. Trends in Cognitive Sciences *9*, 296–305.

Marshuetz, C., and Smith, E.E. (2006). Working memory for order information: Multiple cognitive and neural mechanisms. Neuroscience *139*, 195–200.

Marti, S., and Dehaene, S. (2017). Discrete and continuous mechanisms of temporal selection in rapid visual streams. Nature Communications *8*.

Marti, S., King, J.-R., and Dehaene, S. (2015). Time-Resolved Decoding of Two Processing Chains during Dual-Task Interference. Neuron *88*, 1297–1307.

Masse, N.Y., Yang, G.R., Song, H.F., Wang, X.-J., and Freedman, D.J. (2018). Circuit mechanisms for the maintenance and manipulation of information in working memory.

Massimini, M. (2005). Breakdown of Cortical Effective Connectivity During Sleep. Science *309*, 2228–2232.

McElree, B. (2001). Working memory and focal attention. J Exp Psychol Learn Mem Cogn *27*, 817–835.

McElree, B. (2006). Accessing Recent Events. In Psychology of Learning and Motivation, (Elsevier), pp. 155–200.

Melloni, L., Molina, C., Pena, M., Torres, D., Singer, W., and Rodriguez, E. (2007). Synchronization of Neural Activity across Cortical Areas Correlates with Conscious Perception. Journal of Neuroscience *27*, 2858–2865.

Melloni, L., Schwiedrzik, C.M., Muller, N., Rodriguez, E., and Singer, W. (2011). Expectations Change the Signatures and Timing of Electrophysiological Correlates of Perceptual Awareness. Journal of Neuroscience *31*, 1386–1396.

Merikle, P.M., and Reingold, E.M. (1990). Recognition and lexical decision without detection: unconscious perception? J Exp Psychol Hum Percept Perform *16*, 574–583.

Merikle, P.M., Joordens, S., and Stolz, J.A. (1995). Measuring the Relative Magnitude of Unconscious Influences. Consciousness and Cognition *4*, 422–439.

Merikle, P.M., Smilek, D., and Eastwood, J.D. (2001). Perception without awareness: perspectives from cognitive psychology. Cognition *79*, 115–134.

Meyniel, F., and Pessiglione, M. (2014). Better Get Back to Work: A Role for Motor Beta Desynchronization in Incentive Motivation. Journal of Neuroscience *34*, 1–9.

Mi, Y., Katkov, M., and Tsodyks, M. (2017). Synaptic Correlates of Working Memory Capacity. Neuron *93*, 323–330.

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

Miller, E.K., Li, L., and Desimone, R. (1993). Activity of neurons in anterior inferior temporal cortex during a short-term memory task. J. Neurosci. *13*, 1460–1478.

Miller, E.K., Erickson, C.A., and Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. J. Neurosci. *16*, 5154–5167.

Miller, G.A., Galanter, E., and Přibram, K.H. (1960). Plans and the structure of behavior (New York: Holt).

Milner, B. (1982). Some cognitive effects of frontal-lobe lesions in man. Philos. Trans. R. Soc. Lond., B, Biol. Sci. *298*, 211–226.

Mohler, C.W., and Wurtz, R.H. (1977). Role of striate cortex and superior colliculus in visual guidance of saccadic eye movements in monkeys. Journal of Neurophysiology *40*, 74–94.

Mongillo, G., Barak, O., and Tsodyks, M. (2008). Synaptic Theory of Working Memory. Science *319*, 1543–1546.

Moore, T., Rodman, H.R., Repp, A.B., and Gross, C.G. (1995). Localization of visual stimuli after striate cortex damage in monkeys: parallels with human blindsight. Proc. Natl. Acad. Sci. U.S.A. *92*, 8215–8218.

Moors, P., and Hesselmann, G. (2018). A critical reexamination of doing arithmetic nonconsciously. Psychonomic Bulletin & Review *25*, 472–481.

Mudrik, L., Faivre, N., and Koch, C. (2014). Information integration without awareness. Trends in Cognitive Sciences *18*, 488–496.

Murdock, B.B., Jr. (1962). The serial position effect of free recall. Journal of Experimental Psychology *64*, 482–488.

Muter, P. (1980). Very rapid forgetting. Memory & Cognition *8*, 174–179.

Myers, N.E., Rohenkohl, G., Wyart, V., Woolrich, M.W., Nobre, A.C., and Stokes, M.G. (2015). Testing sensory evidence against mnemonic templates. ELife *4*.

Naccache, L., and Dehaene, S. (2001). Unconscious semantic priming extends to novel unseen stimuli. Cognition *80*, 215–229.

Naccache, L., Marti, S., Sitt, J.D., Trübutschek, D., and Berkovitch, L. (2016). Why the P3b is still a plausible correlate of conscious access? A commentary on Silverstein et al., 2015. Cortex *85*, 126–128.

Naghavi, H.R., and Nyberg, L. (2005). Common fronto-parietal activity in attention, memory, and consciousness: Shared demands on integration? Consciousness and Cognition *14*, 390–425.

Nakamura, K., Makuuchi, M., Oga, T., Mizuochi-Endo, T., Iwabuchi, T., Nakajima, Y., and Dehaene, S. (2018). Neural capacity limits during unconscious semantic processing. European Journal of Neuroscience *47*, 929–937.

Neisser, U. (1967). Cognitive Psychology (New York, NY: Appleton-Century-Crofts [u.a.]).

Norman, D.A., and Shallice, T. (1986). Attention to Action. In Consciousness and Self-Regulation, R.J. Davidson, G.E. Schwartz, and D. Shapiro, eds. (Boston, MA: Springer US), pp. 1–18.

Oberauer, K. (2001). Removing irrelevant information from working memory: a cognitive aging study with the modified Sternberg task. J Exp Psychol Learn Mem Cogn *27*, 948–957.

Oberauer, K. (2002). Access to information in working memory: exploring the focus of attention. J Exp Psychol Learn Mem Cogn *28*, 411–421.

Oberauer, K. (2005). Control of the Contents of Working Memory--A Comparison of Two Paradigms and Two Age Groups. Journal of Experimental Psychology: Learning, Memory, and Cognition *31*, 714–728.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. Computational Intelligence and Neuroscience *2011*, 1–9.

Owen, A.M., Downes, J.J., Sahakian, B.J., Polkey, C.E., and Robbins, T.W. (1990). Planning and spatial working memory following frontal lobe lesions in man. Neuropsychologia *28*, 1021–1034.

Pan, Y., Lin, B., Zhao, Y., and Soto, D. (2014). Working memory biasing of visual perception without awareness. Attention, Perception, & Psychophysics *76*, 2051–2062.

Panagiotaropoulos, T.I., Deco, G., Kapoor, V., and Logothetis, N.K. (2012). Neuronal Discharges and Gamma Oscillations Explicitly Reflect Visual Consciousness in the Lateral Prefrontal Cortex. Neuron *74*, 924–935.

Parthasarathy, A., Herikstad, R., Bong, J.H., Medina, F.S., Libedinsky, C., and Yen, S.-C. (2017). Mixed selectivity morphs population codes in prefrontal cortex. Nature Neuroscience *20*, 1770–1779.

Pascual-Leone, A., and Walsh, V. (2001). Fast Backprojections from the Motion to the Primary Visual Area Necessary for Visual Awareness. Science *292*, 510–512.

Passingham, R.E. (1985). Memory of monkeys (Macaca mulatta) with lesions in prefrontal cortex. Behavioral Neuroscience *99*, 3–21.

Pavlov, I.P., and Anrep, G.V. (2003). Conditioned reflexes (Mineola, N.Y: Dover Publications).

Pedregosa, Fabian, Varoquaux, Gael, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2825–2830.

Persaud, N., McLeod, P., and Cowey, A. (2007). Post-decision wagering objectively measures awareness. Nature Neuroscience *10*, 257–261.

Persuh, M., LaRock, E., and Berger, J. (2018). Working Memory and Consciousness: The Current State of Play. Frontiers in Human Neuroscience *12*.

Pessiglione, M., Schmidt, L., Draganski, B., Kalisch, R., Lau, H., Dolan, R.J., and Frith, C.D. (2007). How the brain translates money into force: a neuroimaging study of subliminal motivation. Science *316*, 904–906.

Pessiglione, M., Petrovic, P., Daunizeau, J., Palminteri, S., Dolan, R.J., and Frith, C.D. (2008). Subliminal Instrumental Conditioning Demonstrated in the Human Brain. Neuron *59*, 561–567.

Petrides, M. (1995). Impairments on nonspatial self-ordered and externally ordered working memory tasks after lesions of the mid-dorsal part of the lateral frontal cortex in the monkey. The Journal of Neuroscience *15*, 359–375.

Petrides, M., and Milner, B. (1982). Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. Neuropsychologia *20*, 249–262.

Pinker, S. (2016). The blank slate: the modern denial of human nature.

Pinto, Y., Sligte, I.G., Shapiro, K.L., and Lamme, V.A.F. (2013). Fragile visual short-term memory is an object-based and location-specific store. Psychonomic Bulletin & Review *20*, 732–739.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. Clin Neurophysiol *118*, 2128–2148.

Polonsky, A., Blake, R., Braun, J., and Heeger, D.J. (2000). Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. Nature Neuroscience *3*, 1153.

Popper, K.R., and Eccles, J.C. (1984). The self and its brain (London: Routledge).

Pribram, K.H., and Broadbent, D.E. (1970). Biology of memory (New York: Academic Press).

Prinz, J. (2000). A Neurofunctional Theory of Visual Consciousness. Consciousness and Cognition *9*, 243–259.

Prinz, J.J. (2010). When Is Perception Conscious? In Perceiving the World, B. Nanay, ed. (Oxford University Press), pp. 310–332.

Quentin, R., King, J.-R., Sallard, E., Fishman, N., Thompson, R., Buch, E., and Cohen, L.G. (2018). Differential brain mechanisms of selection and maintenance of information during working memory.

Quiroga, R.Q., Mukamel, R., Isham, E.A., Malach, R., and Fried, I. (2008). Human single-neuron responses at the threshold of conscious recognition. Proc. Natl. Acad. Sci. USA *105*, 3599–3604.

Ramsøy, T.Z., and Overgaard, M. (2004). Introspection and subliminal perception. Phenomenology and the Cognitive Sciences *3*, 1–23.

Raymond, J.E., Shapiro, K.L., and Arnell, K.M. (1992). Temporary suppression of visual processing in an RSVP task: an attentional blink? J Exp Psychol Hum Percept Perform *18*, 849–860.

Rees, G. (2007). Neural correlates of the contents of visual awareness in humans. Philosophical Transactions of the Royal Society B: Biological Sciences *362*, 877–886.

Rees, G., Kreiman, G., and Koch, C. (2002). Neural correlates of consciousness in humans. Nature Reviews Neuroscience *3*, 261–270.

Rensink, R.A., O'Regan, J.K., and Clark, J.J. (1997). To See or not to See: The Need for Attention to Perceive Changes in Scenes. Psychological Science *8*, 368–373.

Rescorla, R.A. (1988). Pavlovian conditioning. It's not what you think it is. Am Psychol *43*, 151–160.

Ress, D., and Heeger, D.J. (2003). Neuronal correlates of perception in early visual cortex. Nature Neuroscience *6*, 414–420.

Reuss, H., Kiesel, A., Kunde, W., and Hommel, B. (2011). Unconscious activation of task sets. Consciousness and Cognition *20*, 556–567.

Richet, C. (1884). L'Homme Et L'Intelligence: Fragments de Physiology et de Psychologie (Paris: Felix Alcan).

Riggall, A.C., and Postle, B.R. (2012). The Relationship between Working Memory Storage and Elevated Activity as Measured with Functional Magnetic Resonance Imaging. Journal of Neuroscience *32*, 12990–12998.

Roberts, B.M., Libby, L.A., Inhoff, M.C., and Ranganath, C. (2017). Brain activity related to working memory for temporal order and object information. Behavioural Brain Research.

Roelfsema, P.R. (2015). The role of the different layers of primary visual cortex in working memory. Journal of Vision *15*, 1406.

Rose, N.S., LaRocque, J.J., Riggall, A.C., Gosseries, O., Starrett, M.J., Meyering, E.E., and Postle, B.R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. Science *354*, 1136–1139.

Rosenthal, C.R., Andrews, S.K., Antoniades, C.A., Kennard, C., and Soto, D. (2016). Learning and Recognition of a Non-conscious Sequence of Events in Human Primary Visual Cortex. Current Biology *26*, 834–841.

Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., and Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review *16*, 225–237.

Rounis, E., Maniscalco, B., Rothwell, J.C., Passingham, R.E., and Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Cognitive Neuroscience *1*, 165–175.

Ruby, E., Maniscalco, B., Lau, H., and Peters, M.A.K. (2017). On a "failed" attempt to manipulate conscious perception with transcranial magnetic stimulation to prefrontal cortex.

Sackur, J., and Dehaene, S. (2009). The cognitive architecture for chaining of two mental operations. Cognition *111*, 187–211.

Sadaghiani, S., Hesselmann, G., and Kleinschmidt, A. (2009). Distributed and Antagonistic Contributions of Ongoing Activity Fluctuations to Auditory Stimulus Detection. Journal of Neuroscience *29*, 13410–13417.

Sakai, K., Rowe, J.B., and Passingham, R.E. (2002). Active maintenance in prefrontal area 46 creates distractor-resistant memory. Nature Neuroscience *5*, 479–484.

Salamé, P., and Baddeley, A. (1986). Phonological factors in STM: Similarity and the unattended speech effect. Bulletin of the Psychonomic Society *24*, 263–265.

Salti, M., Monto, S., Charles, L., King, J.-R., Parkkonen, L., and Dehaene, S. (2015). Distinct cortical codes and temporal dynamics for conscious and unconscious percepts. ELife *4*.

Schurger, A., and Sher, S. (2008). Awareness, loss aversion, and post-decision wagering. Trends in Cognitive Sciences *12*, 209–210.

Schurger, A., Pereira, F., Treisman, A., and Cohen, J.D. (2010). Reproducibility Distinguishes Conscious from Nonconscious Neural Representations. Science *327*, 97–99.

Schurger, A., Sarigiannidis, I., Naccache, L., Sitt, J.D., and Dehaene, S. (2015). Cortical activity is more stable when sensory stimuli are consciously perceived. Proc. Natl. Acad. Sci. U.S.A. *112*, E2083-2092.

Scoville, W.B., and Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. J. Neurol. Neurosurg. Psychiatry *20*, 11–21.

Serences, J.T., Ester, E.F., Vogel, E.K., and Awh, E. (2009). Stimulus-Specific Delay Activity in Human Primary Visual Cortex. Psychological Science *20*, 207–214.

Sergent, C. (in press). The offline stream of conscious representations. Philosophical Transcactions B.

Sergent, C., and Dehaene, S. (2004). Is Consciousness a Gradual Phenomenon?: Evidence for an All-or-None Bifurcation During the Attentional Blink. Psychological Science *15*, 720–728.

Sergent, C., Baillet, S., and Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. Nature Neuroscience *8*, 1391–1400.

Sergent, C., Wyart, V., Babo-Rebelo, M., Cohen, L., Naccache, L., and Tallon-Baudry, C. (2013). Cueing Attention after the Stimulus Is Gone Can Retrospectively Trigger Conscious Perception. Current Biology *23*, 150–155.

Shallice, T., and Warrington, E.K. (1970). Independent Functioning of Verbal Memory Stores: A Neuropsychological Study. Quarterly Journal of Experimental Psychology *22*, 261–273.

Shanks, D.R. (2017). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. Psychonomic Bulletin & Review *24*, 752–775.

Shepard, R.N., and Cooper, L.A. (1986). Mental images and their transformations (Cambridge, Mass.: MIT Press).

Shepard, R.N., and Metzler, J. (1971). Mental rotation of three-dimensional objects. Science *171*, 701–703.

Silvanto, J. (2017). Working Memory Maintenance: Sustained Firing or Synaptic Mechanisms? Trends in Cognitive Sciences *21*, 152–154.

Silvanto, J., Cowey, A., Lavie, N., and Walsh, V. (2005). Striate cortex (V1) activity gates awareness of motion. Nature Neuroscience *8*, 143.

Simons, D.J., and Ambinder, M.S. (2005). Change Blindness: Theory and Consequences. Current Directions in Psychological Science *14*, 44–48.

Simons, D.J., and Chabris, C.F. (1999). Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events. Perception *28*, 1059–1074.

Sklar, A.Y., Levy, N., Goldstein, A., Mandel, R., Maril, A., and Hassin, R.R. (2012). Reading and doing arithmetic nonconsciously. Proc. Natl. Acad. Sci. USA *109*, 19614–19619.

Sligte, I.G. (2010). Detailed sensory memory, sloppy working memory. Frontiers in Psychology *1*.

Sligte, I.G., Scholte, H.S., and Lamme, V.A.F. (2008). Are There Multiple Visual Short-Term Memory Stores? PLoS ONE *3*, e1699.

Sligte, I.G., Scholte, H.S., and Lamme, V.A.F. (2009). V4 Activity Predicts the Strength of Visual Short-Term Memory Representations. Journal of Neuroscience *29*, 7432–7438.

Smithson, H., and Mollon, J. (2006). Do masks terminate the icon? Quarterly Journal of Experimental Psychology *59*, 150–160.

Soto, D., and Silvanto, J. (2014). Reappraising the relationship between working memory and conscious awareness. Trends in Cognitive Sciences *18*, 520–525.

Soto, D., and Silvanto, J. (2016). Is conscious awareness needed for all working memory processes? Neuroscience of Consciousness *2016*, niw009.

Soto, D., Mäntylä, T., and Silvanto, J. (2011). Working memory without consciousness. Current Biology *21*, R912–R913.

Spaak, E., Watanabe, K., Funahashi, S., and Stokes, M.G. (2017). Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. The Journal of Neuroscience *37*, 6503–6516.

Sperling, G. (1960). The information available in brief visual presentations. Psychological Monographs: General and Applied *74*, 1–29.

Sprague, T.C., Ester, E.F., and Serences, J.T. (2014). Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load. Current Biology *24*, 2174–2180.

Sprague, T.C., Ester, E.F., and Serences, J.T. (2016). Restoring Latent Visual Working Memory Representations in Human Cortex. Neuron *91*, 694–707.

Squire, L.R. (2009). The Legacy of Patient H.M. for Neuroscience. Neuron *61*, 6–9.

Squire, L.R., and Wixted, J.T. (2011). The Cognitive Neuroscience of Human Memory Since H.M. Annual Review of Neuroscience *34*, 259–288.

Squire, L.R., and Zola-Morgan, S. (1988). Memory: brain systems and behavior. Trends in Neurosciences *11*, 170–175.

Sreenivasan, K.K., Curtis, C.E., and D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. Trends in Cognitive Sciences *18*, 82–89.

Standing, L. (1973). Learning 10000 pictures. Quarterly Journal of Experimental Psychology *25*, 207–222.

Stein, T., Kaiser, D., and Hesselmann, G. (2016). Can working memory be non-conscious? Neuroscience of Consciousness *2016*, niv011.

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

Stokes, M., and Spaak, E. (2016). The Importance of Single-Trial Analyses in Cognitive Neuroscience. Trends in Cognitive Sciences *20*, 483–486.

Stokes, M.G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., and Duncan, J. (2013). Dynamic Coding for Cognitive Control in Prefrontal Cortex. Neuron *78*, 364–375.

Stokes, M.G., Wolff, M.J., and Spaak, E. (2015). Decoding Rich Spatial Information with High Temporal Resolution. Trends in Cognitive Sciences *19*, 636–638.

Sugase-Miyamoto, Y., Liu, Z., Wiener, M.C., Optican, L.M., and Richmond, B.J. (2008). Short-Term Memory Trace in Rapidly Adapting Synapses of Inferior Temporal Cortex. PLoS Computational Biology *4*, e1000073.

Tadel, F., Baillet, S., Mosher, J.C., Pantazis, D., and Leahy, R.M. (2011). Brainstorm: A User-Friendly Application for MEG/EEG Analysis. Computational Intelligence and Neuroscience *2011*, 1–13.

Tallon-Baudry, C. (2009). The roles of gamma-band oscillatory synchrony in human visual cognition. Front Biosci (Landmark Ed) *14*, 321–332.

Taulu, S., Kajola, M., and Simola, J. (2004). Suppression of interference and artifacts by the Signal Space Separation Method. Brain Topogr *16*, 269–275.

Thibault, L., van den Berg, R., Cavanagh, P., and Sergent, C. (2016). Retrospective Attention Gates Discrete Conscious Access to Past Sensory Stimuli. PLoS ONE *11*, e0148504.

Tijus, C.A., and Reeves, A. (2004). Rapid iconic erasure without masking. Spat Vis *17*, 483–495.

Tononi, G. (2004). An information integration theory of consciousness. BMC Neurosci *5*, 42.

Tononi, G., and Edelman, G.M. (1998). Consciousness and complexity. Science *282*, 1846–1851.

Tononi, G., and Koch, C. (2008). The Neural Correlates of Consciousness: An Update. Annals of the New York Academy of Sciences *1124*, 239–261.

Tononi, G., and Koch, C. (2015). Consciousness: here, there and everywhere? Philosophical Transactions of the Royal Society B: Biological Sciences *370*, 20140167–20140167.

Tononi, G., Sporns, O., and Edelman, G.M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. Proc. Natl. Acad. Sci. USA *91*, 5033–5037.

Tononi, G., Edelman, G.M., and Sporns, O. (1998). Complexity and coherency: integrating information in the brain. Trends in Cognitive Sciences *2*, 474–484.

Trübutschek, D., Marti, S., Ojeda, A., King, J.-R., Mi, Y., Tsodyks, M., and Dehaene, S. (2017). A theory of working memory without consciousness or sustained activity. ELife *6*.

Tsuchida, A., and Fellows, L.K. (2009). Lesion Evidence That Two Distinct Regions within Prefrontal Cortex are Critical for *n* -Back Performance in Humans. J Cogn Neurosci *21*, 2263–2275.

Tsuchiya, N., and Koch, C. (2005). Continuous flash suppression reduces negative afterimages. Nature Neuroscience *8*, 1096.

Tsuchiya, N., Moradi, F., Felsen, C., Yamazaki, M., and Adolphs, R. (2009). Intact rapid detection of fearful faces in the absence of the amygdala. Nature Neuroscience *12*, 1224.

Tsuchiya, N., Wilke, M., Frässle, S., and Lamme, V.A.F. (2015). No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. Trends in Cognitive Sciences *19*, 757–770.

Tulving, E. (1972). Organization of Memory (New York: Academic Press).

Vergara, J., Rivera, N., Rossi-Pool, R., and Romo, R. (2016). A Neural Parametric Code for Storing Information of More than One Sensory Modality in Working Memory. Neuron *89*, 54–62.

Vogel, E.K., and Machizawa, M.G. (2004). Neural activity predicts individual differences in visual working memory capacity. Nature *428*, 748.

van Vugt, B., Dagnino, B., Vartak, D., Safaai, H., Panzeri, S., Dehaene, S., and Roelfsema, P.R. (2018). The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. Science *360*, 537–542.

Vul, E., Hanus, D., and Kanwisher, N. (2009). Attention as inference: Selection is probabilistic; responses are all-or-none samples. Journal of Experimental Psychology: General *138*, 546–560.

Wager, T.D., and Smith, E.E. (2003). Neuroimaging studies of working memory: a meta-analysis. Cogn Affect Behav Neurosci *3*, 255–274.

Ward, L.M. (2011). The thalamic dynamic core theory of conscious experience. Consciousness and Cognition *20*, 464–486.

Warden, M.R., and Miller, E.K. (2010). Task-Dependent Changes in Short-Term Memory in the Prefrontal Cortex. Journal of Neuroscience *30*, 15801–15810.

Warrington, E.K., and Shallice, T. (1969). The selective impairment of auditory verbal short-term memory. Brain *92*, 885–896.

Watanabe, K., and Funahashi, S. (2007). Prefrontal Delay-Period Activity Reflects the Decision Process of a Saccade Direction during a Free-Choice ODR Task. Cerebral Cortex *17*, i88–i100.

Watanabe, K., and Funahashi, S. (2014). Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. Nature Neuroscience *17*, 601–611.

Weibel, S., Giersch, A., Dehaene, S., and Huron, C. (2013). Unconscious task set priming with phonological and semantic tasks. Consciousness and Cognition *22*, 517–527.

Weiskrantz, L. (1996). Blindsight revisited. Current Opinion in Neurobiology *6*, 215–220.

Williams, M.A., Visser, T.A.W., Cunnington, R., and Mattingley, J.B. (2008). Attenuation of Neural Responses in Primary Visual Cortex during the Attentional Blink. Journal of Neuroscience *28*, 9890–9894.

Wolff, M.J., Ding, J., Myers, N.E., and Stokes, M.G. (2015). Revealing hidden states in visual working memory using electroencephalography. Frontiers in Systems Neuroscience *9*.

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

Wyart, V., and Tallon-Baudry, C. (2009). How Ongoing Fluctuations in Human Visual Cortex Predict Perceptual Awareness: Baseline Shift versus Decision Bias. Journal of Neuroscience *29*, 8715–8725.

Yang, E., Zald, D.H., and Blake, R. (2007). Fearful expressions gain preferential access to awareness during continuous flash suppression. Emotion *7*, 882–886.

Yang, E., Brascamp, J., Kang, M.-S., and Blake, R. (2014). On the use of continuous flash suppression for the study of visual processing outside of awareness. Front Psychol *5*, 724.

Yeh, S.-L., He, S., and Cavanagh, P. (2012). Semantic Priming From Crowded Words. Psychological Science *23*, 608–616.

Zeki, S. (2003). The disunity of consciousness. Trends in Cognitive Sciences *7*, 214–218.

Zeki, S. (2015). A massively asynchronous, parallel brain. Philosophical Transactions of the Royal Society B: Biological Sciences *370*, 20140174–20140174.

Zeki, S.M. (1978). Functional specialisation in the visual cortex of the rhesus monkey. Nature *274*, 423.