A Philosophical Examination of Working Memory

by

Max Hanson Beninger

Department of Philosophy
Duke University

Date:_____
Approved:

_____
Karen Neander, Co-Supervisor

_____
Felipe De Brigard, Co-Supervisor

_____
Owen Flanagan

_____
Marty Woldorff

_____
Carlotta Pavese

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Philosophy in the Graduate School
of Duke University

2019

ABSTRACT

A Philosophical Examination of Working Memory

by

Max Hanson Beninger

Department of Philosophy
Duke University

Date:_____
Approved:

_____
Karen Neander, Co-Supervisor

_____
Felipe De Brigard, Co-Supervisor

_____
Owen Flanagan

_____
Marty Woldorff

_____
Carlotta Pavese

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the Department of
Philosophy in the Graduate School of
Duke University

2019

Abstract

Working memory—the mental capacity to "hold on to" information after it ceases to be perceptually available—is one of the most discussed topics in psychology and neuroscience. Despite the importance of working memory in the sciences, however, there is only a small amount of philosophical research on the topic. The aim of my dissertation is to provide a philosophically-informed account of working memory, and to assess its relationship to other mental phenomena, including attention and consciousness.

In chapter one, I provide a broad historical overview of working memory. I begin by outlining William James' original distinction between "primary" and "secondary" memory, and work my way up to present-day neuroscientific investigations of working memory. One of the main conclusions of this chapter is that there is no working memory "module" in the brain. Instead, working memory is best conceptualized as a functionally-defined process that is potentially realized by multiple neural mechanisms.

In chapter two, I explore the link between working memory and attention. Recent evidence from psychology and neuroscience indicates that attention is (to some extent) involved in the process of working memory maintenance. However, it remains unclear whether the contents of working memory are *always* attended, or if working memory representations can be dynamically shifted in and out of the focus of attention. Drawing on empirical and phenomenological data, I argue that the second view is

correct. Although attention plays an important role in working memory maintenance, working memory representations can persist—at least temporarily—outside the focus of attention.

Chapter three addresses a related question: namely, how working memory relates to consciousness. I distinguish three possible positions on this score: (i) working memory representations are always conscious; (ii) working memory representations can be either conscious or unconscious, but they are all *accessible* to consciousness; and (iii) working memory representations can be either conscious or unconscious, and some are *inaccessible* to consciousness. Based on the available empirical data, I argue in favor of position (ii). Evidence suggests that working memory representations can be unconscious, but such unconscious representations still appear to be consciously accessible, in the sense that they can be brought to consciousness at will.

Finally, in chapter four, I provide a critique of Peter Carruthers' recent sensory-based account of working memory. According to Carruthers, attention only targets "mid-level" sensory areas, and thus the representations held in working memory will necessarily be sensory based in nature. I disagree. I point out that there is some evidence for attentional modulation outside of modality-specific sensory areas. I also highlight several empirical studies which provide preliminary support for the existence of non-sensory (i.e., amodal) working memory representations.

Dedication

To Sam—I couldn't have done it without you.

Contents

List of Figures

Acknowledgements

I would like to thank all the members of my committee—Karen, Felipe, Owen, Marty and Carlotta—for their input over the years. My work has benefitted immensely from their feedback, both written and verbal. I am especially grateful to Karen for her unwavering support. Karen has played a key role in my philosophical development, and I owe much of my success to her sage advice and kindness. My dissertation may not be on teleosemantics, but it is nonetheless heavily influenced by the empirically-oriented approach that Karen encouraged me to pursue. Felipe also deserves special mention. It was his class on attention and consciousness that initially got me thinking about working memory. Since then, Felipe has always been generous with his time, and has helped me in any way he can.

My family and friends have also been wonderfully supportive during my time at Duke. Thanks Mom, Dad, Mari, Grandma, Judith and Sunil for making every trip home a fun and relaxing experience. And thanks to the many friends—Ben, Eric, Jordan, Alissa and the entirety of "cool cats"—who helped distract me when the dissertation got hard to bear. Finally, thank you to my wonderful wife, Sam, for always being there for me, and for reading a substantial portion of my chapters. You have certainty made this dissertation better than it would have been otherwise—and you've made me happier while writing it.

1. Introduction

Briefly scan the following three symbols—♠ Ω ⊗ —then try to keep these symbols in mind for a few moments, without looking back at the page. In performing this task, you just made use of what psychologists and neuroscientists call "working memory". Working memory can be described (at first pass) as the cognitive process that enables the "temporarily maintenance and manipulation of information" (Baddeley, 1992, p. 281; 2014). It is what allows us to mentally "hold on to" representations of stimuli that are no longer currently being perceived. Working memory is a central part of cognition, and it plays a major role in our everyday lives. You use working memory when you hold in mind a telephone number that you have just been told, and you use working memory when solving a math problem in your head (Baddeley, 2014). Indeed, working memory seems to be involved in *any* kind of stimulus-independent thought that requires the maintenance and/or manipulation of information.

While working memory has been studied intensively by psychologists and neuroscientists over the past half-century, it has only recently begun to receive attention from philosophers (Block, 2007; Prinz, 2012; Carruthers, 2013, 2014, 2015; Gomez-Lavin, 2017, *in preparation*). This neglect is unfortunate, as the notion of working memory stands in need of conceptual clarification. Although working memory is discussed widely in scientific circles, there is no general consensus regarding its nature and essential properties (see Miyake & Shah, 1999 for a compilation of dissenting views).

Furthermore, it is also unclear how working memory is related to other mental phenomena, such as attention and consciousness (Soto & Silvanto, 2014; LaRocque, Lewis-Peacock & Postle, 2014). Working memory is thus ripe for philosophical investigation. We need a clear articulation of the relevant issues surrounding working memory, and we need to carefully consider how the empirical evidence bears on the issues in question.

In the present chapter, I do two things: first, I provide an overview of the empirical literature on working memory and, second, I develop a preliminary account of the nature of working memory. I argue that we should reject "reductionist" accounts of working memory that attempt to identify working memory with a specific neural mechanism. Instead, I suggest that working memory is best understood as a functionally-defined process, which exhibits four key properties: (i) information maintenance, (ii) information manipulation, (iii) capacity limitations, and (iv) distractor resistance. Thus, on my view, there need not be a single neural mechanism for working memory in the brain—rather, working memory is unified at the *functional level*. The chapter is organized as follows. Section 1.1 outlines the historical development of the concept of working memory in cognitive psychology, while section 1.2 surveys the neuroscientific literature on the neural underpinnings of working memory. Section 1.3 sets up the dispute between reductionist and anti-reductionist approaches to working memory; it is in this section that I argue for a *functionalist* (anti-reductionist) account of

the nature of working memory. Finally, section 1.4 presents three further unanswered questions concerning working memory. These questions serve as the guiding topics for the following three chapters of my dissertation.

1.1 Historical preliminaries

The origins of the modern concept of working memory can be traced back to William James' *Principles of Psychology* (1890/1983).[1] Here, James draws a distinction between two different kinds of memory: *primary memory* and *secondary memory* (James, 1890/1983, p. 608). Secondary memory, for James, corresponds to our everyday understanding of memory. It refers to information that is encoded, stored (in the absence of consciousness), and then gets returned to consciousness awareness at a later point in time. Primary memory, by contrast, refers to information that is *continuously experienced*—it constitutes our "memory" for things that were just perceived and are still actively being held in mind. To quote James directly:

> An object which is recollected, in the proper sense of that term [i.e., from secondary memory], is one which has been absent from consciousness altogether, and now revives anew. It is brought back, recalled, fished up, so to speak, from a reservoir in which, with countless other objects, it lay buried and lost from view. 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. (James, 1890/1983, p. 608)

[1] Though, even before James, there were various precursor concepts. See Logie (1996).

James' dichotomy between primary and secondary memory clearly presages contemporary classifications of memory, which often separate working memory (or short-term memory) from long-term memory (Atkinson & Shiffrin, 1968; Baddeley, 2014). Interestingly, though, James uses *consciousness* as a criterion to distinguish primary and secondary memory. For James, the main difference between primary and secondary memory is that secondary memories cease to be conscious at some point, whereas primary memories are continuously conscious throughout their lifespan. This emphasis on consciousness was largely abandoned with the rise of scientific psychology. Many subsequent researchers retained James' distinction between a temporary store for information that is currently being used, and a long-term store for passively retained information; however, this distinction is often cashed-out in terms of differences in *information processing*, rather than in terms of *consciousness* (see, e.g., Waugh & Norman, 1965; Atkinson & Shiffrin, 1968).

1.1.1 Atkinson & Shiffrin's model

Jumping ahead several decades (since James), one of the most important models of memory from the 20th century is the "modal model" developed by Richard Atkinson and Richard Shiffrin (1968). Atkinson and Shiffrin drew a tripartite distinction among three different components of memory: (i) the sensory register, (ii) the short-term store, and (iii) the long-term store (Figure 1). Atkinson and Shiffrin's long-term store is roughly analogous to James' secondary memory, whereas the sensory register and short-

term store can perhaps be viewed as different sub-components of what James' called "primary memory".

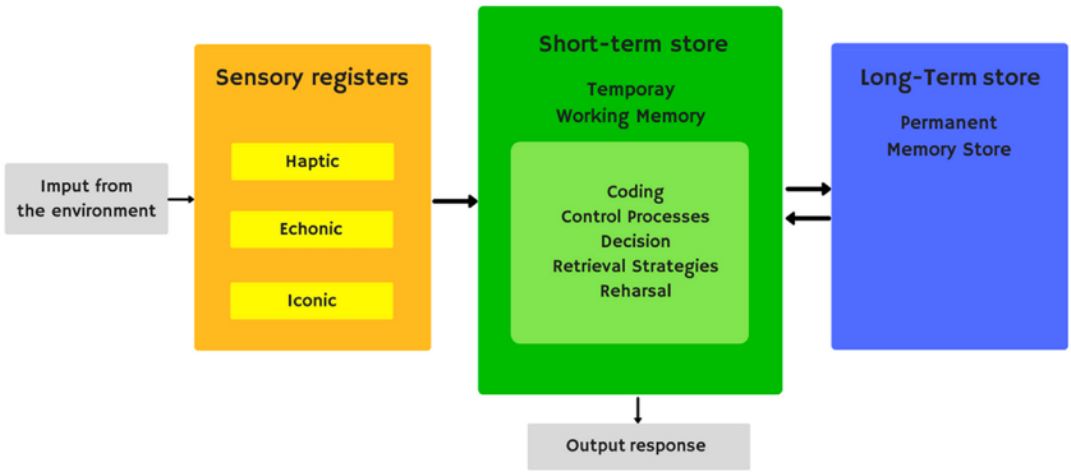

Figure 1: A depiction of Atkinson and Shiffrin's (1968) model. Arrows indicate the direction of information flow. From Camina & Güell (2017, p. 4). Copyright © 2017, Camina & Güell. Reprinted under the terms of the Creative Commons Attribution License (CC BY).

The sensory register is the first step in mnemonic processing: it serves as a temporary buffer that preserves sensory traces of incoming stimuli for brief durations (on the order of milliseconds or seconds). According to Atkinson and Shiffrin, each modality likely has its own dedicated sensory register that operates on a specific kind of perceptual information (Atkinson & Shiffrin, 1968, p. 95-96). Next, there is the short-term store, which receives input from the sensory register. Atkinson and Shiffrin claim that the short-term store "may be regarded as the subject's 'working memory'" (1968, p. 92), which holds information in an accessible state and mediates the transfer of information into the long-term store. Representations in the short-term store are believed to decay

over a 30-second period, but they can be prolonged via an active rehearsal processes (Atkinson & Shiffrin, 1968, p. 90-91). Finally, the long-term store is the end-point of memory processing, serving as a vast repository for stored knowledge. Information in the long-term store is thought to be "relatively permanent", although Atkinson and Shiffrin admit that "it may be modified or rendered temporarily irretrievable as the result of other incoming information" (Atkinson & Shiffrin, 1968, p. 93). Importantly, in Atkinson and Shiffrin's model, information can also flow *backwards* from the long-term store into the short-term store. This corresponds to the process of memory recall, whereby we actively call up information that previously lay dormant within the long-term store.

Atkinson and Shiffrin support their tripartite model by appealing to a variety of empirical findings. First, they note that the distinction between the sensory register and short-term store is motivated by George Sperling's now-famous partial report studies (Sperling, 1960; 1963). In his work, Sperling presented participants with 12 letters (organized in a 3x4 array) for 15-500 milliseconds, followed by a brief delay and reporting phase (Figure 2). Sperling found that participants could recall about four letters from the array in total. Crucially, however, if an auditory cue was presented just after the array—indicating which row of letters was to be recalled—participants were capable of recalling most (3/4) of the letters from the cued row. This indicates that the majority of the letters in the array remained temporarily accessible (until the time of the

auditory cue), even though only a much smaller number could be later reported. Sperling interpreted these results as showing that we have a very brief visual store (termed "iconic memory") with a capacity that exceeds that of ordinary short-term memory (Sperling, 1963, p. 21). If the auditory cue is presented quickly enough after the stimulus array, participants are capable of *selecting* which row will be transferred from iconic memory into the more durable short-term memory store.

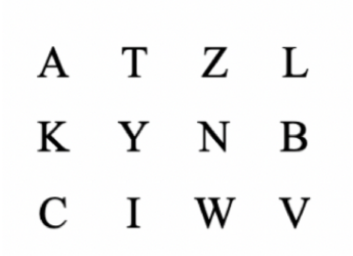

Figure 2: An example of a Sperling array, as used by Sperling (1960).

Atkinson and Shiffrin also present empirical evidence in favor of the distinction between the short-term store and long-term store (Atkinson & Shiffrin, 1968, p. 96-98). One major motivation for distinguishing between the short-term and long-term components of memory comes from studies of amnesic patients with hippocampal lesions. Hippocampal-lesion patients often display severe deficits in long-term memory, being unable to encode and store new memories over prolonged periods of time. However, such patients can still perform well on short-term memory tests, which require them to actively rehearse a set of verbal items for several seconds or minutes (Milner, 1966). According to Atkinson and Shiffrin, the fact that long-term memory can be damaged without any corresponding effects on short-term memory suggests that the

long-term store and short-term store constitute separate memory systems. Additionally, Atkinson and Shiffrin also point to *capacity limitations* as a potential feature that distinguishes the short- and long-term stores. Unlike the long-term store, which is capable of retaining vast amounts of information about one's past experiences, the short-term store can only maintain a few separate items at a time. Following Miller (1956), Atkinson and Shiffrin suggest that the maximum capacity of the short-term store "is usually in the range of five to eight [items]" (Atkinson & Shiffrin, 1968, p. 112).

Atkinson and Shiffrin's model was, and still is, highly influential. The basic distinctions that Atkinson and Shiffrin draw—between sensory memory, short-term memory, and long-term memory—are still widely recognized as delineating separate components of human memory (Baddeley, 2014). Certain aspects of the model have since been revised, however. For instance, Atkinson and Shiffrin assumed that memory encoding involved a linear progression, with any long-term memory encoding requiring the mediation of the short-term store. This assumption was challenged by Shallice and Warrington (1970), who identified a patient (K.F.) with intact long-term memory encoding abilities, yet severely impaired short-term/working memory. As Shallice and Warrington point out, K.F.'s intact long-term memory performance suggests that information may be transferred directly to the long-term store, without necessarily having to pass through short-term memory (Shallice & Warrington, 1970, p. 270).

1.1.2 Fractionating the short-term store: Baddeley & Hitch's multi-component model

Another issue with Atkinson and Shiffrin's model concerns the *unitary* nature of the short-term store: Atkinson and Shiffrin depicted the short-term store as a single "box" that operated over all different types of information. This unitary conception of the short-term store was famously challenged by Baddeley and Hitch in their seminal 1974 paper, *Working Memory*. In this paper, Baddeley and Hitch surveyed results from a number of *dual-task studies*, which required participants to perform two separate tasks designed to load different aspects of short-term memory. In one such experiment, for instance, participants had to remember a list of six letters while also performing a secondary reasoning task. Specifically, participants were shown a list of 32 statements—each matched with a letter string of the form **AB** or **BA**—and their task was to report whether the statement was true or false of the corresponding string, e.g.:

> B is followed by A—BA
> A does not follow B—BA
> A is not followed by B—BA
> (Baddeley, 2014, p. 44)

As hypothesized, Baddeley and Hitch found that participants in the dual-task condition often performed slightly worse on the two tasks, relative to controls who only had to perform one of the tasks by itself. However, as Baddeley and Hitch note, "the degree of disruption observed … was far from massive" (Baddeley & Hitch, 1974, p. 75). Participants could usually still recall *most* of the six remembered letters (which is close to

the supposed capacity limit of working memory) while simultaneously performing the secondary reasoning task. On the basis of these results, Baddeley and Hitch suggest that short-term memory may consist of multiple separate *subcomponents*, with different tasks taxing different parts of the system.

Baddeley and Hitch then go on to develop a new "multi-component model" of working memory (they explicitly opt to use the term "*working* memory" to emphasize the fact that the system in question is involved in the *processing* and *use* of information). On Baddeley and Hitch's multi-component model, working memory has three separate parts: (1) the *phonological loop*, (2) the *visuospatial sketchpad*, and (3) the *central executive* (Baddeley & Hitch, 1974; Baddeley, 2014). The phonological loop and visuospatial sketchpad have purely storage-related functions: they are responsible for the temporary maintenance of verbal and visuospatial information. The central executive, by contrast, acts as an "overseer", guiding the focus of attention and initiating the *rehearsal* and *manipulation* of information held within the two storage buffers (Figure 3).[2] According to Baddeley and Hitch, this model can help us explain participants' ability to perform well in dual-task studies. In the case of the above experiment, remembered letters could be passively sorted in the phonological loop, while the central executive dealt with the concurrent reasoning task (Baddeley & Hitch, 1974, p. 77). Baddeley and Hitch also note

[2] At least initially, Baddeley and Hitch assumed that the central executive was capable of storage too, being able to "recode" information held in the storage buffers (Baddeley & Hitch, 1974, p. 78).

that the proposed separation of the verbal and visuospatial buffers can help explain patterns of dual-task performance in other contexts. For example, Baddeley, Grant Wight and Thomson (1975) showed that participants have severe difficulty performing two concurrent visuospatial tasks, but have no problem performing one visuospatial task combined with a secondary verbal task. This makes sense on the above multi-component model, because verbal and visuospatial information are hypothesized to be stored independently—and can thus coexist without interfering with one another.

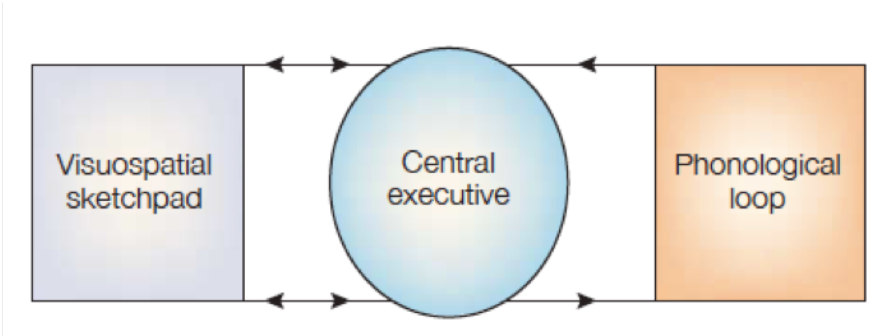

Figure 3: A depiction of Baddeley and Hitch's original multi-component model. From Baddeley (2003, p. 830). Copyright © 2003 Nature Publishing Group. Reprinted with permission.

Baddeley and Hitch's multi-component model has enjoyed great popularity in psychology. Indeed, until the early 2000's, it was generally considered to be the "default" model of working memory (Baddeley, 2003). In recent years, however, the multi-component model has been challenged in a variety of ways. One issue with the multicomponent model concerns its *scope*. As noted by Postle (2006) and Carruthers (2015), there is evidence for working memory in other modalities, such as touch and

smell (Harris, Miniussi, Harris, & Diamond, 2002; Dade, Zatorre, Evans & Jones-Gottman, 2001). This evidence is problematic for the multicomponent model—at least in its original form—because it indicates that working memory encompasses more than just the visuospatial sketchpad and phonological loop. More seriously, the very idea that working memory should be conceived of as a set of storage buffers has recently been called into question. Contemporary neuroscientific results (which will be discussed in more detail in the next section, 1.2) suggest that working memory storage does not take place within specialized buffers, but rather works by offloading representations to primary sensory-processing areas (Postle, 2006; Carruthers, 2015; D'Esposito & Postle, 2015). Thus, although Baddeley and Hitch may have been right that verbal and visuospatial information is stored separately, such information may turn out to be held in sensory cortices instead of specialized storage buffers.

1.2 The neuroscience of working memory

In tandem with psychological research, neuroscientists have also begun to uncover the neural mechanisms that underlie working memory. One of the first neuroscientific studies of working memory was conducted by Joaquin Fuster and Garrett Alexander (1971). Fuster and Alexander trained rhesus monkeys to perform a simple delayed-response task, which involved remembering the location of a piece of food over a short delay (15-60 seconds). During the performance of the task, Fuster and Alexander also recorded the monkeys' neural activity via cortically-implanted

electrodes. They found that certain groups of neurons—specifically, neurons in the prefrontal cortex—exhibited elevated activity over the delay-period. Prefrontal activity during working memory tasks has since been reported by numerous other studies (Niki & Watanabe, 1976; Funahashi, Bruce & Goldman-Rakic, 1989; Goldman-Rakic, 1995). Moreover, the effect is not specific to monkeys: human neuroimaging studies also find that prefrontal regions tend to display heightened activation while information is being retained in working memory (McCarthy et al., 1994; Smith & Jonides, 1999).

The finding that prefrontal areas are active during working memory maintenance initially led many neuroscientists to conclude that the prefrontal cortex was the primary site of working memory storage (see, e.g., Goldman-Rakic, 1992, 1995; Funahashi & Kubota, 1994). According to this view, championed by Patricia Goldman-Rakic (1992, 1995), the prefrontal cortex contains a set of anatomically distinct storage modules—akin to buffers proposed by Baddeley and Hitch (1974)—which are each specialized for a different type of information. This conclusion is suspect, however. The fact that the prefrontal cortex is active during working memory maintenance does not, by itself, show that the prefrontal cortex is operating as a working memory storehouse. After all, the activity in the prefrontal cortex might reflect some *other* function aside from informational maintenance (see Postle, 2006). Ironically, Fuster and Alexander acknowledged this possibility in their seminal 1971 paper. They suggested that

prefrontal activity does not *itself* encode the contents of working memory[3], but is instead "related to the focusing of attention by the animal on information that is being or has been placed in temporary storage" (Fuster & Alexander, 1971, p. 654).

More recent research has helped to illuminate the function of the prefrontal cortex in relation to working memory (Postle, 2006, 2007; D'Esposito & Postle, 2015). As it turns out, the data appear to support Fuster and Alexander's "attentional control" interpretation. That is, prefrontal activity seems to be more related to the guidance of attention than the actual storage of working memory contents. Several lines of evidence support this view. First of all, lesion studies have found that patients with significant prefrontal damage can still perform normally on simple working memory span tests—e.g., tests where participants are required to hold in mind a list of numbers or words (D'Esposito & Postle, 1999). This suggests that the prefrontal cortex may not be critical for working memory storage, since storage can remain intact in patients with widespread prefrontal damage. Notably, prefrontal damage does *sometimes* affect performance on working memory tasks, but it appears to selectively impair performance on tasks that involve shifts of attention or distractor stimuli (Chao & Knight, 1998;

---

[3] Following standard scientific usage, I use the expression "contents of working memory" to refer to the information or representations stored in working memory. This is distinct from how philosophers typically use the word "content". For philosophers, the word "content" is often taken to refer to the intentional object that a mental representation is directed towards (Neander, 2017).

Pasternak, Lui & Spinelli, 2015; Postle, 2017). The fact that prefrontal damage particularly affects tasks that implicate attentional control supports the view that the prefrontal cortex has an executive function, instead of serving as a storehouse for memory representations.

Recent neuroimaging studies employing more sensitive forms of data analysis—such as multivoxel pattern analysis[4]—also indicate that working memory storage occurs outside the prefrontal cortex. These studies find that during working memory maintenance, fine-grained information about working memory items can often be detected in early sensory cortices (Harrison & Tong, 2009; Serences, Ester, Vogel & Awh, 2009; Kumar et al., 2016; see Postle, 2015 for a helpful review). One seminal example comes from a functional magnetic resonance imaging (fMRI) study by Harrison and Tong (2009). In this study, participants were shown two sequential line gratings of different orientations, followed by a cue indicating which grating they were required to remember. After a delay of eleven seconds, participants were shown a third "test"

---

[4] Multivoxel pattern analysis (MVPA) is a recently-developed technique that involves applying machine learning algorithms to neuroimaging data. In a typical MVPA study, a pattern classification algorithm is "trained" to recognize patterns of activity specific to different types of stimuli (e.g., faces and houses) using a set of training data (in the face/house case, the training data would consist of brain scans of participants looking at faces or houses). Next, the classification algorithm is "tested" by applying it to novel imaging data; the aim here is to see whether the algorithm can accurately sort *new* brain scans into the relevant categories. Importantly, MVPA can be applied to neuroimaging data from *specific brain regions*. For instance, we can assess whether activity in V1 can be used to predict whether a participant is looking at a face or a house. This is helpful, because it tells whether the brain region in question (V1) differentially encodes the two classes of stimuli. For a helpful summary of MVPA, see Norman, Polyn, Detre and Haxby (2006).

grating, and their task was to report the test grating's direction of rotation—clockwise or counterclockwise—relative to the initial cued grating. Using machine learning techniques, Harrison and Tong showed that it was possible to *accurately infer* the remembered grating's orientation based on the observed delay-period activity in areas of the visual cortex (V1-V4). In other words, they found that "orientations in working memory can be *decoded* from activity patterns in the human visual cortex" (Harrison & Tong, 2009, p. 632; emphasis added). The fact that visual cortical activity is predictive of the content of working memory—in this case, the orientation of the remembered grating—suggests that these areas help store visual working memory representations.

On the basis of the above findings (and others like them) many researchers now believe that the main function of the prefrontal cortex in working memory is to serve as an executive attentional controller (Curtis & D'Esposito, 2003; Postle, 2006; Carruthers, 2015; Lara & Wallis, 2015). According to this view, the prefrontal cortex generates attentional signals that bring about activity in posterior sensory areas (Gazzaley & Nobre, 2012; Carruthers, 2015). This model is often referred to as the "sensory recruitment model" because it takes working memory to employ sensory regions to represent the relevant memoranda (D'Esposito & Postle, 2015). The sensory-recruitment model stands in stark contrast to Baddeley and Hitch's original multicomponent model of working memory: whereas the multi-component model takes working memory to consist of specialized storage buffers, the sensory-recruitment model tells us that

working memory offloads the storage of information to the very same sensory areas that are involved in encoding perceptual input. It is worth noting that the term "*sensory*-recruitment" may be overly restrictive, however. While evidence suggests that working memory representations are *often* stored in a sensory format, it remains possible that working memory may also maintain more abstract representations too. Indeed, in my fourth chapter I will argue for a "multi-level" model of working memory storage, according to which working memory can include *both* sensory representations *and* non-sensory (i.e., amodal) representations (see also Christophel et al., 2017).

Finally, before we move on, it is necessary to talk about one more development in the neuroscientific literature. Until recently, it was standardly assumed that working memory is an *active process*—that is, that working memory representations are always encoded in the above-baseline firing of neural populations (see, e.g., Luck & Vogel, 2013). In the past few years, however, a number of studies have challenged this pervasive assumption. For example, Lundqvist and colleagues (2016) used intracranial electrodes to record neural activity in monkeys during the performance of a visual working memory task. They found that the spiking of neural populations was *intermittent*—occurring at regular intervals—rather than remaining constant over the entire delay (Lundqvist et al., 2016, p. 152). According to Lundqvist and colleagues, this result challenges the "sustained activity" view, showing that working memory maintenance is also characterized by periods of inactivity. They further suggest that

previous reports of sustained activity may have been an experimental artifact, resulting from looking at *average activity* of neurons across the entire delay period (Lundqvist, Herman & Miller, 2018, p 7014).

Additionally, several recent human neuroimaging studies also support a dissociation between working memory and neural activation (Lewis-Peacock, Drysdale, Oberauer & Postle, 2012; Rose et al., 2016; Wolff, Jochim, Akyürek & Stokes, 2017). For instance, Lewis-Peacock and colleagues (2012) used multivariate decoding techniques to assess the strength of working memory representations, depending on whether they were currently being prioritized by attention. They found that neural representations tended to vary in the strength as a function of behavioral relevance. Indeed, if a representation was not immediately relevant for the task at hand, the level of activation for this representation would drop to baseline (that is, there would be no evidence of neural activation). However, such representations could still be "brought back online"—i.e., reinstantiated in the form of active patterns of neural firing—if they later *became* relevant for a subsequent portion of the task. (See Chapter 2, §4.2 for a more in-depth summary of the results of Lewis-Peacock et al.)

These results have prompted researchers to rethink the identification of working memory with persistent activity. Indeed, there is now a growing movement to recognize that working memory may depend, in part, on *synaptic plasticity* (Mongillo, Barak, & Tsodyks, 2008; LaRocque et al., 2014; D'Esposito & Postle, 2015; Trübutschek et al., 2017;

Lundqvist, Herman & Miller, 2018). The idea, roughly speaking, is that information may be transiently stored via "activity-silent" synaptic changes, possibly involving alterations in the residual calcium concentrations of previously-active assemblies (Mongillo et al., 2008). These synaptic changes are hypothesized to retain information about the remembered stimulus for short-durations, ranging from milliseconds to seconds. Moreover, such synaptic memories are thought to be *recoverable*, in the sense that they can be returned to state of heightened activation via top-down executive control. Although the existence of synaptic forms of working memory remains somewhat speculative—owning to the fact that it is primarily motivated by null findings—the notion is gaining increasing traction within the field.

1.3 The metaphysics of working memory

In the previous two sections, I provided a brief history of the scientific study of working memory. A critical issue still remains to be addressed, however: we do not have a full account of what exactly working memory *is*, ontologically speaking, or what its essential properties are. As philosophers might put, we are still lacking an account of the "metaphysics" of working memory. The present section aims to provide a preliminary account of the metaphysics (or nature) of working memory.

1.3.1 Reductionism vs. anti-reductionism

A central question concerning the nature of working memory is whether it is unified at the neural level. That is, we can ask whether working memory is reducible to

a *single* neural process or mechanism in the brain, or if it is a higher-level kind that is potentially realized by multiple neural mechanisms.[5] There has been considerable disagreement about this in the literature. On the one hand, some researchers adopt a reductionist approach, claiming that there is indeed a specific neural correlate for working memory—at least in humans and non-human primates (Goldman-Rakic, 1992, 1995; Carruthers, 2015). On the other hand, some researchers adopt an anti-reductionist approach, claiming that multiple neural mechanisms contribute to working memory maintenance (Soto & Silvanto, 2014, 2016; D'Esposito & Postle, 2015; Gomez-Lavin, 2017).

Patricia Goldman-Rakic's prefrontal buffer view is an admirably clear example of the reductionist approach (Goldman-Rakic, 1992, 1995). According to Goldman-Rakic, working memory is constituted by the persisting firing of neurons in the prefrontal cortex, with different subsets of prefrontal neurons coding for different stimulus features. She claims that "neurons in the prefrontal cortex possess what we call 'memory fields': when a particular target disappears from view, an individual prefrontal neuron switches into an active state, producing electrical signals at more than twice the baseline rate" (Goldman-Rakic, 1992, p. 113). On Goldman-Rakic's view, then, working memory

[5] We could frame this question in a different way, by asking whether working memory is a "natural kind" (Carruthers, 2015, p. 180; Gomez-Lavin, 2017), however this terminology would only confuse things, since there are multiple conflicting accounts of what constitutes a natural kind (see Bird, 2018).

representations are identified with above-baseline activity in select memory neurons in the prefrontal cortex. A more contemporary reductionist account of working memory is provided by Peter Carruthers (2015). Unlike Goldman-Rakic, Carruthers takes working memory representations to be stored in sensory regions of the brain; however, Carruthers still thinks that there is a reasonably specific neural process that is common to all instances of working memory. According to Carruthers, working memory involves representations in sensory areas being targeted by top-down attention and thereby getting "globally broadcast" via a prefronto-parietal workspace network (Carruthers, 2015, p. 76-9). Although much more could be said about Carruthers' account (see chapter 4), the basic point here is that he takes working memory to be identifiable with a particular neural process—namely, the (active) broadcasting of sensory representations.

There are, however, major difficulties facing such reductionist views. First, as we saw in section 1.2, Goldman-Rakic's prefrontal buffer view is challenged by studies showing that working memory representations are often stored in early sensory cortices. More seriously, both reductionist views—Goldman-Rakic's prefrontal buffer view and Carruthers' broadcasting view—seem to be at odds with recent research on "activity-silent" working memory. Recall, for instance, that a number of studies have found that neural activity during working memory tasks can be discontinuous, sometimes dropping to baseline (Lewis-Peacock et al., 2012; Lundqvist et al., 2016, 2018). Such findings suggest that working memory may depend not only on active neural

mechanisms (like persistent firing), but also on synaptic plasticity (Mongillo et al., 2008; Lundqvist et al., 2018). Indeed, in light of the recent work on activity-silent working memory, many researchers are now explicitly endorsing the view that working memory depends on multiple heterogenous neural mechanisms. For instance, Soto and Silvanto (2016, p. 2) assert that "[t]here are at least a few neural mechanisms for WM functions, including: synaptic mechanisms for information maintenance (i.e., through calcium kinetics in task-relevant neural substrates) (Mongillo et al. 2008), persistent neural firing (Sreenivasan et al. 2014), and oscillatory network coherence (Palva and Palva 2012)."

I am inclined to agree with Soto and Silvanto's anti-reductionist conclusion. While we may not yet have conclusive proof of activity-silent working memory, the evidence certainly seems to be pointing in that direction. Moreover, *a priori* considerations also marshal against a reductionist view: it seems implausible that a highly-complex and multifaceted phenomenon like working memory would rely on only *one* single neural mechanism (Poldrak & Yarkoni, 2016, p. 599). Importantly, however, rejecting a reductionist account of working memory does not mean we have to abandon the concept of working memory altogether. Another option is available. Specifically, we could take working memory to be a high-level *functional* (or *computational*) kind that abstracts away from the implementation details (Persuh, LaRock & Berger, 2018; De Brigard, 2012; Wu, 2014a, Ch. 2; Marr, 1982). This move is analogous to the "functionalist" approach in the philosophy of mind, which takes mental

states/processes to be identified with functional roles, rather than specific neural states/processes. On a functionalist account of working memory, we are still committed to the *existence* of working memory; however, instances of working memory are unified in virtue of their functional properties, rather than some shared neural mechanism.[6] Of course, merely asserting that working memory is a functional kind does not, by itself, settle the issue of the nature of working memory. After all, if working memory *is* a functional kind (as I suggest), then we will need to identify its relevant functional properties.

1.3.2 A functional account of working memory

So, what *are* the essential functional characteristics of working memory? A natural answer—which I take to be a good starting point—is that working memory involves the *maintenance* and *manipulation* of information. After all, these two characteristics feature prominently in most (if not all) definitions of working memory (Baddeley, 1992; Miyake and Shah, 1999; Gomez-Lavin, *in preparation*).

---

[6] It is worth noting that this functionalist approach is in line with the conception of working memory originally put forth by Baddeley and Hitch (1974). Baddeley and Hitch were working within the tradition of cognitive psychology: their research attempted to probe the nature of working memory by showing how it operated in the context of different behavioral tasks. While Baddeley and Hitch certainly cared about the neural implementation of working memory, their primary interest was in delineating the *functional properties* of working memory, including its storage limitations and how different components of working memory (e.g., the visuospatial sketchpad and phonological loop) interacted with one another.

The notion of maintenance is fairly self-explanatory. To say that working memory "maintains" information is just to say that it is capable of preserving the relevant information over time, and that this information can be subsequently used to guide behavior. Manipulation, however, requires a bit more discussion. When we say that the representations of working memory are "manipulable" we usually mean that they can be *altered* in some way. For instance, we can rearrange a sequence of remembered letters to be in alphabetical order (Fougnie & Marois, 2007), and we can mentally rotate a three-dimensional shape to see how it might look from a different perspective (Shepard & Metzler, 1971). Importantly, I take manipulation to be a *personal-level phenomenon*—that is, it is something voluntarily performed by the subject. Thus, a low-level malfunction of the visual system that results in a change in the content of a representation will not count as a genuine case of manipulation, on my view, since it was not undertaken by the subject. Another important caveat to make is that working memory representations do not *always* have to be manipulated (Persuh et al., 2018). After all, many classic working memory tasks merely require participants to maintain a set of stimuli, without manipulating them in any way (see, e.g., Luck & Vogel, 1997). Strictly speaking, then, the correct thing to say is that working memory representations are *available* for manipulation, rather than always being manipulated (see also Persuh et al., 2018).

Unfortunately, however, the above two properties—maintenance and manipulability—may not be sufficient to delineate working memory. After all, even *long-term memories* involve some kind of informational maintenance, and long-term memories are at least potentially available for manipulation too (though perhaps less immediately than working memory representations). To remedy this, I suggest that we appeal to another classic feature of working memory: *capacity limitations*. By specifying that working memory is capacity-limited in nature we clearly distinguish it from long-term memory, which seems to have a nearly unlimited storage capacity (Cowan, 2008a). Importantly, in claiming that working memory is capacity-limited I am not taking a stand on the *exact nature* of these capacity limits, which is a highly contentious issue in the empirical literature. Some researchers claim that the capacity of working memory is best understood as a series of discrete slots, with most people being able to store around four or five separate representations at a time (Luck & Vogel, 1997; Cowan, 2010). By contrast, other researchers hold that working memory capacity is best characterized as a continuous resource, which can be flexibly allocated depending on the behavioral context (Brady, Konkle & Alvarez, 2011; Ma, Hussain & Bays, 2014). I am not going to try to adjudicate this dispute here. What ultimately matters for the present purposes—and what almost all theorists agree on—is that working memory is indeed severely limited in its capacity.

Finally, we should add one more functional property to our list: *resistance to distraction*. A key difference between working memory and sensory memory stores concerns their behavior in response to new sensory input. Sensory memory stores, like iconic memory, tend to be erased when new visual/auditory information is presented (Phillips, 1974; Cowan, 2008b; Carruthers, 2015). Working memory representations, however, appear to be more resilient, and can remain intact despite the appearance of new irrelevant stimuli (Sakai, Rowe & Passingham, 2002; Sligte, Scholte & Lamme, 2008). Following the standard terminology in the literature, I will refer to this property of working memory as "distractor resistance" (Sakai et al., 2002). By building distractor resistance into our functional characterization of working memory, we can guarantee that there will be a clear functional demarcation between working memory, and other less-durable forms of short-term memory (like sensory memory). To sum up, then, I have suggested that working memory exhibits the following four essential characteristics:

1. Maintenance of information
2. Manipulation (or *manipulability*) of information
3. Capacity limitations
4. Resistance to distraction

Putting these four characteristics, together we thus arrive at the following functional definition: *working memory is the cognitive process responsible for the maintenance and*

potential manipulation of a highly circumscribed amount of information (no longer perceptually available), in a format which is resistant to distraction. This account of working memory is by no means entirely novel; many researchers appeal to various subsets of these properties when providing their own definitions of working memory (Baddeley, 1992; Miyake & Shah, 1999; Luck & Vogel, 2013; Cowan, 2017). However, it is important to bring all four of these characteristics together and to explicitly label them as necessary conditions for working memory.

It is worth noting that the above definition does not specify a maximum duration for working memory. This absence is intentional. Although working memory representations do decay over time (often on the order of seconds), it is possible to *refresh* them by means of a rehearsal process (Atkinson & Shiffrin, 1968; Carruthers, 2015). For this reason, there may be no absolute time limit for working memory. We may hypothetically be able to keep information in working memory for minutes or even hours—though this rarely happens during the course of our everyday lives. Additionally, in contrast to James (1890/1983), my functional definition does not say anything about the relation between working memory and consciousness. Thus, on my definition, it is at least *possible* that working memory could operate without conscious awareness. The question of whether there is any such thing as unconscious working memory will be taken up in Chapter 3.

At this point, I want to briefly consider a possible objection to my view. One potential worry regarding any functionalist account of a psychological phenomenon is that it may end up being too broad, including more than it should. A telling example is provided by Wayne Wu (2011, p. 97): Wu points out that if the phenomenon of attention is merely characterized as a *selection process*, then even a machine that sorts gum balls according to size would qualify as an instance of attention. The philosopher Javier Gomez-Lavin (2017) has recently suggested that functionalist accounts of working memory—like my own—may fall prey to the same kind of overgeneralization. He claims that even a *retinal afterimage* will satisfy the criteria of maintenance and manipulation, since "[a]ctivation can persist in the retina after the withdrawal of the stimulus . . . and cells in the retina can inhibit their neighbors, thus enhancing edge contrast and producing the undulating Mach band illusion" (Gomez-Lavin, 2017, p. 3). This is an important concern to address. After all, we obviously do not want our account of working memory to be so general as to include things like afterimages!

Luckily, however, I believe we can provide a principled reason for excluding afterimages on the basis of the above functionalist account. Recall that, on my functionalist account, manipulation is supposed to be a *personal-level phenomenon*, carried out by the subject. While afterimages may involve some kind of low-level "manipulation", insofar as the retinal cells can bias each other's activity, this kind of manipulation clearly is not under the voluntary control of the subject. So, strictly

speaking, afterimages are not manipulable in the way that working memory representations are. At this point, one might wonder how we are to draw the line between personal-level manipulation and the sub-personal biasing of information-processing. This is a major question in its own right, which I cannot fully address here. Nevertheless, personal-level manipulation likely requires—*at least minimally*—top-down control via executive regions in the frontal cortex (Funahashi, 2001; Buehler, 2018). Since retinal cells do not receive projections from frontal executive regions, retinal afterimages will not be directly available for voluntary manipulation.[7] In sum, then, afterimages do not pose a problem for my functionalist account. On my view, afterimages do not count as genuine instances of working memory because they are inaccessible to the kind of top-down executive control processes that are required for voluntary, personal-level manipulation.

1.3.3 Eliminating working memory?

I end this section by considering a different kind of worry regarding the functionalist account. One might be inclined to think that if what we call "working memory" is undergirded by multiple neural mechanisms, then we should replace the concept of working memory with new concepts that *do* correspond to specific neural

[7] Of course, information processed in the retina may eventually *propagate* to regions of the visual cortex where it can be maintained and manipulated by top-down signals. But at this point the label of working memory may indeed be warranted.

mechanisms (Craver, 2004; see also Poldrak & Yarkoni, 2016 for discussion). This line of reasoning is based on a mechanistic view of psychological kinds. If a concept does not line up with a specific neural mechanism, so the argument goes, then it should be eliminated—or at very least broken down into sub-concepts.

I reject this eliminativist approach for two reasons. First, it is not obvious that psychological concepts must correspond to specific neural mechanisms in order to be scientifically respectable (De Brigard, 2012). Indeed, many foundational concepts in psychology—such as *attention*, *learning*, and *executive function*—are likely disunified at the neural level, yet these concepts nevertheless continue to be used and to guide meaningful scientific research (De Brigard, 2012; Poldrak & Yarkoni, 2016; Chun, Golomb & Turke-Browne, 2011). The situation may be similar when it comes to working memory: perhaps the concept of working memory will continue to be a useful high-level classification despite the fact that there is no single working memory mechanism in the brain. A related point can be made by reflecting on the philosophical literature. As mentioned above, one of the dominant philosophical theories of mind is the doctrine of *functionalism*. According to functionalism, psychological states and processes are identified with functional roles, which can be *multiply realized* by different physical substrates (Levin, 2018; Aizawa & Gillett, 2009). If some version of functionalism is correct—as many contemporary philosophers of mind believe—then the neural disunity

of working memory is no threat to its existence. Working memory can still be considered a genuine functional kind, even if it has multiple distinct neural realizers.

Second, eliminativism is only a viable option insofar as it offers an *alternative* to our current cognitive ontology (Poldrak & Yarkoni, 2016). That is, we cannot plausibly eliminate the concept of working memory unless we have a new conceptual schema to replace it. The problem is that, at present, no such alternative is available. As Poldrak and Yarkoni (2016, p. 600) put it, "there is no guarantee that there is any viable replacement for the concept of working memory that would both (a) map cleanly onto underlying biological structures and (b) remain sufficiently compact and psychologically interpretable to be useful in practice." This is not to say, of course, that working memory will *never* be superseded. It is certainly possible that—in the future—we may revise ontology and do away with the notion of working memory altogether. Until an alternative framework is proposed, however, working memory will inevitably remain an important part of our scientific conception of the mind. As this last point highlights, my project in this chapter is to some extent a descriptive one. My aim has been to develop a precise analysis of the concept of working memory *as it is currently used by cognitive scientists*—not to make speculations about its future prospects.

1.4 Three questions

In this chapter, I did two things. First, I provided an overview of the empirical literature on working memory; second, I developed a functionalist account of working

memory that aimed to capture the core features of working memory as understood by contemporary cognitive scientists. On the account I proposed, working memory has four distinct functional characteristics: (i) information maintenance, (ii) information manipulation (or manipul*ability*), (iii) capacity limitations, and (iv) distractor resistance. I also defended my functionalist account against the charge that it was too liberal (Gomez-Lavin, 2017), and explained why my functionalist account is (at least for the time being) preferable to eliminativism about working memory. With this work behind us, I end the chapter by introducing three remaining questions about working memory. These questions will serve as the basis for the following three chapters of my dissertation.

1.4.1 How is working memory related to attention?

As we saw earlier in section 1.2, recent models of working memory tend to emphasize the role of attention in working memory. Indeed, the predominant view in the literature is that attention is instrumental in maintaining working memory representations over time (Awh & Jonides, 2001; Lepsien & Nobre, 2006a; Postle, 2006; Chun, 2011; Lepsien, Thornton & Nobre, 2011; Gazzaley & Norbre, 2012; Kiyonaga & Egner, 2013; Carruthers, 2015; D'Esposito & Postle, 2015). Even if we grant that attention plays a role in working memory maintenance, however, an important question still remains: namely, how *tight* is the connection between working memory and attention? Does attention need to be *continually* directed at working memory representations in

order to ensure their prolonged maintenance, or can attentional modulation be more sporadic, merely "refreshing" working memory representations at regular intervals?

Opinions are divided on this issue. Some researchers hold that working memory maintenance requires sustained attention to the remembered representations over the entire delay period (Chun, 2011; Carruthers, 2015). Others, however, claim that attention can be dynamically shifted among the contents of working memory—thus, allowing for working memory representations to persist temporarily outside the focus of attention (Baars, 1997; Oberauer, 2002; Myers, Stokes & Nobre, 2017). The second chapter of my dissertation addresses this dispute. Drawing on behavioral, neural and phenomenological data, I argue in favor of the second attention-shifting position. On my view, attention can be rotated among the contents of working memory, with some representations being temporarily retained in an unattended state. This conclusion is important for two reasons. First, it serves as a necessary corrective to those who claim that working memory is just "internal attention" (Chun, 2011; Kiyonaga & Egner, 2013; Carruthers, 2015). Second, it can help inform broader philosophical conceptions of reasoning and thought. If I am correct, when we perform cognitive tasks—like solving a math problem or evaluating an argument—we do not necessarily attend to all the relevant information at once. Rather, our attention may often move among items based on their current relevance.

1.4.2 How is working memory related to consciousness?

A similar (but distinct) issue concerns how working memory relates consciousness. There is a long history of viewing working memory and consciousness as overlapping processes (see Baars & Franklin, 2003); however, the exact nature of the connection between working memory and consciousness remains a contentious topic. There are at least three possible views on this score, which have been endorsed by various authors:

1. The representations held in working memory are *always* conscious (Bor & Seth, 2012; Carruthers, 2015; Stein, Kaiser & Hesselmann, 2016).
2. The representations in working memory can be either conscious or unconscious, but they are all *accessible* to consciousness (Baars 1997, 2001; Kintsch, Healy Hegarty, Pennington & Salthouse, 1999; Gilchrist & Cowan, 2010).
3. The representations in working memory can be either conscious or unconscious, and some are *inaccessible* to consciousness (Soto, Mäntylä & Silvanto, 2011; Soto & Silvanto, 2014).

The third chapter of my dissertation is devoted to adjudicating between these competing positions. I argue that we should reject the first position on the above list, as there is compelling evidence that working memory representations are often maintained in the absence of *immediate* conscious awareness. The situation gets trickier, however, when trying to decide between positions two and three. There are a handful of studies purporting demonstrate the existence of subliminal (i.e., consciously *in*accessible) working memory (Soto et al., 2011; Soto & Silvanto, 2014; Trübutschek et al., 2017), yet

the exact import of these studies is not yet clear. It remains to be seen whether such subliminal processes meet *all* the necessary criteria for working memory, including manipulability and capacity limitations (Persuh et al., 2018). As such, I suggest that we should provisionally adopt position two: working memory representations can be either conscious or unconscious, but they are (to the best of current our knowledge) *consciously accessible*.

1.4.3 Is working memory "sensory based"?

A final question concerns the nature of the *contents* of working memory. In several recent publications, the philosopher Peter Carruthers (2014, 2015; 2017a) has argued for a *sensory-based* account of working memory. On Carruthers' view, the representations held in working memory always take the form of sensory images, depending on activity in "mid-level" sensory areas of the cortex (Carruthers, 2015, p. 14). Although Carruthers still acknowledges the *existence* of non-sensory representations (like concepts and propositional attitudes), he insists that such representations cannot gain direct access to the working memory system. Carruthers' main argument for his sensory-based account rests on claims about the neural architecture of working memory and attention. According to Carruthers, attention is required for information to be boosted above the threshold for entry into working memory (Carruthers, 2015, p. 88-90). Moreover, Carruthers also holds that attention only ever targets "mid-level sensory processing areas" (Carruthers, 2014, p. 147). Putting these two claims together,

Carruthers arrives at the conclusion that only sensory-based representations can ever be held in working memory.

Carruthers' sensory-based account is important to consider, as it has major implications concerning the nature of cognition. If Carruthers is right, non-sensory representations never enter directly into working memory—or, as a result, conscious reflection (Carruthers, 2015, p. 229). My fourth chapter thus critically examines Carruthers' sensory-based account. In this chapter, I ultimately challenge Carruthers view. First, I show that Carruthers' main argument for his sensory-based account is flawed, resting on the empirically dubious premise that attention only targets mid-level sensory areas. Second, I contend that there is in fact some positive evidence for the existence of non-sensory working memory representations. This evidence comes from several sources, including electrophysiological studies with monkeys (Nieder, 2012; Vergara, Rivera, Rossi-Pool & Romo, 2016), human neuroimaging studies (Lee, Kravitz & baker, 2013), and studies of individuals with aphantasia (Jacobs, Schwarzkopf & Silvanto, 2018).

2. Attention to working memory representations: Sustained or sporadic?

In recent years, there has been a growing interest in the relationship between working memory and attention. Indeed, there is now a large body of evidence indicating that attention is, to some extent, involved in the process of working memory maintenance (Awh & Jonides, 2001; Awh, Vogel & Oh, 2006; Lepsien & Nobre, 2006a; Lepsien, Thornton & Nobre, 2011; Gazzaley & Norbre, 2012; Kiyonaga & Egner, 2013; Carruthers, 2015; but see Fougnie, 2008 for a dissenting view). There is, however, still an ongoing debate regarding the degree of overlap between working memory and attention. Some researchers claim that working memory maintenance requires *sustained* attention, such that the representations in working memory are continuously attended in parallel (Chun, 2011; Carruthers, 2015). By contrast, other researchers claim that attention can be dynamically shifted among the contents of working memory, with some representations being temporarily retained in an unattended state (Baars, 1997; Oberauer, 2002; Myers, Stokes & Nobre, 2017). My aim in the present chapter is to carefully examine the connection between working memory and attention. I ultimately argue that a version of the second view is correct: while attention does play an important role in working memory maintenance, working memory representations can

persist—*at least temporarily*—outside the focus of attention.[8] Thus, on my view, attentional modulation is needed to maintain representations in working memory, but such attentional modulation can be sporadic rather than continuous.

This chapter proceeds as follows. In section 2.1, I provide a brief overview of the concepts of working memory and attention. Next, in section 2.2, I survey some of the empirical literature showing that attention plays an important role in working memory maintenance. In section 2.3, I address the central question of whether working memory representations are *always* attended. Drawing on empirical and phenomenological data, I argue for a negative answer: contrary to the sustained-attention view, working memory representations can be dynamically shifted in and out of the focus of attention depending on their current behavioral relevance. Section 2.4 examines a separate (but related) issue—namely, the *capacity* of attention. Drawing on earlier proposals (Eriksen & St. James, 1986; Cowan et al., 2005), I suggest that attention functions as a flexible resource, sometimes "zooming in" to focus on a single working memory item, and sometimes "panning out" to encompass multiple items simultaneously. Finally, in

[8] Here I use the term "focus of attention" simply to refer to that which is currently being attended to. On this definition, items outside the focus of attention count as being fully *un*attended. This qualification is important, as some authors might prefer a narrower definition, according to which the term "focus of attention" refers only to center (as opposed to the periphery) of one's attentional field (see Watzl, 2011).

section 2.5, I respond to a conceptual objection to the view I endorse, raised by both Peter Carruthers (2015) and Ned Block (2007).

2.1 Background

Before we assess how working memory and attention are related, a few words should be said about the individual natures of working memory and attention. The present section thus provides a brief overview of the two concepts.

2.1.1 Working memory

The term "working memory" gets used in a variety of different ways (Cowan, 2017). At its core, however, working memory is usually understood as *the cognitive process responsible for the maintenance and potential manipulation of a highly circumscribed amount of information, which is no longer perceptually available* (Baddeley, 1992; D'Esposito, 2007; Marois, 2015). Roughly speaking, working memory is what allows us to hold information "in mind"—like a telephone number or visual image—over short periods of time. Aside from being responsible for maintenance and manipulation, working memory has at least two other distinguishing functional characteristics. First, working memory is highly limited in its storage capacity. According to one popular estimate, for instance, working memory is only capable of storing around four or five discrete representations simultaneously (Cowan, 2010). Next, a second key feature of working memory is *resistance to distraction*. That is, information stored in working memory can be

retained despite the presentation of new sensory input (Sakai, Rowe & Passingham, 2002).

The neural correlates of working memory have been the subject of much empirical research over the past 50 years. Early observations of elevated prefrontal activity during working memory tasks initially prompted researchers to hypothesize that the prefrontal cortex was the locus of working memory storage (Goldman-Rakic, 1992, 1995). In the past decade or so, however, the "prefrontal buffer" view of working memory has been challenged by a variety of novel findings (see Postle 2006, 2017). One particularly compelling source of evidence comes from recent multivariate neuroimaging studies: such studies have shown that, during the performance of working memory tasks, memoranda-specific activity can be detected outside the prefrontal cortex, in early sensory areas (see Postle, 2015 for a review). For example, Harrison and Tong (2009) found that information about the orientation of remembered line gratings was coded in patterns of activity in the early visual cortex (V1-V4). Similarly, Riggall and Postle (2012) found that information about the direction of motion of a remembered stimulus was coded in visual area MT—but *not* in more anterior frontal regions. Importantly, these are not isolated findings. The existence of working-memory-related activity in sensory areas has now been confirmed by numerous studies, especially within the visual domain (Serences, Ester, Vogel & Awh, 2009; Albers, Kok, Toni, Dijkerman & Lange, 2013).

The discovery that sensory areas encode fine-grained representations of remembered stimuli has led to a reevaluation of the neural correlates of working memory. According to more contemporary neural models of working memory, there is no specialized site of working storage. Instead, working representations are thought to be held in the sensory-processing areas where they are in initially encoded (Postle 2006, 2007; Carruthers, 2015). The prefrontal cortex is still considered to be involved in working memory; however, its primary function is thought to be *executive* in nature, rather than storage-related.[9] Specifically, the prefrontal cortex is thought to produce activity in other storage-related areas by means of top-down signaling (Gazzaley & Nobre, 2012; Carruthers, 2015; Postle, 2017). Importantly, as we will see in section 3, these frontal top-down signals are often conceived of as *attentional* signals.

2.1.2 Attention

Like working memory, the exact nature of attention remains debated within the scientific and philosophical literature (Mole, Smithies & Wu, 2011; De Brigard, 2012). For the purposes of the present chapter, I am going to adopt the influential functional characterization of attention provided by Chun, Golomb and Turk-Browne (2011; see also De Brigard, 2012). According to Chun and colleagues, all instances of attention

[9] Areas of the frontal cortex may still be responsible for storing high-level abstract representations, however (Lee, Kravitz & Baker, 2013). See chapter 4.

share at least three basic functional characteristics: (i) *selectivity*, (ii) *filtering* and (iii) *modulation*. First, attention is *selective*, in the sense that it prioritizes some pieces of information over others. The brain is constantly being bombarded by a large number of competing inputs, and a central function of attention is to select which pieces of information will receive further processing. The flip-side of selection is *filtering*: in addition to selecting some information contents for further processing, attention also *filters out* other contents, blocking them from consuming valuable cognitive resources. Finally, the third key characteristic of attention is *modulation*: once a piece of information is selected, attention is thought to modulate the processing of that information in a faciliatory manner. Modulation typically involves enhancing the activity of the selected neural representations (Ruff, 2011; Carruthers, 2015, p. 61). It is worth noting that the modulatory effects of attention are well-documented in the empirical literature. Numerous studies have shown that attention to specific objects, spatial locations or sounds can facilitate neural activity in the brain regions that process the relevant sensory information (Woldorff et al., 1993; Reynolds & Chelazzi, 2004; Ruff, 2011).

Additionally, Chun and colleagues (2011) also draw a distinction between two broad categories of attention: *external attention* and *internal attention*. External attention encompasses the selection and modulation of externally produced information coming in through the senses. In other words, external attention refers to outwardly-directed perceptual attention, including attention to perceived objects, features, sounds, etc.

Internal attention, by contrast, encompasses the selection and modulation of information that is generated *internally*—i.e., information which is not perceptually available. Attending to one's memory of having seen the *Mona Lisa* five years ago would constitute an instance of internal attention, for example, since the relevant information content is being generated via an internal recollection process, rather than being directly perceived. The notion of internal attention is particularly relevant in the present context, as the selection and modulation of working memory representations will presumably be the result of such internal attention.

As with working memory, the neural correlates of attention have been studied extensively. Although questions still remain, a large amount of empirical research indicates that frontoparietal cortices are instrumental in the control of attention (Corbetta & Shulman, 2002; Ruff, 2011; Ptak, 2012). Many neuroimaging studies have found heightened activity in frontal and parietal areas during attentionally demanding tasks (Corbetta & Shulman, 2002; Giesbrecht, Woldorff, Song & Mangun, 2003; Naghavi & Nyberg, 2005). Additionally, lesions to frontal and parietal cortices have been found to produce certain types of attentional deficits, such as hemineglect (Driver & Mattingley, 1998; Rossi, Bichot, Desimone & Ungerleider, 2007). Some studies have even used a combination of intervention and neuroimaging techniques to show that artificial stimulation of frontal and parietal areas can produce attentional modulations in sensory areas. For instance, Moore and Armstrong (2003) found that microstimulation of the

frontal eye fields in monkeys selectively enhances the responsivity of neurons in visual area V4. Similarly, Ruff and colleagues (2006, 2007) found that applying transcranial magnetic stimulation (TMS) to various frontal and parietal sites in humans can also modulate activity in retinotopic areas of the visual cortex (V1-V4).

Researchers also commonly distinguish between two separate attentional networks: the top-down attentional network and the bottom-up network (Corbetta & Shulman, 2002; Vossel, Geng & Fink, 2014; Carruthers, 2015). The top-down network is located in *dorsal* frontoparietal areas and is responsible for goal-directed attentional processing. The bottom-up network, in contrast, is located in *ventral* frontoparietal areas and is responsible for automatic, stimulus-driven attentional capture (Corbetta & Shulman, 2002). Importantly, the top-down/bottom-up distinction is orthogonal to Chun and colleagues' earlier distinction between external and internal attention (Chun et al., 2011; De Brigard, 2012). External attention can be *either* top-down or bottom-up. In some instances, we deliberately direct our attention towards an external target, while in other instances the external target automatically grabs out attention in a bottom-up fashion. Internal attention, by contrast, is likely to be top-down in nature, since internal attention—by definition—involves a lack of relevant sensory input.

2.2 Evidence that working memory maintenance involves attention

Let us now move on to the main topic of discussion: the relationship between working memory and attention. Although working memory and attention have historically been treated as separate phenomena, it is becoming increasingly clear that they are closely related. Indeed, there is a growing body of evidence indicating that attention plays an important role in working memory maintenance (Awh & Jonides, 2001; Lepsien & Nobre, 2006a; Lepsien et al., 2011; Gazzaley & Norbre, 2012; Kiyonaga & Egner, 2013; Carruthers, 2015). In the present section, I highlight some of the key empirical findings—both behavioral and neural—supporting the link between working memory maintenance and attention.

An early demonstration of attentional involvement in working memory maintenance comes from a study by Awh, Jonides and Reuter-Lorenz (1998; for helpful summary see Awh & Jonides, 2001). In this study, participants were given a spatial working memory task, interspersed with a secondary color-judgment task. Participants were briefly shown a circle (the memory cue), and tasked with remembering its location. During the 5-second memory delay, a second colored circle appeared on the screen—at which point participants had to immediately report its color (red or blue). On some trials, the colored circle appeared at a different location from the memory cue, while, on other trials, the colored circle overlapped the original location of memory cue. In the

former case, participants had to move their focus of attention *away* from the cued location to perform the color-judgement task, whereas in the latter case participants could retain their initial focus of attention (since the colored circle overlapped the cued location). Finally, after the delay, a third circle appeared (the memory probe), and participants had to indicate whether its location was the same as the location of the original memory cue. Interestingly, Awh and colleagues found that distracting attention had a negative effect on memory performance: participants' accuracy on the spatial working memory task was *significantly lower* when they had to shift their focus of attention, compared to their performance in the non-shifting condition. As Awh and colleagues point out, this result supports a link between attention and working memory maintenance. If shifting attention *away* from the cued location decreases memory performance, then attention presumably "plays a beneficial functional role in the active maintenance of location information" (Awh & Jonides, 2001 p. 121).

Additional evidence that working memory maintenance involves attention comes from an experiment by Todd, Fougnie and Marois (2005), which examined how working memory load affects participants' susceptibility to inattentional blindness.[10] In the experiment, participants were briefly presented with an array of colored discs—one

[10] Inattentional blindness is a phenomenon whereby participants fail to notice a salient stimulus when their attention is occupied with another task.

disc in the low-load condition and four discs in the high-load condition—and instructed to remember their location and color. Next, after a delay of five seconds, a probe disc appeared and participants had to determine whether it "matched the location and color of one of the discs in the sample display" (Todd et al., 2005, p 996). Crucially, on some trials, an unexpected stimulus (a white clover symbol) was briefly flashed during the delay period. At the end of these critical trials, participants were also asked whether or not they had noticed the presentation of this unexpected stimulus.[11] The interesting result was that participants in the high-load condition were less likely to detect the stimulus—that is, they showed higher rates of *inattentional blindness*—relative to participants in the low-load condition. This finding supports the view that working memory maintenance and attention are related. The fact that participants in the high-load condition are more susceptible to inattentional blindness suggests that working memory "uses up" attentional resources, thereby limiting the amount of attention that participants can direct towards the external environment.

Moving on, a wealth of neural data also supports a link between working memory maintenance and attention (see Carruthers, 2015, Ch. 4). Numerous

---

[11] During each trial, participants were also given a secondary verbal memory task, which involved rehearsing two auditorily-presented digits. The aim of this manipulation was to ensure that the primary visual task was in fact carried out by visual working memory, rather than being partially offloaded to verbal working memory.

neuroimaging studies have found that there is increased activation in the frontoparietal attentional network—particularly in dorsal (top-down) attentional areas—during the performance of working memory tasks (Todd & Marois, 2004; Naghavi & Nyberg, 2005; Majerus, Péters, Bouffier, Cowan & Phillips, 2018). Furthermore, disruptions to the functioning of the frontoparietal attentional network (due to lesions) have been shown to produce deficits in working memory performance (D'Esposito & Postle, 1999; Curtis, 2006; Pasternak, Lui & Spinelli, 2015). Interestingly, frontal lesion patients appear to have particular difficulty with working memory tasks that involve distractor stimuli (D'Esposito & Postle, 1999). This result makes sense, given the role of attention in filtering out irrelevant information; if an individual's attentional control mechanisms are compromised, then it will presumably be harder for them to ignore irrelevant distractors.

One functional magnetic resonance imaging (fMRI) study by Lepsien and Nobre (2006b) provides an especially striking demonstration of attentional involvement in working memory. In this study, participants were tasked with remembering both a scene and a face over a delay of several seconds. During the delay, however, participants were presented with retrocues[12] that instructed them to orient their attention towards

[12] A retrocue is a cue that is presented *after* the relevant stimulus array, directing participants attention to one of the previously perceived (and currently remembered) items.

one item—either the face or the scene—for the purposes of an upcoming memory test. Fascinatingly, Lepsien and Nobre found that orienting attention to faces or scenes modulated activity in content-related brain areas. Specifically, attending to remembered faces enhanced activity in the fusiform gyrus (an area dedicated to processing facial information), whereas attending to remembered scenes enhanced activity in the parahippocampal gyrus (an area dedicated to processing scene information). This study thus provides powerful evidence that attention operates on the contents of working memory: it demonstrates that attention can modulate working memory representations "on the fly", enhancing the activity of representations that are relevant for the task at hand. Similar findings have also been documented by Kuo, Stokes and Nobre (2012), using electroencephalography (EEG) rather than fMRI.

On the basis of the results summarized above, many researchers have begun to view working memory and attention as interrelated processes. Indeed, the emerging consensus in cognitive neuroscience is that attentional mechanisms play a key role in the maintenance of working memory representations. According to this view, we hold information in working memory by "allocating attention to internal representations" (D'Esposito & Postle, 2015, p. 115; see also Curtis & D'Esposito, 2003; Postle, 2006; Chun 2011; Gazzaley & Nobre, 2012; Carruthers, 2015; Postle, 2017). The presumed role of such attentional modulation is to prolong the activation of working memory representations in the absence of sensory input. The resulting model of working

memory is often referred to as the *sensory recruitment model* to emphasize the fact that working memory representations are typically offloaded to primary sensory processing areas (D'Esposito & Postle, 2015). Attention plays an equally central role in the model, however, since it is *attention* that is ultimately responsible for keeping working memory representations in a continued state of heightened accessibility.

2.3 Working memory representations outside the focus of attention

In the previous section, we saw that there is strong evidence that working memory maintenance involves the deployment of attention. Indeed, according to the most popular contemporary model of working memory—the sensory recruitment model—attention plays a central role in sustaining working memory representations. This does not completely settle the relationship between the two phenomena, however. There is still a further question regarding the *tightness* of the connection between attention and working memory maintenance. In particular, it remains unclear whether working memory representations must be *continuously attended* during the maintenance process. Some theorists, such as Chun (2011) and Carruthers, (2015), have suggested that the contents of working memory are indeed *always* attended. On this view, maintaining a set of representations in working memory involves consistently attending to them over the entire delay period. There is another possibility, however: attention may only be

needed to *periodically refresh* the contents of working memory.[13] On this view, attention may shift amongst the representations in working memory, temporarily leaving some representations outside the focus of attention. A number of theorists have endorsed versions of this second view, including Baars (1997), Oberauer (2002) and Myers et al. (2017; see also LaRocque, Lewis-Peacock & Postle, 2014).

Is there any support for the view that working memory representations can temporarily persist outside the focus of attention? I believe that the answer is "yes". In the present section, I argue in favor of the existence of unattended working memory representations, drawing on an array of phenomenological, behavioral and neural evidence. Ultimately, I conclude that Baars (1997), Oberauer (2002) and Myers et al. (2017) are correct: at any given moment, the focus of attention may only contain a fraction of the total contents of working memory. As we will see, however, there is some debate about how unattended working memory representations are stored in the brain.

2.3.1 Phenomenological and behavioral considerations

One initial motivation for thinking that working memory representations are not always attended comes from our inner phenomenological experience of working memory maintenance. To borrow an example from Bernhard Baars (1997, p. 44), try holding in mind the following set of numbers—42, 71, 89, 36—for ten seconds. If you are

[13] See Camos et al. (2018) for a helpful review of the literature on attentional refreshing.

like me, you presumably accomplished this task by mentally rehearsing the numbers in inner speech. In this case, however, it doesn't seem like one is attending to *all the numbers at once*. Rather, as Baars (1997, p. 44) notes, one typically has the inner experience of attending to each number in quick succession, attending first to the number 42, then to 71, and so on. A similar point applies—though perhaps less obviously—in the context of visual working memory. Briefly examine the three shapes below and then try to hold them in memory for a few moments (Figure 4). One way to hold these shapes in mind is to "chunk" them together, forming a single mental image of the visual display. However, another option is also available: you could visualize each shape *sequentially*, repeatedly rotating them in and out of the "mind's eye". When this second strategy is employed, it seems to us as though attention is being *shifted* between the relevant items, one after another. Of course, I do not expect an appeal to phenomenology to settle the debate by itself. Nevertheless, such phenomenological observations should at least persuade us to take seriously the idea that attention and working memory can come apart.

Figure 4: Three random shapes.

An experiment by Garavan (1998) also provides some behavioral evidence that meshes with the above phenomenological observations (see also Fortney, 2018). In this experiment, participants were shown a random sequence of triangles and rectangles. Their task was to keep two mental counts in working memory—one for the total number of each type of shape. Importantly, the rate of display was "self-paced", with participants pressing a button to indicate that they were ready to advance from one shape to the next (Garavan, 1998, p. 265). This allowed Garavan to measure the time it took for participants to update their respective mental counts. The interesting result of the experiment was that it took participants more time to update their mental counts following "stimulus switches" (i.e., instances where a triangle was presented after a rectangle, or vice versa) relative to "stimulus no-switches" (i.e., instances where the same type of shape was presented twice in a row). Specifically, participants' response times were 100-500 milliseconds slower after stimulus switches, when compared to stimulus no-switches. Garavan proposed that this response time cost was likely due to *shifts in attention* (Garavan, 1998, p. 2177): participants' response times were slower on switch trials because they had to *move their attention* from the previously-relevant count to the newly-relevant count (see also Fortney, 2018, p. 126). Conversely, no such response time costs were incurred on no-switch trials because participants did not have to switch their focus of attention from one count to the other. Garavan's response time

data thus lends empirical support to the intuition that we sometimes shift attention among the contents of working memory.

2.3.2 Neural evidence for unattended working memory representations

Moving on, neural evidence for unattended working memory representations comes from an fMRI study by Lewis-Peacock, Drysdale, Oberauer and Postle (2012; see also LaRocque, et al., 2014; Gomez-Lavin, *in preparation*). Lewis-Peacock and colleagues attempted to dissociate working memory and attention with a clever experimental design that involved multiple retrocues. Participants were initially shown two stimuli, consisting of two of the following: (1) a word; (2) a pronounceable pseudoword; or (3) a pair of oriented lines (Lewis-Peacock et al., 2012, p. 70). The presentation of the stimuli was followed by a retrocue (a set of arrows) indicating which stimulus was immediately relevant for an upcoming test. After a delay, participants were then shown a memory probe and asked to report whether it matched the cued stimulus.[14] This was not the end of the trial, however. After the first memory test, *yet another retrocue was presented*, preparing the participants for a second memory test. Crucially, the second retrocue could either stay on the same item, or it could "switch" to the originally uncued item, thus requiring participants to shift attention to the item that had previously been

[14] What constituted a "match" differed depending on the nature of the cued stimulus: for words, a match was any synonym; for pseudowords, a match was any rhyming pseudoword; and for line segments, a match was any similarly-oriented pair of line segments.

ignored (Lewis-Peacock et al., 2012, p. 69). After the second retrocue, another delay period ensued, followed by a second memory test.

During the performance of the task, Lewis-Peacock and colleagues measured participants' neural activity using fMRI. With the aid of sophisticated pattern classification techniques (multivariate pattern analysis), they were able to isolate patterns of activity that were specific to each class of remembered stimulus (words, pseudowords, and lines). This allowed them to examine how the neural representation for a given stimulus changed, depending on whether or not it was currently being prioritized by attention. The results were striking. Before the first retrocue, both items exhibited "decodable" patterns of neural activity (Lewis-Peacock et al., 2012, p. 73). After the onset of the first retrocue, things changed: while the *cued* item was still actively represented, the observable neural signatures of the *uncued* item disappeared entirely. Crucially, however, the uncued item remained accessible, and could be "brought back online" (i.e., reinstantiated in active format) if it was subsequently selected by the *second* retrocue (Lewis-Peacock et al., 2012, p. 73). This finding thus provides neural support for a dissociation between working memory maintenance and attention. It appears that uncued working memory representations can be temporarily shifted outside the focus of attention—indicated by a drop in neural activation—without thereby ceasing to be remeberable. (For similar findings see LaRocque, Lewis-Peacock, Drysdale, Oberauer & Postle, 2013; LaRocque, Riggall, Emrich & Postle, 2016; Rose et al., 2016).

At this point, a skeptic might raise the following objection: how do we know that the unattended representations in the above studies were in fact being stored in *working memory*, rather than ordinary long-term memory? This is an important objection to consider. After all, if unattended representations were simply being offloaded to long-term memory, we would have no grounds for concluding that unattended working memory exists. Luckily, a recent study by Rose et al. (2016) provides evidence against this alternate interpretation. In this study, participants performed a double-retrocue task—similar to the one employed by Lewis-Peacock et al., 2012)—while their neural activity was recorded using EEG. Additionally, during each trial, a TMS pulse was also applied to targeted regions of the brain; the aim of this manipulation was to see whether it was possible to *manually reactivate* (via TMS) latent, unattended memory representations. Fascinatingly, the technique worked. TMS applied to targeted brain areas could reactivate representations of uncued items that were not currently being attended. Crucially, however, unattended representations could only be reactivated by TMS *so long as they were potentially relevant*—that is, so long as they might be needed for an upcoming memory test. After the trial was over, subsequent TMS pulses had no effect. As the authors point out, these results suggest that the unattended memory representations are maintained in a state of heightened accessibility, which is distinct from the default state of long-term memory (Rose et al., 2016, p. 1138). After all, if

unattended representations were merely being stored in ordinary long-term memory, then they should be equally (in)accessible both during the trial and after it finished.

2.3.3 How are unattended working memory representations stored in the brain?

Suppose, then, that we thus accept the existence of unattended working memory representations. At this point, a further question arises: namely, how are such unattended working memory representations stored in the brain? This question is particularly pressing because Lewis-Peacock et al. (2012) failed to find any above-baseline neural activity associated with unattended working memory representations. Such a result appears to contradict the longstanding assumption that working memory representations are encoded in the active firing of neural populations (Luck & Vogel, 2013).

One possibility, suggested by Lewis-Peacock and colleagues (2012), is that unattended working memory representations may actually rely on short-term *synaptic plasticity*, rather than sustained neuronal spiking. On this view, the initial processing of a stimulus can generate changes in synaptic connectivity—perhaps resulting from elevated neuronal calcium concentrations—which temporarily buffer mnemonic information (Mongillo, Barak & Tsodyks, 2008). If attention is shifted away, this "activity-silent" working memory representation will decay over a short period of time; however, if attention is shifted *back* to the representation before it decays, the

representation may be returned to its original state of heightened activity. Although synaptic models of unattended working memory remain speculative, they are garnering increasing support. Indeed, many neuroscientists are now endorsing the idea that synaptic plasticity may play an important role working memory maintenance (Mongillo et al., 2008; LaRocque et al., 2014; D'Esposito & Postle, 2015; Trübutschek et al., 2017; Lundqvist, Herman & Miller, 2018). Crucially, the short-term synaptic mechanisms hypothesized to be involved in working memory (e.g., increased calcium levels) are still thought to be distinct from the mechanisms of ordinary long-term memory, such as long-term potentiation (see LaRocque et al., 2015; Rose et al., 2016).

A synaptic model of unattended working memory is not the only option, however. Alternatively, it could be the case that unattended working memory representations *are* encoded via active neural firing, but this activity is instantiated in a format that is difficult to detect using contemporary neuroimaging techniques—thus explaining previous null results (LaRocque et al., 2014). At least one recent fMRI study by Christophel and colleagues (2018) supports this alternative interpretation. In this study, participants performed a double-retrocue task which involved shifting attention between two Gabor patches (i.e., oriented line-gratings) held in working memory. Using a highly-sensitive variant of MVPA, Christophel and colleagues were able to identify stimulus-specific patterns of activity corresponding to *both* attended and unattended Gabor patches. Interestingly, however, different brain regions were involved in the

storage of attended versus unattended items. Information about the *attended* Gabor patch was encoded in the visual cortex, whereas information about *both the attended and unattended* Gabor patches was encoded in the intraparietal sulcus and frontal eye-fields. Christophel and colleagues thus propose that working memory is an active process, but that the locus of working memory storage varies depending on the attentional status of the representations being held in mind. On their view, "sensory cortex maintains a high-resolution representation of the currently attended memory item, whereas parietal cortex has low-resolution representations of both attended and unattended items" (Christophel et al., 2018, p. 496).

Unfortunately, based on the current state of the literature, it is not yet clear which of these two views is correct. Unattended working memory representations could be encoded via short-term synaptic changes; but they could also be encoded via active spiking patterns that are difficult to detect using current neuroimaging techniques (or even some combination of both). Further research is required to clarify this issue. Nevertheless, setting aside questions of neural implementation, the studies mentioned above all converge on the idea that there is an important distinction to be made between attended and unattended working memory representations. Indeed, even the one study that found above-baseline activity corresponding to uncued items (Christophel et al., 2018) suggests that these items are encoded in a different manner from items in the current focus of attention.

2.4 The capacity of attention

In the previous section, I outlined several studies that support a dissociation between working memory and attention (Garavan, 1998; Lewis-Peacock et al., 2012; Rose et al., 2016; Christophel et al., 2018). Notably, in these studies, participants appeared to focus their attention to *one* memory item—such as a word or a line-grating—leaving other items temporarily outside the focus of attention. This might prompt readers to wonder whether attention always operates serially, selecting one working memory representation at a time, or if it can sometimes focus on multiple representations simultaneously. The capacity of attention is itself a highly-debated topic. Some authors claim that the focus of attention is restricted to a single item (Baars, 1997, 2001; Oberauer, 2002), while others claim that attention can be simultaneously directed towards multiple items, up to a limit of around four (Cowan, 2010, 2011; Carruthers 2015). The present section is devoted to adjudicating this issue. I ultimately suggest that there is some truth to both views: attention may *sometimes* focus on a single item, but, in other contexts, it may be simultaneously divided between multiple items.[15]

---

[15] What, one might ask, constitutes a single "item"? Unfortunately, there is no wholly unambiguous answer to this question. In the visual domain, an item is often taken to be a single spatiotemporally discrete object. However, related visual stimuli—like the letters C A T—may potentially be "chunked" into a single higher-order item (see footnote 3). A full account of the capacity of attention will inevitably require saying more about the identity conditions of individual items. However, this is beyond the scope of the present chapter.

To begin, the studies outlined in section 2.3 provide some initial motivation for "single-item" focus of attention. For instance, Garavan (1998) found that alternating between two mental counts in working memory incurred a reaction time cost, suggesting that participants had to take time to shift their attention from one count to the other (see section 2.3.1). Relatedly, in the retrocue paradigm developed by Lewis-Peacock et al. (2012), it was exclusively the *cued* item that exhibited an observable activation trace; a natural interpretation of this finding is that only the presently-cued item occupies the current focus of attention (Lewis-Peacock et al., 2012, p. 73; see section 2.3.2). This is not the end of the story, however. There is also competing evidence for a "multi-item" focus of attention (Cowan, 2011).[16] Ironically, as noted by LaRocque and colleagues (2014, p. 9), the study by Lewis-Peacock et al. (2012) itself provides some indirect support for a multi-item focus. Recall that, up until the point of the first retrocue, "there [was] evidence … for simultaneous active neural representations of both items in memory" (LaRocque et al., 2014, p. 9). This suggests that *before* the first retrocue appeared, participants were actively attending to both items simultaneously.

Additional support for a multi-item capacity comes from research on divided visual attention. For instance, Awh and Pashler (2000) found that participants are capable of attending to—and encoding the identity of—two noncontiguous target

[16] The terminology I am using here ("single-item focus" vs. "multi-item focus") comes from Cowan (2011).

objects (numbers or shapes) in a briefly-presented stimulus array. Awh and Pashler argue that the participants in their study must have been attending to both objects simultaneously, because the duration of stimulus presentation (~69ms in one experiment) was too short to allow for shifts in attention.[17] Another provocative example of divided attention is provided by McMains and Somers (2004). McMains and Somers had participants divide their attention between two spatially-distinct visual streams (each consisting of a rapidly-presented series of letters and numbers) while recording their neural activity with fMRI. Fascinatingly, McMains and Somers found that attending to the two streams simultaneously boosted activity in separate regions of the visual cortex, which corresponded to the spatial locations of the attended streams. This finding supports the view that attention can be "split" between two different stimulus locations at the same time.

At first glance, the observed pattern of results might seem somewhat confusing. Some findings appear to support a single-item focus of attention, while others appear to support a multi-item focus (or even the existence of separate attentional foci). How can we reconcile these two competing views? The solution, I claim, is to recognize that attention is a "flexible resource" that can be deployed in different ways in different

---

[17] The presentation of the array was also followed by a mask, which should have disrupted any lingering sensory memory. This rules out the possibility that participants attended to one object during stimulus presentation and then shifted their attention to a sensory trace after stimulus offset.

contexts (Alvarez & Franconeri, 2007). We may *sometimes* choose to deploy our attention in a serial fashion, shifting a highly focused beam of attention between individual items. In other contexts, however, we may choose to deploy our attention in a more diffuse manner, spreading it across multiple items simultaneously (Figure 5). This proposal is similar to the idea that attention functions as a "zoom lens" (Eriksen & St. James, 1986; Cowan et al., 2005). On the zoom lens model, attention can either "zoom in" to provide a detailed representation of a single item, or "zoom out" to provide a more coarse-grained representation of multiple items at once. (I hesitate to fully endorse the "zoom lens" model, however, because it carries with it the implication that there is only a *single* spotlight of attention that varies in size. The studies of divided attention suggest that there may occasionally be *two separate* spotlights of attention, which operate independently of one another (McMains and Somers, 2004).)

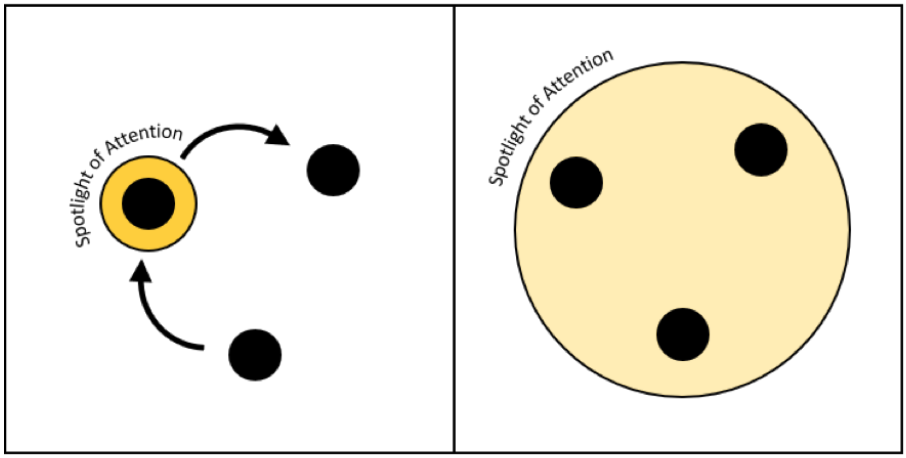

Figure 5: An illustration of two different attentional strategies. Attention may sometimes be rapidly shifted from one item to the next (left). In other contexts, however, attention may be spread across multiple items simultaneously (right).

The view that attention can be flexibly deployed has several advantages. First, it provides a neat account of the experimental data, allowing us to explain both cases where attention appears to be focused on a single item, and cases where attention appears to be focused on multiple items simultaneously. Second it also meshes with our inner phenomenological experience. Intuitively, we do seem to use different attentional strategies in different contexts: sometimes we attend to items in quick succession (such as when we are rehearsing numbers or words in inner speech), and sometimes we try to focus our attention on several items at once (such as when we try to visualize multiple items in a single visual scene). Additionally, there is also some direct empirical evidence that attention can vary in scope depending on the context. Müller and colleagues (2003) had participants attend to sets of 1, 2 or 4 spatially-distinct visual items while

undergoing fMRI. The authors found that, as the number of attended items increased, larger portions of the visual cortex were modulated by attention. At the same time, however, the modulatory effects of attention were *weaker* when spread across multiple items. Thus, it appears that participants can allocate attention on an as-needed basis, varying its scope depending on the task demands (Müller et al., 2003, p. 3561).

At this point, an important question arises: if we can adopt different attentional strategies for maintaining representations within working memory, then what factors determine which strategy we will employ in a given context? In other words, *why* do we sometimes attend to working memory representations serially, while other times attending to multiple items simultaneously? Unfortunately, I know of no empirical research that directly addresses this question. Nevertheless, it is possible to make a few tentative suggestions. First, the *modality* of the remembered information might be an important factor. On the one hand, auditory information seems to lend itself to serial-attention strategy, as we know that people have trouble simultaneously attending to two auditory streams at once (Moray, 1959). Indeed, the strategy of *rehearsal*—which operates in a serial fashion—seems to be particularly entrenched in the auditory domain. On the other hand, visual information seems more amenable to a divided-attention strategy, since it appears to be possible for people to attend to multiple spatially-distinct visual stimuli at once (see above). Next, another potential factor in determining one's attentional strategy might be *the mode of stimulus presentation*. If participants are

presented with stimuli in a *sequential* fashion, this might prompt them to employ a serial strategy, since no two items are ever apprehended at once. By contrast, if participants are presented with multiple items *simultaneously*, this might prompt them to attend all the items in parallel.

Of course, I don't expect this to be the final word on the capacity of attention. Future research will surely tell us much more about the nature and scope of our attentional capacity limits. My aim in this section has simply been to provide a plausible explanation for why we sometimes observe differences in attentional capacity across different experimental contexts.

2.5 Conceptual issues

I want to end by considering a possible challenge to my view regarding the relationship between working memory and attention. One of the central claims of this chapter is that, although working memory and attention are closely related, working memory representations can be temporarily stored *outside* the focus of attention. Some authors, however, reject the possibility of unattended working memory representations on conceptual grounds. For instance, Peter Carruthers (2015, p. 91) takes top-down attention to be a defining feature of working memory maintenance, and thus he claims that only attended representations should be considered to be a part of working memory. Carruthers admits that representations may persist *in some capacity* outside the focus of attention, but he insists that such representations are distinct from genuine,

"attention-dependent" working memory (Carruthers, 2015, p. 89). A similar view is also articulated by Ned Block (2007, p. 539). Block reserves the term "working memory" for representations that are currently being attended and globally broadcast throughout the brain, preferring to use the more generic term "short-term memory" to describe short-lived mnemonic representations outside the focus of attention.

Disagreements about what counts as working memory are hard to settle because they ultimately boil down to differences in conceptual preferences. If Carruthers and Block wish to use the term "working memory" to refer only to attended representations, they are certainly within their rights to do so. It would be impossible to refute their choice of terminology. That said, I think there are some reasons for preferring a more inclusive conception of working memory (e.g., one which allows for unattended working memory representations) over Carruthers' and Block's restrictive conception. First of all, it should be noted that most empirical researchers do not view sustained attention as a necessary condition for working memory maintenance. Indeed, as Javier Gomez Lavin (2017) points out, many leading theorists—including Alan Baddeley, the father of modern working memory research—allow that working memory may encompass both attended and unattended contents (Baars, 1997; Cowan, 1999; Oberauer, 2002; Awh et al., 2006; Fougnie, 2008; Baddeley, 2010; D'Esposito & Postle, 2015; Hitch, Hu, Allen & Baddeley, 2018; Myers et al., 2017). Thus, from a sociological standpoint, an

inclusive conception of working memory appears to be more prominent than a restrictive one.[18]

Second, it also worth noting that many "canonical" examples of working memory likely rely on a mixture of attended and unattended representations. Consider, for instance, the act of mentally rehearsing a phone number. As we discussed earlier (in section 2.3.1), this kind of rehearsal process appears to involve the sequential shifting of attention from one number to the next. If verbal rehearsal is taken to be a function of working memory—which is almost universally agreed upon—then working memory will presumably end up including both attended and unattended representations. Indeed, one of the morals of the studies by Garavan (1998), Lewis-Peacock et al. (2012) and Christophel et al. (2018) is that even fairly simple working memory tasks can involve shifts of attention from one item to another (see also Gomez-Lavin, *in preparation*). Carruthers and Block could perhaps try to redescribe such cases as instances where participants are relying on *both* working memory and some other form of memory. That is, they might claim that when attention is shifted among representations—such as when we engage in verbal rehearsal—the representations in

[18] Of course, the fact that most people believe *X* does not mean that *X* is true. In this case, however, since we are dealing with a terminological dispute, it makes sense to consider how the concept of working memory is actually deployed in the literature.

question actually being rapidly shunted in an out of working memory. However, this amounts to a serious revision of the concept, and thus requires further justification.

I conclude that working memory—*as it is standardly conceived of*—encompasses both attended and unattended representations. It should be emphasized, however, that the issues discussed in this chapter are important regardless of the exact terminology that one chooses to adopt. We have seen that the temporary maintenance and manipulation of information involves dynamic interactions between attention and mnemonic representations, with attention often being shifted between the relevant representations. This is a crucial discovery, which deserves careful consideration from both neuroscientists and philosophers. How such theorists choose to describe the relevant phenomena does not alter the significance of the empirical findings.

2.6. Conclusion

In this chapter, I critically examined the relationship between working memory and attention. Drawing on behavioral, neural and phenomenological data, I argued for two related conclusions: (i) attention plays an important role in working memory maintenance, but (ii) working memory representations can also temporarily persist outside the focus of attention. Thus, on the view I advocate, attention modulation is needed to maintain representations in working memory, but such attentional modulation can be sporadic rather than continuous. One unanswered question—which requires further research—concerns the *neural implementation* of working memory

representations outside the focus of attention. As we saw, a number of neuroimaging studies failed to find any evidence of neural activity for unattended working memory items (Lewis-Peacock et al., 2012; LaRocque et al., 2013, 2016; Rose et al., 2016). This has prompted some researchers to suggest that unattended working memory representations are "activity silent", being encoded via short-term changes in synaptic weights. Alternatively, however, it is also possible that unattended working memory representations display weak neural activity, which is simply difficult to detect using standard neuroimaging techniques (Christophel et al., 2018).

3. Working memory, consciousness, and conscious accessibility

In the previous chapter, I addressed the question of how working memory relates to attention. The present chapter moves on to discuss a different but related issue: how working memory relates to *consciousness*.[19] Almost everyone agrees that working memory and consciousness are related in some way, but there is significant disagreement over the details. Surveying the literature, there appears to be at least three different positions that one might take regarding working memory and consciousness:

1. The representations held working memory are always conscious (Bor & Seth, 2012; Carruthers, 2015; Stein, Kaiser & Hesselmann, 2016)
2. The representations in working memory can be either conscious or unconscious, but they are all *accessible* to consciousness (Baars 1997, 2001; Kintsch, Healy Hegarty, Pennington & Salthouse, 1999; Gilchrist & Cowan, 2010)
3. The representations in working memory can be either conscious or unconscious, and some are *inaccessible* to consciousness. (Soto, 2011; Soto & Silvanto, 2014)[20]

---

[19] It is important to recognize that these are in fact different issues. There is abundant evidence that attention can be deployed unconsciously (Cohen, Cavanagh, Chun & Nakayama, 2012), and even some evidence that consciousness can occur in the absence of attention (van Boxtel, Tsuchiya & Koch, 2010). As such, my conclusions regarding working memory and attention do not automatically transfer to the case of consciousness. Further work is needed to figure out how working memory and consciousness are related to one another.

[20] Technically, there is also a fourth option: working memory contents are *never* consciously experienced. I am not going to discuss this view in detail, as it has not received much attention or support in the literature (though see Jacobs & Silvanto, 2015). In Section 3, I outline some evidence that working memory and consciousness are (partially) overlapping, which marshals against such a "never-conscious" view.

The first position is the easiest to grasp: it says that *all* working memory representations are conscious, all of the time. The second position is slightly weaker. It allows that working memory representations may sometimes be stored unconsciously, but insists all working memory representation are *potentially conscious*, in the sense that they can be brought into conscious awareness at will. Finally, the last (and perhaps most contentious) position posits the existence of full-blown *subliminal* working memory. On this view, perceptual representations below the threshold for conscious access can still be retained in working memory, even in circumstances where we are unable to bring these representations to consciousness.[21]

My aim in this chapter is to adjudicate among the above three positions. I argue that we should reject the first position, as there is compelling evidence that working memory representations can be maintained in the absence of *immediate* conscious awareness. It is harder to decide between positions two and three, however. There is some preliminary evidence for subliminal memory maintenance (e.g., Soto & Silvanto, 2014; Trübutschek et al., 2017), however it is unclear whether the kind of memory involved meets all the functional conditions for *working* memory (Persuh, LaRock & Berger, 2018). As such, I arrive at a provisional conclusion: working memory

---

[21] Elsewhere (Beninger, *forthcoming*) I have tried to defend the existence of full-blown subliminal working memory. In this chapter I am more circumspect. Although I find the idea of subliminal working memory intriguing, I am no longer sure that it truly meets the requisite functional criteria for being considered *working* memory.

representations can be retained in an unconscious-but-accessible state, yet further evidence is needed to unequivocally establish the existence of subliminal working memory. The chapter is organized as follows. Section 3.1 clarifies the notion of "consciousness" I will be operating with, which eschews Ned Block's (1995, 2007, 2011) proposed distinction between phenomenal consciousness and access consciousness. Section 3.2 outlines some empirical evidence supporting a connection of some kind between working memory and consciousness. Section 3.3 then argues against the "always-conscious" view of working memory, demonstrating that working memory representations are sometimes retained in an unconscious (yet consciously accessible) state. Finally, section 3. 4 addresses the possibility of subliminal working memory.

3.1 Different kinds of consciousness?

The first thing we need to do is to clarify what is meant by the term "consciousness".[22] This matter is particularly pressing, as it has been argued—most prominently by Ned Block (1995, 2007, 2011)—that consciousness is non-unitary. Specifically, Block distinguishes between two different kinds of consciousness, which he calls *phenomenal consciousness* and *access consciousness* (Block, 1995). In the present section, I critically evaluate Block's dissociative view of consciousness. I ultimately reject

[22] See chapter 1 for an analogous discussion of the concept of working memory.

Block's proposed dissociation, arguing that genuine conscious states are *both* phenomenally experienced *and* cognitively accessed. This will set the stage for our subsequent discussion of the relationship between working memory and consciousness.

In his now-famous 1995 article, "On a confusion about the function of consciousness", Ned Block argues that there are in fact two different varieties of consciousness: (i) *phenomenal consciousness*, and (ii) *access consciousness*. Block defines phenomenal consciousness in experiential terms. For Block, phenomenal conscious states are states with qualitative character, such that "there is *something it is like* to be in [them]" (Block, 1995, p. 262; emphasis added). On the other hand, the notion of access consciousness is defined in information processing terms, rather than in terms of phenomenality. According to Block, "[a] perceptual state is access-conscious, roughly speaking, if its content—what is represented by the perceptual state—[…] gets to the Executive System, whereby it can be used to control reasoning and behavior" (Block, 1995, p. 229). In short, access conscious states are states that are made available to frontally-based executive centers, and can thus be used for various high-level cognitive activities (such as reasoning, reporting, action planning, etc.). Block explicitly ties access consciousness to *reportability*, claiming that access conscious states should be reportable in virtue of their widespread distribution throughout the cortex (Block, 2005). No such

criterion is imposed on phenomenal consciousness, however. On Block's view, some aspects of perception may be phenomenally conscious yet unreportable.[23]

Importantly, Block maintains that the distinction between phenomenal consciousness and access consciousness is not merely *possible*, but *actual* (Block, 1995, 2007, 2011). That is, he claims that phenomenal and access consciousness can come apart in certain experimental contexts. In particular, Block points to Sperling's (1960) partial report studies as a putative example of phenomenal consciousness without access consciousness. In Sperling's studies, participants were presented with an array of 12 letters (organized into a 3x4 array) for 15-500 milliseconds, immediately followed by an auditory retrocue directing their attention to one of the three rows (Figure 6). Sperling found that participants could only report about 3-4 letters from the grid in total, but—crucially—they could reliably report the letters from whichever row was selected by auditory retrocue (Sperling, 1960). Interestingly, participants also *claimed* to have seen the entire array, even though they could only report the identity of some of them (Block 2011, p. 567; Sperling, 1960). According to Block, Sperling's results thus support the conclusion that "perceptual consciousness overflows cognitive access" (2011, p. 567). Participants (allegedly) have a fleeting phenomenally conscious experience of all the

---

[23] It is worth noting that Block sometimes slips from describing access conscious states as *accessed* to describing them as merely *accessible* (Block, 1995; Carruthers, 2017b). Following Carruthers (2017b), I will adopt the former interpretation, according to which access conscious states (as Block understands them) are in fact *actually accessed* by the executive processes in question.

letters in the array before the retrocue, yet they can only *cognitively access* a subset of these phenomenally conscious representations.[24]

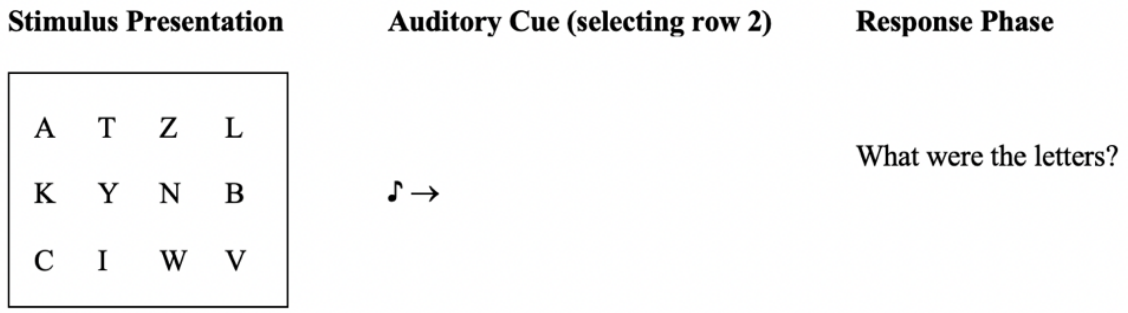

Figure 6: The Sperling partial report paradigm (Sperling, 1960). Participants are shown a letter-array, immediately followed by an auditory retrocue selecting one of the three rows. After a brief delay, participants then have to report the identity the cued letters.

Block's overflow argument has been widely criticized in the literature, however. One popular line of response to the overflow argument is to challenge Block's assumption that all the letters are consciously experienced in perfect detail before the retrocue. Instead, it may be that consciousness arises at a "late" stage in processing, only *after* the cued row has been selected (Kouider, de Gardelle, Sackur & Dupoux, 2010; Sergent, 2018). On this interpretation, participants are only truly phenomenally

[24] Block's overflow argument is sometimes framed as an argument that "phenomenology overflows *working memory*" (Brown, 2014, p. 1). This description is apt, insofar as the argument aims to establish that phenomenology overflows the amount of information that can be accessed, and thereby encoded in working memory. It should be emphasized, however, that working memory and cognitive access are not necessarily the same thing. Representations may need to be accessed in order for them to be *encoded* in working memory, but, once a representation is *in* working memory, they are not necessary accessed continuously. Indeed, as I will argue in Section 4, some working memory representations may be temporarily buffered in an *accessible* state, without thereby being accessed.

conscious of the identity of the letters that get selected by the auditory retrocue, and thus phenomenal consciousness *does not* overflow cognitive access. This interpretation still has to explain why participants *think* they saw all the letters, but this can be accounted for by appealing to "blurry" or "gist-like" phenomenology (Overgaard, 2018). Specifically, it may be that—in addition to consciously accessing the identity of the cued letters—participants can also access an indistinct gist-like representation of the surrounding characters, lacking specific letter information (Cohen & Dennett, 2011; Kouider et al., 2010; Overgaard, 2018). This appeal to gist-like phenomenology preserves the connection between conscious experience and cognitive access, since the vague experience of seeing *something* in the periphery of one's visual field *is* cognitively accessible and reportable (Kouider et al., 2010, p. 304). It is worth noting that there is some empirical support for the gist interpretation, coming from a study by de Gardelle, Sackur & Kouider (2009). De Gardelle and colleagues found that participants failed to notice when letters in the uncued rows were replaced by nonsense symbols, suggesting that participants are not aware of the *identity* of the uncued letters (as Block supposes).

Moving on, a more general worry about Block's phenomenal/access distinction has been raised by Cohen and Dennett (2011). Cohen and Dennett argue that the hypothesis of pure phenomenal consciousness without access consciousness is in fact scientifically intractable. To see the issue, try to imagine how we could possibly test for the existence of unaccessed phenomenally conscious brain states in a non-question

begging way. We cannot rely on participants' subjective reports—the standard marker of conscious experience—because reported information is necessarily accessed by higher-order executive centers (and thus does not qualify as unaccessed) (Cohen & Dennett, 2011, p. 362). But without relying on some kind of report, say Cohen and Dennett, there is no way forward: if we cannot *ask* participants to report on their experience, there is no way to know for sure whether a given pattern of brain activity is phenomenally conscious or not. As they put it, the existence of pure, unaccessed phenomenal consciousness "cannot be empirically confirmed or falsified and is thus outside the scope of science" (Cohen and Dennett, 2011, p. 358). One could of course insist that this is merely an epistemic limitation, and continue to hold that unaccessed conscious states exist, even if we cannot confirm or refute it. But this would be to leave the realm of scientifically-informed consciousness research and enter the realm of speculative metaphysics. Thus, practically speaking, we appear limited to the study of accessed and reportable conscious experiences.

Finally, Block's distinction between phenomenal and access consciousness has also been challenged on conceptual grounds. As Owen Flanagan (1992) points out, conscious experience seems bound up with the notion of a subject who is *having* the experience. On Flanagan's view, consciousness cannot come apart from cognitive access, because conscious experience *requires* that the relevant content be accessed by the subject in question. To quote Flanagan directly:

> Phenomenal feel necessarily involves access to whatever it is we feel. This access may be epistemically impoverished, as in the case of the prosopagnosiac … But all cases of phenomenal awareness, even if the awareness has no propositional content, are cases in which the agent has access to information about what state he is in. (Flanagan, 1992, p. 147)

From Flanagan's perspective, then, the idea of a conscious state that is completely inaccessible to the subject is conceptually problematic. Phenomenal consciousness is supposed to reflect what it is like *for some subject*, and this subject presumably has to access the conscious content in some way. Flanagan rightly notes that there might be instances where such access is "impoverished", such as in the case of prosopagnosic individuals who cannot consciously access detailed information about faces. But, even here, the individuals in question can still consciously access and report having a "fuzzy" visual experience of faces—it is just harder for them to discriminate between the faces of different individuals (Flanagan, 1992, 148; see also Fine, 2012). If we accept Flanagan's claim that conscious experiencing must come along with a subject who has access to it (which seems difficult to refute), Block's proposed dissociation between phenomenal and access consciousness appears to break down.

In light of the above considerations, I contend that we should reject Block's dissociative view of consciousness. Typical conscious states are *both* phenomenally experienced (having a distinct qualitative "feel" to them) *and* accessed by the subject in some way. Additionally, there are good *pragmatic* reasons for focusing on states that are both experienced and cognitively accessed, since only these accessed states are

scientifically tractable (Cohen & Dennett, 2011). For the remainder of this paper, then, I will take consciousness to involve both phenomenal experience and cognitive access. This ultimately makes our job easier. In assessing the relationship between working memory and consciousness, we do not have worry about addressing two separate questions—i.e., how working memory relates to phenomenal consciousness *and* how working memory related to access consciousness. Instead, we can simply ask how working memory related to consciousness, *tout court*, which is both experienced and accessed.

3.2 Evidence that working memory and consciousness are related

Now that we have addressed the phenomenal/access consciousness debate, we can move on to consider the relationship between working memory and accessible conscious experience (henceforth just "consciousness" or "conscious access"). The claim that working memory and consciousness are related in some way is not new. Many researchers, past and present, tend to assume that that the information in working memory is frequently consciously experienced (Baars & Franklin, 2003; Gilchrist & Cowan, 2010). However, the relationship between working memory and consciousness is often assumed, rather than explicitly argued for. My aim in the present section is to *justify* this assumption by appealing to a pair of recent empirical studies (Jantz et al., 2014; Koivisto, Ruohola, Vahtera, Lehmusvuo and Intaite, 2018).

Presumably, the main reason why most people take working memory and consciousness to be related has to do with inner phenomenological experience. When you hold a piece of information in working memory—say a line from a book you are reading—it *feels like something* to do so, with the rehearsed words being heard in the "mind's ear" (Baars, 1997, p. 41-44). This putative connection between working memory and conscious imagery has been empirically demonstrated in a recent study by Jantz et al. (2014). Jantz and colleagues had participants remember a set of numbers—either two numbers or six numbers—over an 11 second delay, followed by a memory recall test. During the delay, participants were also asked to monitor their own inner phenomenology, pressing a button whenever they experienced any memory-related mental imagery (either visual or auditory). As expected, Jantz and colleagues found that participants frequently reported having mental imagery, pressing the button at regular intervals throughout the delay. Interestingly, imagery rates were also modified by the number of items being held in working memory. In the low-load (two item) condition, participants pressed the "imagery button" an average of 5.32 times over the delay, whereas, in the high-load (six item) condition, participants pressed the button an average of 6.79 times over the delay (Jantz et al., 2014, p. 92). The presumed explanation of this effect, according to the authors, is that participants have to engage in more frequent (or faster) rehearsal when tasked with remembering more items, thereby leading to a greater amount of conscious imagery (Jantz et al., 2014, p. 97).

Next, a recent study by Koivisto and colleagues (2018) also provides empirical support for an association between working memory and consciousness. In one experiment, Koivisto and colleagues employed a dual task paradigm, which required participants to perform a subtraction task in working memory (counting backwards by threes) while also attempting to detect low-contrast Gabor patches briefly flashed on a computer screen (Figure 7). The aim of this design was to see if the subtraction task would interfere with conscious visual perception. The authors reasoned that, if working memory and consciousness are overlapping processes, then taxing working memory should lead to deficits in conscious perceptual processing. Additionally, Koivisto and colleagues also monitored participants' neural activity using electroencephalography (EEG) to see how working memory load modulated the signals associated with conscious perception.

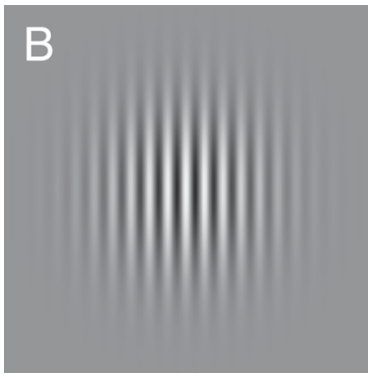

Figure 7: An example of a Gabor patch. From Arranz-Paraíso & Serrano-Pedraza (2018, p. 5) Copyright © 2018, Arranz-Paraíso & Serrano-Pedraza. Reprinted under the terms of the Creative Commons Attribution License (CC BY).

Interestingly, Koivisto and colleagues found that the backwards-counting task had a negative impact on conscious perceptual detection. Participants in the backwards-counting condition detected fewer low-contrast targets relative to controls who performed the detection task while merely *maintaining* a number, without manipulating it in any way (Koivisto et al., 2018, p. 93). Moreover, the concurrent counting task also had a noticeable effect on the event-related potentials produced by the to-be-detected perceptual stimuli. Specifically, it was found that P300 component—a waveform often associated with conscious perception (Sergent, Baillet & Dehaene, 2005)—was attenuated when participants were required to count backwards at the same time. These findings thus indicate that working memory and consciousness rely on a shared pool of resources, at least to some extent. It should be noted that, although Koivisto et al. found that conscious perceptual detection was disrupted by the counting task (which required working memory *manipulation*), no such deficits were observed for pure working memory *maintenance* (Koivisto et al., 2018, p. 89). As such, Koivisto and colleagues suggest that consciousness may be particularly involved in situations where working memory representations are required to be *used* or *manipulated*. They write that, "visual consciousness and WM share resources at a relatively late stage of conscious processing, which involves active manipulation of contents" (Koivisto et al., 2018, p. 86).

In sum, then, there appears to be persuasive evidence for an overlap between working memory and consciousness. In line with our phenomenological intuitions,

experimental work has shown that participants reliably report having conscious mental imagery during the performance of working memory tasks (Jantz et al., 2014). Moreover, working memory tasks that require manipulation (such as counting backwards) can disrupt conscious perception and its electrophysiological correlates, indicating that working memory and consciousness share some neural machinery (Koivisto et al., 2018).

3.3 Working memory representations are not always consciously accessed

Now that we have established that working memory and consciousness are related, we can proceed to consider the exact nature of this relationship. There is a tendency—especially in the philosophical literature—to assume that working memory representations are conscious by nature. For instance, Peter Carruthers holds that "the contents of working memory are always conscious" (Carruthers, 2015, p. 12), and, similarly, Wayne Wu (2014b, p. 163-164) writes that "working memory … memoranda are conscious at least in the access sense".[25] On the other hand, many researchers in psychology and neuroscience adopt a less stringent view, according to which working memory representations can be buffered outside of consciousness while still remaining *consciously accessible* (Baars, 1997, 2001; Kintsch, Healy Hegarty, Pennington & Salthouse, 1999; Gilchrist & Cowan, 2010). For example, Bernard Baars (1997, p.41-43) describes

[25] See Appendix A for a discussion of Jesse Prinz's view (Prinz, 2012), which differs slightly from both that of Carruthers and Wu.

working memory as theater stage, with only some of the "actors" (i.e., working memory representations) in the spotlight of consciousness at any one time. On Baars' view, representations outside the spotlight are unconscious, but they can be *brought to consciousness*, if the spotlight is shifted onto them. Similarly, Kintsch and colleagues (1999, p. 430-431) propose that, "[a]lthough there is a temptation to equate working memory to consciousness, such a simple equation of working memory and consciousness may not be warranted … [C]onsciousness may be a subset of the information that is maintained or that which is *accessible*." As stated at the outset, I endorse the second position, which allows that working memory representations can be unconscious while still remaining consciously accessible. The present section argues for this position on empirical grounds.

One piece of evidence supporting the existence of unconscious working memory view comes from an old behavioral study by Sternberg (1966; see also Baars, 2001). In this study, Sternberg had participants perform a simple delayed-response task that involved remembering a sequence of digits, ranging in length from one to six (Figure 8). Participants were initially presented with images of the to-be-remembered digits, followed by a two-second delay period. After the delay, participants were then shown a new "test" digit; their task was to indicate, as fast as possible, whether or not the test digit had appeared in original sequence (this was done by pulling one of two leavers). Sternberg then looked at participants reaction times on the task, to see if they varied as a

function of sequence length. Interestingly, he found that participants' response times increased linearly with the length of the remembered sequence. That is, it took participants *longer* to discern whether the test digit had appeared in original sequence when there were more digits to remember. Sternberg interpreted this finding as evidence that participants are not immediately conscious of *all* the remembered digits, but rather have to *search through* the digits—in a serial fashion—to see if there is a match. As he puts it, "[t]he linearity of the latency functions suggests that the time between test stimulus and response is occupied, in part, by a serial-comparison (scanning) process. An internal representation of the test stimulus is compared successively to the symbols in memory, each comparison resulting in either a match or a mismatch" (Sternberg, 1966, p. 653).

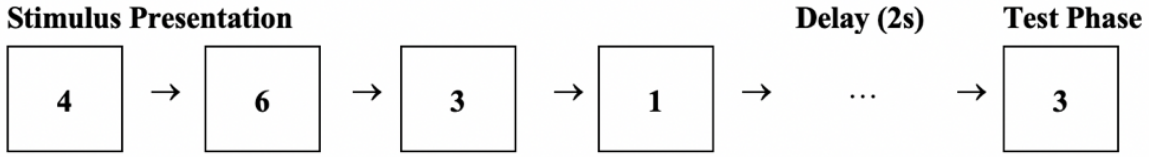

Figure 8: An illustration of the behavioral task from Sternberg (1966).

At this point skeptic might respond that the reaction time costs do not directly bear on the question of consciousness. Perhaps participants *do* shift attention amongst the remembered digits—but even the unattended representations may remain "peripherally" conscious (see Fortney, 2018 for this kind of view). This interpretation could potentially account for Sternberg's results, since it would still predict a reaction time cost for longer lists (owing to the attentional scanning process), even though all the

working memory representations remain conscious, to at least some degree. There are problems with this alternative interpretation, however. First of all, it is worth noting that this interpretation appears to be at odds with our internal phenomenology. When rehearsing a set of numbers (say, a phone number) in working memory, it doesn't seem like we are we are "peripherally" conscious of *all* of the digits not currently being rehearsed. Rather, consciousness feels like it is operating in a sequential manner, jumping from one digit to the next. Indeed, Baars notes that, anecdotally, "[m]ost people report that at any given moment, whatever number is *being said* in their inner speech is conscious; numbers that are not momentarily being said are not" (Baars, 1997, p. 44).

Next, the always-conscious view has trouble explaining the results from the above study by Jantz et al. (2014). Recall that Jantz et al. found that participants reported *more frequent* memory-related imagery on high-load (six item) trials, when compared to low-load (two item) trials. This finding makes no sense from the perspective of the always-conscious view: if participants are continually conscious of *all* working memory representations, the frequency of imagery should remain constant (and, indeed, be *unceasing*) regardless of the number of items being held in mind. By contrast, if we endorse a "scanning" view—whereby consciousness repeatedly scans through the remembered digits—we are in a better position to explain Jantz et al.'s findings. Participants in the high-load condition may experience more frequent episodes of mental imagery because they have to scan through the remembered digits at a faster

rate, with smaller temporal gaps between each successive conscious image. Ultimately, then, Sternberg's scanning interpretation appears to be far more plausible that the alternative. The patterns of imagery that participants report in the study by Jantz and colleagues suggests that the contents of working memory are "rehearsed effortfully … one bit at a time, but in a reiterative fashion" (Jantz et al. 2014, p. 97).

Moving on, further evidence for a dissociation between working memory and consciousness comes from the aforementioned study by Koivisto et al. (2018). Koivisto and colleagues found that a working memory task requiring *manipulation* (i.e., counting backwards) disrupted the conscious perception of low-contest targets. However, they also found that conscious perception was unaffected by pure maintenance. In particular, participants in a high-load maintenance condition (who had to remember seven letters) were just as good at detecting low-contrast perceptual stimuli as participants in a low-load maintenance condition (who only had to remember one letter) (Koivisto et al., 2018, p. 89). The fact that increasing the number of maintained items does not interfere with conscious perception suggests that mere maintenance may not always require consciousness. After all, seven items are thought to be around the upper limit of working memory capacity; if *all* these letters were being consciously accessed at once, we would presumably expect this to tax one's conscious resources, and thereby negatively impact perceptual detection. Since perceptual detection remained unaffected, it seems reasonable to assume that some (or all) of the remembered letters must have

been buffered unconsciously, thus freeing up conscious resources for the perceptual detection task. A plausible interpretation of Kovisto et al.'s results, then, is that—whereas working memory *manipulation* may require consciousness—mere working memory maintenance does not (see also Baars, 2003).

Finally, the existence of unconscious working memory is also indirectly supported by studies of "activity silent" working memory. As discussed in chapters 1 and 2, a number of studies have found evidence of working memory maintenance without continuous neural activity (Lewis-Peacock, Drysdale, Oberauer & Postle, 2012; Lundqvist et al., 2016; Rose et al., 2016). For instance, Lewis-Peacock et al. (2012) found that neural representations for remembered items tended to vary in the strength as a function of behavioral relevance, with representations often dropping to baseline if they were not immediately needed for the task at hand. Nevertheless, these representations remained *accessible*, and could be "reactivated" if they were needed for a subsequent memory test (see chapter 2 for a more detailed account of Lewis-Peacock et al.'s findings). A natural interpretation of this finding (and others like it) is that working memory representations can be temporarily stored in an inactive—and hence unconscious state—while still remaining *accessible* to consciousness.

One might challenge this interpretation on the grounds that a lack of *observed* activity does imply an *actual* lack activity. That is, it might be that the supposedly "silent" working memory representations are in fact encoded via weak activity that is

difficult to pick up using contemporary neuroimaging techniques (Christophel et al., 2018). This is a fair point. But even if the unattended representations in question are realized by weak activity (rather than some activity-silent synaptic mechanism), it is still highly unlikely that such representations are *conscious*. Much of the current research on the neural correlates of consciousness indicates that conscious brain states exhibit widespread "global" activation patterns, that are fairly recognizable (Sergent et al., 2005; Dehaene & Changeux, 2011). Moreover, even weak *subliminal* signals often display observable neural signatures (Dehaene et al. 2001). As such, the fact that working memory representations sometimes go completely "silent" (at least, from the perspective of our contemporary imagining techniques) gives us good reason to think that they are not consciously experienced.

The findings outlined in this section thus all converge on the same conclusion: working memory representations are not *always* consciously accessed. Instead, such representations may be temporarily buffered in an unconscious state for short periods of time. Note, however, that the unconscious representations in question still appear to be consciously accessible, in the sense that they can be "called back" to consciousness at will. In Sternberg's study, for instance, participants had no trouble judging whether the test digit had in fact appeared in the initial memory sequence—it just took them *time* to bring each of the remembered digits back to consciousness.

3.4 Subliminal working memory

So far, I have argued working memory is not *always* consciously experienced, but can sometimes be maintained in an unconscious-yet-accessible state. I now move on to consider an even more radical possibility: namely, that working memory can function *subliminally*, retaining information that is too weak to be brought to conscious awareness in the first place (Soto, Mäntylä & Silvanto, 2011; Soto & Silvanto, 2014; Trübutschek et al., 2017).[26] Although the notion of subliminal working memory might seem bizarre at first glance, there are a growing number of empirical studies that appear to support it (for reviews, see Soto & Silvanto, 2014; Beninger, *forthcomming*). In the present section, I outline some of the evidence for subliminal working memory, mainly focusing on studies by Trübutschek and colleagues (2017, 2018). I argue that, while these studies are provocative, it remains unclear whether the kind of memory involved meets all the functional conditions for *working* memory, as outlined in chapter 1 (see also Persuh et al., 2018).

Studies of subliminal working memory typically proceed by presenting participants with a target stimulus that has been rendered unconscious (via visual

---

[26] Many of the studies in question speak of "unconscious working memory" rather than "subliminal working memory". I prefer the term "subliminal" because it emphasizes that the representations in question are not merely unconscious, but also *inaccessible to consciousness*. Indeed, "unconscious working memory" is ambiguous, failing to distinguish between unconscious-yet-accessible representations, and full-blown subliminal representations.

masking, attentional blink, or continuous flash suppression) followed by a delay period and a subsequent memory test. The rationale behind this method is that, if participants can perform at above-chance levels on the memory test, then they must have maintained a subliminal representation of the target over the delay (Soto et al., 2011). One recent example of this method comes from a study by Trübutschek and colleagues (2017). In this study, Trübutschek and colleagues presented participants with a grey square which had been masked using a metacontrast display (Trübutschek et al., 2017, p. 4; Figure 9). Participants were tasked with trying to remember the location of the subliminally-perceived square over a 2.5-4 second delay, and then with reporting its location in a subsequent test phase (there were 20 possible locations to choose from, organized in a circle around a fixation point). Participants were also asked to indicate whether or not they had seen the square—using a four-point perceptual awareness scale—in order to ensure that the stimulus was in fact *subliminal*.

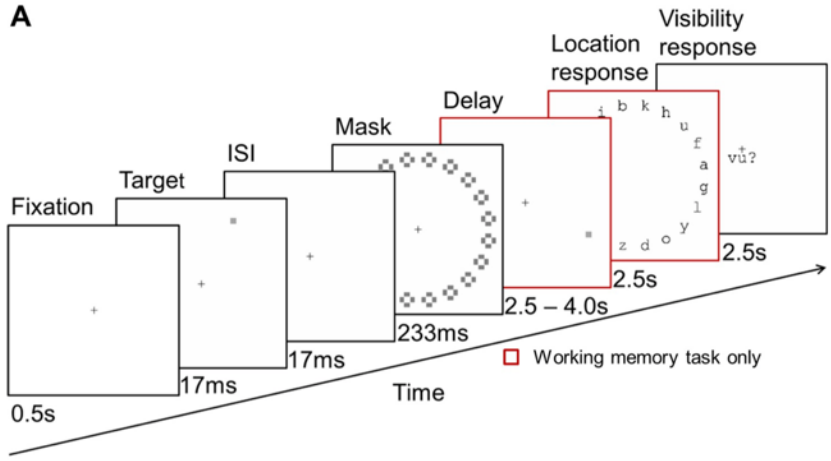

Figure 9: The behavioral paradigm used by Trübutschek et al. (2017). From Trübutschek et al. (2017, p. 4). Copyright © 2017, Trübutschek, et al. Reprinted under the terms of the Creative Commons Attribution License (CC BY).

Trübutschek et al. found that participants were able to identify the location of the subliminally-perceived square at a rate that exceeded chance (22% vs. 5% chance), indicating that participants could retain subliminal information about the memory item. Additionally, participants performed equally well in a secondary distractor condition—in which an irrelevant distractor square was presented during the delay—thus demonstrating that the subliminal memory effect was not easily disrupted. Trübutschek et al. interpret their results as supporting the existence of subliminal working memory. They write that, "our findings support the existence of genuine working memory in the absence of … conscious perception" (Trübutschek et al., 2017, p. 21). Is unclear whether this conclusion is warranted, however. Can we indeed be sure that the finding reflects the operation of subliminal *working* memory, rather than some other distinct type of memory phenomenon (such as sensory memory or priming; see Persuh et al. 2018)? To

answer this question, I suggest that we return to the four functional criteria for working memory that I laid out in chapter 1: (i) information maintenance, (ii) information manipulation (or manipul*ability*), (iii) capacity limitations, and (iv) distractor resistance. If subliminal working memory indeed exists, then it should exhibit all four of these functional characteristics.

Looking at the study by Trübutschek et al. (2017), it is obvious that conditions (i) and (iv) are satisfied: participants clearly *maintained* subliminal information, and the retained information was also *resistant to distraction*. Things are less clear, however, when it comes to the other two criteria. First, the participants in Trübutschek et al.'s study were merely required to passively hold the locations of the targets in mind, and thus it remains unknown whether the subliminal information in question is available for manipulation. Indeed, Persuh and colleagues (2018) have recently argued that we should be skeptical of *all* purported demonstrations of subliminal working memory—including Trübutschek et al. (2017)—precisely because the studies in question fail to meet the criterion of manipulability. They state:

> [W]e take it that perhaps the key feature of WM … is that information encoded in WM is available for use in ongoing tasks … Many purported studies of unconscious [working memory] that do not involve information manipulation are open to the criticism that the information gleaned from unconsciously perceived targets merely primes participants or is stored in (fragile) VSTM" (Persuh et al. 2018, p. 8-9).

I agree with Persuh et al. that the original study by Trübutschek et al. (2017) does not speak to the question of manipulability. Intriguingly, however, the same research

team has recently conducted another study (currently in preprint) which comes closer to demonstrating the manipulation of subliminally perceived information (Trübutschek, Marti, Ueberschär & Dehaene, 2018). In this study, the authors had participants perform a similar task—involving the localization of a subliminally perceived target square—with one major difference. Specifically, rather than having participants report the *original* location of the subliminally-perceived square, Trübutschek and colleagues instead asked participants to try to *rotate* the square left or right 120°, and then to report the result of this transformation. Fascinatingly, Trübutschek and colleagues found that participants *still* performed better than chance (albeit to a lesser degree), despite the fact that they were being asked to rotate and item that they had never consciously perceived. This finding thus appears to provide some preliminary evidence that the subliminal memory items may be available for manipulation, at least in some cases.

Unfortunately, however, even this result may not be *completely* definitive. For one thing, I have claimed (in chapter 1) that working memory manipulation is a personal-level, voluntary procedure. It is hard to determine—on the basis of the behavioral results alone—whether or not the "manipulation" observed in the above study was undertaken voluntarily. It is worth noting that participants could "choose" to rotate the target square either left or right (based on the experimental instructions), which seems to *suggest* that they have some degree of voluntary control. But it is hard to say for sure, given the current paucity of evidence. Additionally, there is another possible

interpretation of the results, which does not appeal to the manipulation of subliminal contents. Specifically, participants may have accomplished the rotation task by generating a *conscious guess* of the memory target—which is primed by their subliminal perception—and then proceeded to rotate this guess in working memory (Trübutschek et al. 2018 p. 18-19; Stein, Kaiser & Hesselmann, 2016). On this interpretation, it is not truly the subliminal memory representation that is available for manipulation; rather, participants are spontaneously generating a *new* conscious "guess" representation during the test phase, and performing manipulations on this representation instead (Jacobs & Silvanto, 2015). All in all, then, the manipulability of subliminal memory representations remains underdetermined. While the study by Trübutschek et al. (2018) is suggestive, further research is needed before we can draw any strong conclusions.

Finally, we still have to consider the third criterion: *capacity limitations*. As far as I know, no studies to date have investigated the capacity limits of subliminal working memory. There is one recent study—again by Trübutschek, Marti and Dehaene (2019)—which found that participants could successfully maintain the locations of *two* target squares over a 2.5 second delay. However, as the authors themselves note, the memory load of two items "[falls] well within the capacity limits of conscious working memory" (Trübutschek et al. 2019, p. 8). As such, we still lack evidence that so-called "subliminal working memory" has a capacity limit that mirrors that of ordinary conscious working memory (i.e., somewhere in the ballpark of four to seven items). To be fair, it seems

highly likely that such subliminal maintenance will turn out to be highly limited in capacity, given the inherent difficulty of remembering subliminal information. But, strictly speaking, we do not yet have evidence to support this conclusion.

Summing up, then, subliminal memory unambiguously satisfies two of the functional criteria for working memory—maintenance and distractor resistance. But we still cannot be sure whether it satisfies the conditions of personal-level (voluntary) manipulability and capacity limitedness (see also Persuh et al., 2018). As such, I suggest that, for the time being, we should remain skeptical of subliminal working memory. It may turn out that subliminal working memory does indeed exist, but, at present, the evidence remains inconclusive. My view here is similar to that of Persuh et al. (2018), who also recommend caution when interpreting the above studies of subliminal working memory. However, my view differs from theirs in one crucial respect. Specifically, Persuh et al. do not draw a distinction between unconscious-yet-accessible and subliminal working memory, and thus they end up advocating for a wholesale skepticism about unconscious working memory. My position is different in that I endorse the existence of unconscious-yet-accessible working memory. I am only skeptical of working memory for full-blown subliminal information, which is consciously *inaccessible*.

3.5 Conclusion

I began this chapter by outlining three possible positions on the relationship between working memory and consciousness:

1. The representations held working memory are always conscious.
2. The representations in working memory can be either conscious or unconscious, but they are all *accessible* to consciousness.
3. The representations in working memory can be either conscious or unconscious, and some are *inaccessible* to consciousness.

As we have seen, the first position appears to be incorrect. Studies by Sternberg (1966) and Jantz et al. (2014) indicate that we frequently scan through the items in working memory, with consciousness "illuminating" representations one after another (notably, this is in line with my view on working memory and attention, which allows for attentional scanning within working memory). Deciding between positions two and three is more difficult, however, as the study of subliminal "working memory" is relatively new, and therefore underdeveloped. Subliminal memory certainly exhibits *some* of the functional characteristics of working memory—including information maintenance and distractor resistance—but it remains unclear whether subliminal memory representations are subject to voluntary manipulation, and whether subliminal memory mechanisms have the same capacity limitations as ordinary consciously-accessible working memory. For this reason, I suggest that we provisionally adopt

position two: working memory representations can be either conscious or unconscious, but they are (to the best of current our knowledge) *consciously accessible*.

4. Is working memory sensory-based?

Working memory is believed to be "a temporary system under attentional control that underpins our capacity for complex thought" (Baddeley, 2007, p. 1). It is what enables us to hold information in mind (often consciously) for use in cognitive tasks, such as reasoning and decision making (Baddeley, 2003; 2007). In several recent publications, Peter Carruthers (2014; 2015; 2017a) has argued for a *sensory-based* account of working memory. The main tenet of Carruthers' view is that the representations held in working memory are always grounded in sensory imagery, depending on "activity in mid-level sensory areas of the brain" (Carruthers, 2014, p. 150). While non-sensory representations like concepts and propositional attitudes still *exist*, on Carruthers' view, these kinds of representations cannot gain access to working memory (except indirectly, by being "bound into" mid-level sensory representations (Carruthers, 2015, p. 101)). One major upshot of this view, is that "conscious reflection is sensory based" (Carruthers, 2015, p. 51). Since working memory is assumed to undergird our capacity for conscious reflection—and working memory is itself sensory-based—reflection always manifests in the form of sensory images.[27]

---

[27] On Carruthers' view, *all* working memory representations are conscious (Carruthers, 2015, p. 82-88). I disagree with Carruthers on this issue (see chapter 3), however I am not going to press the point here. My aim in this chapter is to argue against the *sensory-based* aspect of Carruthers position.

In the present chapter, I critically evaluate Carruthers' sensory-based account of working memory. Drawing on various findings from cognitive science, I argue that Carruthers' view is incorrect: not only is Carruthers' main argument for the sensory-based account suspect, but there is actually positive evidence for the existence of *non-sensory* working memory representations, instantiated outside of sensory cortices. As such, I suggest that we need a more permissive account of working memory that allows for the existence of *both* sensory and non-sensory forms of working memory. The overall structure of the chapter is as follows. In section 4.1, I outline Carruthers' sensory-based account of working memory, beginning with a brief summary of his views on consciousness and attention. In section 4.2, I critique Carruthers' main argument for his sensory-based account—the "targets of attention" argument. In section 4.3, I present several lines of empirical evidence that support the existence of non-sensory working memory. Finally, in section 4.4, I sketch an alternative to Carruthers' position and defend it against two further objections.

Before we get started, two clarifications are in order. First, it is necessary to say a bit more about the distinction between sensory representations and non-sensory representations. A sensory representation, as I am using the term, refers to a representation that is: (i) instantiated within a dedicated perceptual-processing area of the cortex (e.g., the occipital lobe in the case of vision), and (ii) couched in a modality-specific code (Prinz, 2002, Ch. 5; Dove, 2009). A non-sensory representation, by contrast,

is just a representation that lacks these properties. Non-sensory representations are instantiated in higher-level association areas, which lie outside of the primary perceptual systems, and they are "amodal" in the sense that they can be triggered by inputs from multiple modalities. Second, it should be noted that the main issue at stake in this chapter—whether working memory is sensory based—is different from the philosophical debate over the existence of cognitive phenomenology (Flanagan, 1992, p. 67-8; Bayne & Montague, 2011). Carruthers' claim that working memory is sensory based is a claim about the *representational format* of working memory representations, not a claim about the putative existence of cognitive qualia. Some of my conclusions may bear indirectly on the cognitive phenomenology debate, but I won't try to draw such connections here.[28]

4.1 Carruthers' sensory-based account

In this section, I provide an overview Carruthers' sensory-based account of working memory. I begin by articulating Carruthers' views regarding the nature of consciousness and attention (since this is important background information), and then go on spell out his sensory-based account of working memory in more detail.

---

[28] For discussion of how neuroscientific considerations may inform the cognitive phenomenology debate, see Kemmerer (2015) and McClelland and Bayne (2016).

4.1.1 Consciousness and attention

In formulating his account of working memory, Carruthers relies on the "global broadcasting" theory of consciousness developed by Bernhard Baars (2002) and Stanislas Dehaene (2014). The central tenet of the global broadcasting theory is that information reaches conscious awareness when it gets widely distributed (or "globally broadcast") throughout the brain (Carruthers, 2015, p. 48). On this theory, the difference between conscious and unconscious information processing is ultimately a matter of scope. A representation will remain unconscious so long as it stays encapsulated within modular sensory-processing areas; by contrast, a representation will become conscious when it gets broadcast to many different brain areas, "including those for forming memories, issuing in affective reactions, as well as a variety of systems for inference and decision making" (Carruthers, 2015, p. 48; Dehaene & Naccache, 2001). As Carruthers notes, the global broadcasting theory gains support from a number of empirical studies examining the neural correlates of conscious visual perception (Dehaene et al., 2001; Sergent, Baillet & Dehaene, 2005; Dehaene & Changeux, 2011). These studies find that when a visual stimulus goes *unseen* (due to visual masking or the attentional blink), there is localized activity in posterior visual-processing areas, such as the occipital cortex and posterior temporal cortex. When the stimulus is *consciously seen*, however, this local activity is accompanied by widespread reverberating activity across frontal and parietal cortices (see Dehaene, 2014, Ch. 4).

At this point, a further question arises: if consciousness is the result of global broadcasting, then what determines whether or not a given representation will be globally broadcast? Carruthers' answer is that *attention* is the key mechanism that enables global broadcasting. He tells us that, "attention operates by boosting the neural activity of some groups of neurons while simultaneously suppressing the activity of competing populations, resulting in global broadcast of the information encoded in the former set" (Carruthers, 2013, p. 10375). According to Carruthers, attention is controlled by a distributed frontoparietal network, which primarily includes the dorsolateral prefrontal cortex, frontal eye-fields and intraparietal sulcus (Carruthers, 2015, p. 60-64). This network helps to initiate the global broadcasting by sending top-down, activity-enhancing signals towards the relevant (to-be-broadcast) sensory representations. Carruthers is explicit that attention is a "*necessary condition* for global broadcasting to occur" (Carruthers, 2015, p. 16, emphasis added). He stops short of saying that attention is *sufficient* for global broadcasting, however, as there is mounting evidence that attention can sometimes modulate subliminal stimuli without bringing them to consciousness (e.g., Kentridge, Nijboer & Heywood, 2008).

Importantly, Carruthers also maintains that attention "has an exclusively sensory focus" (Carruthers, 2015, p. 92). On his view, the attentional-control network is structured such that it can only direct attention towards so-called "mid-level" sensory areas. Carruthers is not always explicit about what constitutes a mid-level sensory area;

however, it is clear that he takes such areas to be located within dedicated perceptual systems, situated somewhere between the lowest and highest processing stages. When it comes to vision, for instance, Carruthers tells us that mid-level visual areas "include the regions that process color, form, motion, spatial layouts, and faces, but not the primary visual projection area V1, and not the regions of temporal and parietal association cortices that receive output from mid-level regions" (Carruthers, 2014. p. 147). Carruthers' bases his claim that attention exclusively targets mid-level sensory areas on an apparent *lack of evidence*. He insists that, while we have an abundance of evidence for attentional modulations in sensory areas, "there is no evidence that they can be directed toward [non-sensory] association areas", such as the frontal cortex (Carruthers, 2015, p. 91). According to Carruthers, the best explanation for this lack of evidence is that attention *can only* target mid-level sensory areas.

4.1.2 Working memory as stimulus-absent broadcasting

Now that we have surveyed Carruthers' views on consciousness and attention, we are ready to proceed to his account of working memory. According to Carruthers, working memory simply reflects the endogenous (stimulus-absent) operation of the global broadcasting system (Carruthers, 2015, p. 76-79). In the case of ordinary conscious perception, the mid-level sensory representations that get attended and globally broadcast are driven by external stimuli. In the case of working memory, by contrast, mid-level sensory representations are attended and broadcast *in the absence of* external

stimulation. To illustrate, suppose that you are currently entertaining a visual representation of an object—say, of a coffee cup—that is about to be removed from view. While the cup is still being perceived, there will be an exogenously-produced sensory representation in your visual cortex, which can be attended and broadcast in the normal way. Once the cup *leaves* your visual field, information about the cup will cease to be perceptually available. Yet, according to Carruthers, this representation can still be maintained and broadcast in working memory so long as you continue to direct attention towards it.

Again, Carruthers bolsters his view with empirical evidence. He points out that many neuroimaging studies of working memory find activity in mid-level sensory areas during memory maintenance, suggesting that participants are indeed actively sustaining sensory representations (see, e.g., Postle, 2006; Harrison & Tong, 2009; Sreenivasan, Curtis & D'Esposito, 2014). Moreover, working memory also appears to "[use] the same top-down attentional mechanisms … that are implicated in conscious forms of perception" (Carruthers, 2015, p. 104; Naghavi & Nyberg, 2005). The natural conclusion to draw from these findings, says Carruthers, is that working memory and conscious perception share the same neural substrates: working memory constitutes the "off-line" operation of the perceptual broadcasting system (Carruthers, 2015, p. 74).

So how does this lead to the conclusion that working memory is *sensory-based*? Ultimately, the reason has to do with Carruthers' dual commitments regarding (i) the

nature of working memory and (ii) the sensory focus of attention. Carruthers believes that working memory depends on the deployment of attention, since attention is the mechanism that enables representations to be globally broadcast. Yet, in addition, Carruthers *also* holds that attention only ever targets mid-level sensory areas. Putting these two claims together, Carruthers arrivers at the conclusion that only mid-level, sensory-based representations are capable of being attended and broadcast within working memory. As Carruthers puts it:

> The generally accepted current picture [of working memory], then, is of a set of executive systems that deploy attentional and other resources to recruit activity in mid-level sensory areas of the brain, resulting in globally broadcast representations that can be sustained, manipulated, or replaced by further actions of the executive . . . But [working memory] is a sensory based system. What figures within it are not (in general) propositional attitudes, but rather visual images, auditory images, imagined movements, and so forth. (Carruthers, 2014. p. 150)

There is one important caveat to this conclusion, however. Although Carruthers holds that working memory is sensory-*based*, he allows that non-sensory representations—like concepts and judgements—may still feature *indirectly* in working memory, by being "bound into the contents of working memory images" (Carruthers, 2015, p. 16). Unfortunately, Carruthers does not provide much detail about the nature of such binding. He seems to have in mind something like cognitive penetration, whereby amodal conceptual areas causally influence activity in lower-level sensory areas (Vetter & Newen, 2014.) In the case of *visual* binding, for instance, Carruthers says that "[i]nformation from conceptual regions is projected back to earlier stages of processing,

influencing the latter, and becoming a component in the perceptual state that follows visual recognition" (Carruthers, 2015, p. 66). This qualification does not change the overall picture too much, though. On Carruthers view, it is still the mid-level sensory vehicles that get targeted by attention and which serve as the basis for global broadcasting. Any contribution from amodal areas must be relayed back to mid-level sensory areas and be incorporated into an imagistic sensory representation. Carruthers remains adamant that "*purely amodal* attitudes are incapable of figuring among [working memory's] contents" (Carruthers, 2015, p. 15; emphasis added).

4.2 The targets of attention argument

Now that we have summarized Carruthers view and the reasoning behind it, let us turn a critical eye to his main argument for the conclusion that working memory is sensory based. This argument, dubbed "the targets of attention" argument, has already been informally outlined above. Nevertheless, it will be helpful to provide an explicit formulation, so that we can isolate and examine the relevant premises. Here is a straightforward formulation of the argument, which seems to capture Carruthers' basic line of reasoning (see Wu (2014b) and Chudnoff (2016) for slightly different versions):

(1) Attention is required for representations to be globally broadcast, and hence for them to be held in working memory.
(2) Attention only targets mid-level sensory areas.
(3) Therefore, only sensory-based representations can be held in working memory.

Premise 1 encapsulates Carruthers' view regarding how information is encoded and stored in working memory—for Carruthers, information only gains entry to working memory when it has been attended and globally broadcast. Premise 2 is a restatement of Carruthers' claim that attention is solely directed at mid-level sensory areas. Recall that Carruthers supports this second premise by appealing to an (alleged) absence of evidence. According to Carruthers, we don't have any evidence for attention outside of sensory areas, and the best explanation for this lack of evidence is just that attention *can't* be directed to non-sensory parts of the cortex.

Carruthers argument is certainly provocative, but is it convincing? I contend that the answer is "no". The main problem with the argument, on my view, is that the second premise is empirically dubious. Contrary to Carruthers' claim that attention only targets mid-level sensory areas, there *are* in fact a handful of studies that demonstrate attentional effects in non-sensory regions of the cortex (Lau, Rogers, Haggard & Passingham, 2004; Ester, Sutterer, Serences & Awh, 2016; Rowe, Friston, Frackowiak & Passingham, 2002). In what follows, I outline two such studies by Lau and colleagues (2004) and Ester and colleagues (2016). As we will see, these studies put serious pressure on Carruthers' view that attention exclusively targets mid-level sensory areas.[29]

---

[29] It is worth noting that one might also challenge Carruthers' first premise, arguing that even subliminal (i.e., non-broadcasted) representations can be stored in working memory (see Chapter 3). I won't pursue this line of response here, however.

The first example of attention outside of sensory areas comes from a study by Lau, Rogers, Haggard and Passingham (2004), which examined the possibility of *attention to intention*. In this study, participants were asked to perform a Libetian (1999) temporal-judgement task while undergoing functional magnetic resonance imaging (fMRI). Participants were instructed to press a button whenever they felt like it and to subsequently make reports about the timing of their motor actions (as gauged by the position of a light moving rapidly around a clock face). There were two conditions: (i) the attention-to-intention condition and (ii) the attention-to-movement condition. In the former, participants had to report the time at which they initially felt the *intention* to move, whereas, in the later, participants had to report the time at which their motor movement *actually* occurred (Lau et al., 2004, p. 1208). Lau and colleagues then contrasted to the fMRI data in the two conditions to see if attending to one's *intention* (rather than merely attending to one's movement) corresponded to any changes in neural activation.

Intriguingly, Lau and colleagues found that attending to intention modulated activity in the pre-supplementary motor area (pre-SMA), such that the pre-SMA was *more active* in the attention-to-intention condition relative to the attention-to-movement condition. Lau and colleagues further note that this boost in activity "cannot be due to preparing for or initiating a self-paced action", since these aspects of the task were matched across the two conditions (Lau et al., 2004, p. 1209). This finding is important in

the present context because the pre-SMA is located in the frontal cortex and clearly lies outside of mid-level sensory areas. The fact that attending to one's intention enhances pre-SMA activity thus suggests that—*contra* Carruthers—attention *can* target non-sensory association areas of the frontal cortex. It is worth noting that Lau and colleagues also examined the functional connectivity between the pre-SMA and the dorsal prefrontal cortex (dorsal PFC), since the dorsal PFC is a prime candidate for driving top-down attentional modulation. As hypothesized, they found that dorsal PFC and the pre-SMA and were more strongly coupled in the attention-to-intention condition relative to the attention-to-movement condition (Lau et al., 2004, p. 1209). Although the analysis was correlational (rather than causal), the finding that the dorsal PFC and pre-SMA and are activated synchronously is consistent with the view that the dorsal PFC is responsible for top-down modulation in the pre-SMA.

Another example of attentional effects outside of traditional sensory areas comes from a more recent study by Ester, Sutter, Serences and Awh (2016). Ester and colleagues had participants perform a visual discrimination task, while simultaneously measuring their neural activity with fMRI. The task involved monitoring a flickering square-wave grating for potential changes in its orientation/luminance (Figure 10). In one condition (the "attend-orientation condition"), participants were specifically tasked with reporting changes in the grating's orientation, while ignoring changes in its luminance. Conversely, in a second condition (the "attend-luminance condition"),

participants were tasked with reporting changes in the grating's luminance, while ignoring changes in orientation (Ester et al. 2016, p. 8189). This manipulation allowed Ester and colleagues to pull apart the effects of attention to orientation from the effects of attention to luminance.

Figure 10: A depiction of a square wave grating. Redrawn after Ester et al. (2016).

In analyzing the data, Ester and colleagues began by identifying the regions of the brain that were sensitive to changes in orientation of the grating. With the aid of multivariate pattern analysis, they showed that orientation-specific information was encoded throughout the brain, ranging from the occipital cortex to regions of the parietal cortex and (even) the frontal cortex (Ester et al., 2016, p. 8192). Next, Ester and colleagues examined the effect of attention on these orientation-specific representations, contrasting the levels of activation across the attend-orientation condition and the attend-luminance condition. The intriguing result was that, in the attend-orientation condition, the neural representations of orientation were amplified in several frontal/parietal sites—including the left precentral sulcus and right inferior parietal

lobule (Ester et al., 2016, p. 8194). Notably, the precentral sulcus is located in the frontal cortex, and certainly does not qualify as mid-level sensory areas in Carruthers' sense. Indeed, Ester and colleagues take their results to show that "feature-based attentional modulations are distributed across the visual processing hierarchy, including regions typically associated with attentional control rather than sensory processing" (Ester et al., 2016, p. 8197). It should be noted that Ester and colleagues' study does not provide evidence for *pure* non-sensory attention, since the attended-to feature—i.e., orientation—is still a visual property. Nevertheless, the study establishes that attentional modulation is not restricted to mid-level sensory regions (as Carruthers would have it), but rather permeates higher-level association areas.

In light of the above two studies, I conclude that Carruthers' second premise is empirically suspect. There is in fact some evidence for attentional modulations in higher-order brain regions that lie outside of mid-level sensory areas. As such, Carruthers' main argument for his sensory-based account is unconvincing. To be fair to Carruthers, it is true that the evidence for non-sensory forms of attention is fairly sparse, especially when compared to the vast empirical literature on sensory attention (Petersen & Posner, 2012). However, this merely illustrates that non-sensory attention is understudied—not that it doesn't exist.[30]

---

[30] A different kind of objection to Carruthers' argument is raised by Wayne Wu (2014b). Wu challenges the *validity* of Carruthers' argument: he claims that, even if attention *initially* targets mid-level sensory areas,

4.3 Evidence for non-sensory working memory

So far in this chapter I have been on the defensive, challenging Carruthers' argument for his sensory-based account of working memory. In this section, I now move to consider the positive case for non-sensory forms of working memory. Some philosophers might be inclined to think that the existence of non-sensory working memory can be proven on the basis of introspection alone. For instance, Charles Siewert (1998) takes it to be introspectively obvious that we can consciously entertain thoughts without any corresponding sensory imagery. Unfortunately, however, introspection is unlikely to provide a definitive conclusion, as different philosophers appear to have different introspective intuitions. In contrast to Siewert, other philosophers—like Tye (1995) and Prinz (2012)—*deny* that they have experiences of non-sensory thoughts. For this reason, I am going to focus instead on *empirical* evidence for non-sensory working memory. Luckily, as we will see, there are in fact a number of empirical findings that support the existence of non-sensory working memory.

---

attentional effects may *percolate up* to higher-level brain regions—thereby leading to the broadcasting of amodal representations (Wu, 2014b, p. 170). On Wu's view, then, amodal representations can still gain access to working memory by being "triggered" by attentionally-modulated activity in sensory areas. This is a partial concession to Carruthers' view, since Wu is taking it for granted that attention circuitry only projects to mid-level sensory areas. But, for Wu, the sensory focus of working memory is compatible with the genuine existence of amodal working memory. My view is somewhat different from Wu's, as I hold that attention can *directly* target non-sensory areas. In the study by Lau et al. (2004), for instance, it seems likely that attentional signals are being projected straight to the pre-SMA, since we know that "there are direct anatomical connection between the [dorsal] PFC and the pre-SMA" (Lau et al., 2003, p. 1209).

4.3.1 Electrophysiological studies

The first strand of evidence for non-sensory working memory derives from electrophysiological studies with monkeys. In one landmark study, Nieder (2012) tested monkeys' ability to remember small numerosities while recording activity from neurons in the lateral PFC. On each trial, monkeys were presented with either (i) a number of visual dots or (ii) a number of auditory tones (the number of tones/dots could range from 1-4) (Nieder, 2012, p. 11860). Next, after a brief memory delay, a "test" display appeared on the screen, consisting of an array of simultaneously-presented dots. The monkeys' task was to indicate whether the number of dots in the test display *matched* the number of items that had previously been presented in either the visual or auditory modality (Nieder, 2012, p. 11860) (Figure 11). As expected, the monkeys performed the task with a high degree of accuracy, answering correctly around 90% of the time. More importantly, however, Nieder found that neurons in the prefrontal cortex "coded" for specific numerosities over the delay (Nieder, 2012, p. 11860). Certain neurons fired most vigorously for presentations of one auditory/visual item, whereas other neurons fired most vigorously for presentations of two auditory/visual items (and so on for set sizes of three and four). Crucially, the neurons in question "[responded] to numerosity irrespective of the sensory modality" (Nieder, 2012, p. 11860): that is, the same neurons that encoded information about a specific numerosity in the visual domain also encoded information about the same numerosity in the auditory domain. This study thus

supports the existence of non-sensory (or "amodal") working memory. Nieder's results show that neurons in the monkey lateral PFC encode and store information in a modality-independent format, which abstracts away from sensory details.

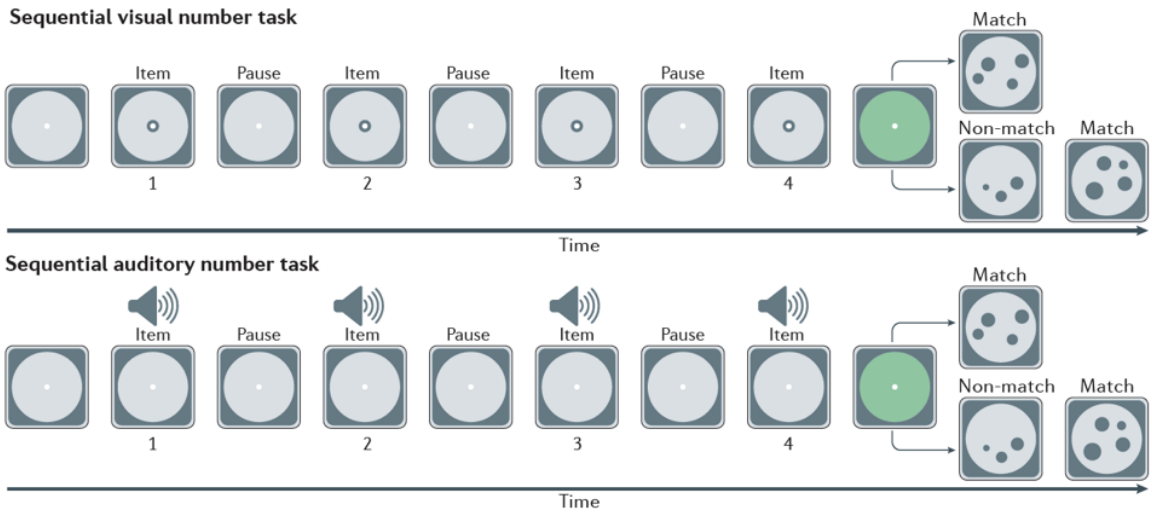

Figure 11: The behavioral task from Nieder (2012). Monkeys were presented with a series of tones or dots. Then, after a brief delay, they had to indicate whether or not a test array matched the number of previously-presented visual/auditory items. From Nieder (2016, p. 368). Copyright © 2016 Nature Publishing Group. Reprinted with permission.

Another recent study by Vergara, Rivera, Rossi-Pool and Romo (2016) corroborates the existence of amodal working memory neurons in the frontal cortex. Vergara and colleagues had monkeys perform a delayed-response task involving auditory and vibrotactile stimuli. On each trial, monkeys were presented with a sequence of tones or tactile stimulations, oscillating on and off at a particular frequency (between 8Hz and 32Hz). This was followed by a delay of three seconds, after which a second "test" sequence was presented (either in the auditory or tactile modality). The

monkeys' task was "to indicate whether the frequency of the second stimulus was higher or lower than the first" (Vergara et al., 2016, p. 55). During the performance task, Vergara and colleagues recorded the activity of 205 individual neurons in the pre-SMA.[31] Intriguingly, they found that a significant portion of these neurons encoded information about the frequency of the remembered sequence over the delay. The firing rates of these "frequency neurons" were linearly proportional to the remembered frequency, with higher frequencies often corresponding to higher firing rates (Vergara et al., 2016. p. 55). Again, however, the neurons did not show a preference for modality: their firing rate for an *auditory* sequence of frequency *X* was the same as their firing rate for a *tactile* sequence of frequency *X* (Vergara et al., 2016, p. 61). As the authors point out, this finding "supports the existence in pre-SMA of a supramodal (modality independent or amodal) neural code for short-term memory representations" (Vergara et al., 2016, p. 61).

One possible response Carruthers could make at this point would be to deny that the representations in question reflect the actual *contents* of working memory. That is, he might try to argue that the relevant activity observed in the dorsal PFC and pre-SMA occurs outside the attentionally-mediated workspace of working memory, and instead merely serves to *coordinate* activity in mid-level sensory areas (see Prinz, 2002, p. 137). I

---

[31] Note that this is the same area where Lau et al. (2004) found evidence of attentional modulations related to intention.

find this response problematic for two reasons. First, as we saw in section 4.2, there is evidence that the brain areas in question—like the pre-SMA—*are* in fact subject to attentional modulation. Thus, denying that representations in the pre-SMA are a genuine part of working memory would simply be an *ad hoc* move to preserve Carruthers' view. Second, there is reason to think that non-sensory representations in the frontal cortex may actually be particularly *well-poised* to be globally broadcast in working memory. The global broadcasting theory of consciousness (which Carruthers endorses) holds that the connective hubs that form the basis of the global broadcasting network "are particularly predominant in the prefrontal cortex" (Dehaene 2014, p. 177; see also Dehaene & Naccache, 2001; Dehaene & Changeux, 2011). As such, non-sensory representations in the frontal cortex are in fact in an ideal position to be globally broadcast, since they are situated within (or at least adjacent to) a central node of the global broadcasting system.

4.3.2 Neuroimaging data

Moving on, there is also some evidence for non-sensory working memory in humans, coming from an fMRI study by Lee, Kravitz and Baker (2013). Lee and colleagues had participants perform a working memory task that involved remembering either imagistic or categorical information about a visually-presented object. On each trial, participants were initially shown an image of a to-be-remembered object, such as a

clock, watch, motorcycle, or scooter.[32] This was followed by a brief three-second delay and a memory test, which could proceed in one of two ways. One test condition (the "imagistic condition") required participants to "indicate whether an object fragment presented after the delay belonged to the cued object or not" (Lee & Baker, 2016, p. 6). Another test condition (the "categorical condition") instead required participants to "indicate whether a whole object presented after the delay was from the same subcategory or not" (Lee & Baker, 2016, p. 7). For instance, if the remembered object had been a clock, participants might be shown an image of a *different* clock and asked whether it belonged to the same category CLOCK. The main purpose of this manipulation was to pull apart imagistic and categorical forms of working memory: in the imagistic condition participants had to retain fine-grained visual details about the remembered object, whereas in the categorical condition participants instead had to retain the abstract category of the remembered object (Lee et al., 2013, p. 997).

Lee and colleagues then examined the fMRI data using multivariate pattern analysis to see if the two types of tasks—imagistic and categorical—recruited different brain regions. Fascinatingly, they found a dissociation between the brain areas involved. In the imagistic condition, stimulus-specific patterns of activity were observed in extrastriate cortex (specifically, the posterior fusiform cortex); conversely, in the

---

[32] Participants were also shown an image of another irrelevant object to establish a baseline level of activity for non-remembered objects.

categorical condition, stimulus-specific patterns of activity were identified in the dorsal prefrontal cortex (see Lee et al., 2013, p. 997). Notably, there was no overlap across the two tasks. The prefrontal cortex exhibited stimulus-specific activity *only* in the categorical task, and extrastriate areas exhibited stimulus-specific activity *only* in the visual task. These results thus indicate that working memory differentially encodes visual and abstract information, with the visual cortex mediating visual working memory, and the prefrontal cortex mediating abstract, categorical working memory (Lee et al., 2013, p. 998).

Importantly, Lee and colleagues' findings are in conflict with the predictions made by Carruthers' sensory-based account. According to Carruthers' account, working memory necessarily depends on activity in mid-level sensory areas. As such, his account predicts that we should find evidence for relevant mid-level sensory representations across *all* working memory tasks, including tasks are typically thought to be conceptual in nature (Carruthers, 2015, p. 92). However, this is not what Lee and colleagues found. Recall that in the categorical condition, information about the remembered stimulus was *only* present in the dorsal prefrontal cortex— *not* in extrastriate sensory areas. This suggests that the observed prefrontal activity is not merely a source of top-down control but rather serves as a *genuine vehicle for the contents of working memory*. All in all, then, the study by Lee and colleagues (2013) provides further support for the existence of non-sensory working memory representations in frontal areas of the brain.

4.3.3 Aphantasia

Finally, the existence of non-sensory working memory is also motivated by recent work on individuals with *aphantasia*. Aphantasia is a newly discovered neuropsychological condition characterized by a lack of visual mental imagery (Zeman et al., 2010; Jacobs, Schwarzkopf & Silvanto, 2018; Keogh & Pearson, 2018). Individuals with aphantasia are capable of visually perceiving the world, but they lack the ability to form stimulus-independent mental images (as indexed by mental imagery questionnaires). Importantly, aphantasia can be either congenital or acquired. In some cases, individuals report a complete lack of mental imagery since birth (Jacobs et al., 2018), whereas in other cases, individuals report a loss of mental imagery following some sort of neurological event (Bartolomeo, 2008; Zeman et al., 2010). Aphantasia is important in the present context because it affords us with a chance to test the predictions made by Carruthers' sensory-based account. Carruthers holds that working memory is always grounded in mid-level sensory imagery, and thus his account seems to predict that aphantasic individuals—who *lack* such imagery in the visual domain—should perform poorly on tests of visuospatial working memory.

Interestingly, however, this prediction turns out to be false. Individuals with aphantasia often perform fairly well on visuospatial working memory tasks (Zeman et al., 2010; Jacobs et al., 2018). In one recent study, for instance, Jacobs and colleagues (2018) had an aphantasic individual ("AI") and healthy controls perform a working

memory task that involved remembering the dimensions of visually-perceived shapes. Participants were shown a shape (a diamond, parallelogram or triangle) followed by a brief mask; then, after a 4-second delay, they were shown a probe dot, and had to indicate "whether they believed the dot to be within or outside of the boundaries of the original geometric shape" (Jacobs et al., 2015, p. 65). Surprisingly, AI performed nearly as well as controls on this task, answering correctly 87% of the time (control accuracy was 90%). As Jacobs and colleagues point out, this result suggests that AI found an alternative way to perform the visuospatial task, relying on some non-visual form of memory. They go on to suggest that AI may have encoded the information about the initially-presented shape in a *spatial code*, "which represents the spatial relations between the presented visual items during the encoding stage" (Jacobs et al., 2018 p. 71). On this interpretation, AI rapidly transferred information about the perceived shape into a non-sensory spatial format, and then held this spatial representation in memory over the delay. (A similar example of preserved working memory performance in an aphantasic individual is documented by Zeman et al., 2010.)

The evidence concerning aphantasia is admittedly preliminary, and thus should be taken with a grain of salt. Nevertheless, the fact that aphantasics can compensate for their lack of mental imagery dovetails nicely with the other results discussed above. We have already seen that frontal regions are capable of encoding abstract representations of number (Nieder, 2012), frequency (Vergara et al., 2016) and object category (Lee et al.,

2013). As such, it makes sense that individuals who lack mental imagery would be able to develop other (non-imagistic) ways to store information, involving representations couched in a non-sensory format.

4.4 The positive picture

In the previous section, I outlined a variety of empirical findings that support the existence of non-sensory forms of working memory. These studies, I argued, challenge Carruthers' view that working memory is always sensory-based. But where does this leave us? If we reject Carruthers' sensory-based account, what should we replace it with? I this section, I briefly sketch an alternative model. Following Christophel and colleagues (2017) and Lee and Baker (2016), I suggest that working memory should be conceived of as a distributed process that can recruit both sensory and non-sensory forms of representation. Additionally, I also defend this model against two final objections raised by Carruthers.

4.4.1 A multi-level model of working memory storage

In contrast to Carruthers' sensory based account, I suggest that we should opt for a *flexible* conception of working memory that allows for storage at various different "levels of abstraction" (Christophel et al., 2017, p. 111). On this "multi-level" model, no stage of processing (such as the "intermediate level") is privileged with respect to working memory. Instead, working memory operates in a dynamic fashion, recruiting representations from various different brain regions (sensory and non-sensory)

"depending on the nature of the information maintained" (Lee & Baker, 2016, p. 6; Christophel et al., 2017; Pearson & Keogh, 2019). Thus, while fine-grained sensory representations will be stored in modality-specific sensory processing areas, more abstract representations (such as representations of numerosity or category membership) will be stored in more anterior amodal association areas. Importantly, on this view, we can still retain the basic idea that working memory representations are encoded and maintained via top-down attention. We just have to acknowledge that attention can target *both* sensory representations areas *and* non-sensory representations (as was argued for in section 4.2).

This multi-level model of working memory storage is gaining increasing support in the literature. Indeed, many researchers now acknowledge that working memory can involve various different types of representation. In a recent landmark review of the working memory literature, Christophel and colleagues (2017) report that:

> There is no evidence for a single site of working memory storage. Rather persistent neuronal activity that is informative about a currently memorized stimulus can be found in sensory, parietal, and prefrontal brain regions. Working memory entails a gradient of abstraction from sensory areas reflecting low-level sensory features to prefrontal regions encoding more abstract, semantic and response-related aspects of stimuli. (p. 111).

A similar view is also endorsed by Lee and Baker (2016):

> We suggest that the ability to maintain representations during working memory is a general property of cortex, not restricted to specific areas . . . [T]he common finding of maintained information in visual, but not parietal or prefrontal, cortex may be more of a reflection of the need to maintain specific types of visual information and not of a privileged role of visual cortex in maintenance. (p. 1).

It should be noted that, in endorsing a multi-level model of working memory storage, I am not making any claims about the *relative frequency* with which we deploy non-sensory representations. For all I have said, Carruthers may be right that we *often* rely on mid-level sensory representations to perform working memory tasks. I am only committed to the view that working memory *sometimes* contains non-sensory representations. In fact, it may even be the case that people *vary* in the extent to which they employ sensory/non-sensory forms of working memory. Pearson and Keogh (2019) have recently made such a suggestion: they propose that people may use different working memory strategies depending on their "cognitive preferences and profiles" (p. 3). Pearson and Keogh provide a helpful example to illustrate their point (see Pearson & Keogh, 2019, Figure 2b). Suppose that participants are tasked with remembering the shape and color of three visually-presented objects—a red hexagon, a yellow star and a blue circle. Some participants may indeed complete this task by employing an imagistic strategy, maintaining detailed visual representations of the three colored shapes. However, say Pearson and Keough, other participants may prefer to employ a non-sensory strategy, maintaining the conceptual representations RED HEXAGON, YELLOW STAR, and BLUE CIRCLE. Although Pearson and Keough's "cognitive strategies" approach remains somewhat speculative, is an interesting idea that merits further empirical research.

4.4.2 Two final objections

In his 2015 book, *The Centered Mind*, Carruthers spends some time considering the kind of multi-level model that I have just proposed. As you might expect, however, Carruthers ends up rejecting the multi-level model (which calls the "multiple-modes model"), arguing that falls prey to two empirically-based objections. In the last portion of this chapter, I outline and respond to these two final objections.

4.4.2.1 The self-knowledge objection

Carruthers' first objection to the multi-level model has to do with self-knowledge. He begins by claiming that if non-sensory representations *could* enter working memory, then it should be "trivially easy" for us to gain immediate introspective knowledge of our own propositional attitudes—all we would have to do is token the relevant attitudes in working memory (Carruthers, 2015, p. 113). According to Carruthers, however, there is evidence that we *lack* this kind of introspective access to our attitudes. He points to a number of empirical studies, which appear to show that we often *infer* our own attitudes on the basis of sensory/behavioral cues and context.

One such strand of evidence comes from studies employing the counter-attitudinal essay task, wherein participants are tasked with writing an essay supporting a view contrary to their own beliefs (Festinger, 1957; Elliot & Devine, 1994). These studies find that there is often a change in participants' self-reported beliefs after writing counter-attitudinal essays, with participants subsequently endorsing a position that is

more in line with the one they argued for in the essay. For instance, a student who wrote an essay in favor of tuition hikes might subsequently report *actually believing* that tuition hikes would be a good idea (despite having denied so in the past) (Carruthers, 2017a, p. 234). According to Carruthers, this result suggests that participants are using their own behavior (in this case, the act of writing an essay) to inform their self-ascriptions of belief. After all, if participants were capable introspecting their beliefs *directly*, says Carruthers, we wouldn't expect the mere act of writing an essay to produce such major changes in their attitude reports (Carruthers, 2015, p. 116). Carruthers (2017a) also appeals to a study by Wells and Petty (1980), who had participants listen to a radio broadcast while either nodding or shaking their heads. Wells and Petty found that participants were more likely to agree with the message of the broadcast if they were nodding their heads, rather than shaking them (Wells & Petty, 1980, p. 226). Again, Carruthers claims that the best explanation for this finding is that we use interpretive strategies to infer our own attitudes. The fact that people gauge their level of agreement based on their overt head movement suggests that they must rely on behavioral/sensory cues to know what they are thinking.

I have no quibble with the empirical results that Carruthers discusses. However, I do not think they show what Carruthers wants them to show—namely, that we are *incapable* of introspecting our own attitudes. This response is clearly articulated by Georges Rey (2013). Rey points out that the above findings do not rule out the possibility

that we use a mixture of *both* sensory/behavioral/circumstantial cues (what he calls "SBC cues") *and* introspection to ascribe attitudes to ourselves. While Carruthers' data may show that we *occasionally* privilege SBC cues over introspection, it remains possible that we have introspective access to our attitudes in many other contexts. To quote Rey (2013) directly:

> All that's essential to Introspectionism is that it's *possible* for people *sometimes* to attribute attitudes to themselves in a way that involves more than SBC data, even if, contrary to one's impression, that pure ability may be deployed haphazardly in ordinary life. Occasional failures in performance are hardly sufficient to show a lack of underlying competence. (p. 267)

As Rey makes clear, then, Carruthers' data fail to refute the multi-level model of working memory. Even if we grant that people occasionally use SBC cues to ascribe attitudes to themselves (as the above studies suggest), this remains compatible with the view that attitudes are also frequently introspectable, and can gain direct access to working memory.

Carruthers himself complains that Rey's "mixed-methods" proposal is *ad hoc* and fails to explain why people introspect their attitudes in some contexts but not others (Carruthers, 2015, p. 116). However, I think this characterization is unfair. Consider, by analogy, the case of sensory representations: sensory representations clearly vary with respect to their introspectability, sometimes entering conscious awareness while other times remaining unconscious (Dehaene, Changeux, Naccache, Sackur & Sergent, 2006). Given that sensory representations are not always introspectable, it seems perfectly

reasonable to assume that the same kind of introspective variability will apply to attitude states as well. Thus, far from being *ad hoc*, Rey's proposal nicely captures the intuition that our attitudes can be more or less introspectable depending on the context. Carruthers may be right that we don't yet have a full explanation of *why* attitude-introspectability varies in the way that it does. But this isn't really a *problem* with Rey's view—it merely reflects our current lack of knowledge.

4.4.2.2 The missing variance objection

Carruthers second objection to the multi-level model concerns the link between working memory and general fluid intelligence (or "fluid *g*", for short).[33] Carruthers starts by noting that there appears to be an observable correlation between working memory and fluid *g*, such that higher scores on tests of working memory correspond with higher performance on tests of fluid *g* (see, e.g., Conway, Kane and Engle, 2003; Colom, Rebollo, Palacios, Juan-Espinosa & Kyllonen, 2012). Indeed, this correlation appears to be particularly robust: estimates indicate that the correlation between working memory ability and fluid *g* is somewhere "between 0.6 and 0.9" (Carruthers, 2015, p. 129).

---

[33] Psychologists typically distinguish two aspects of general intelligence: *crystalized* general intelligence and *fluid* general intelligence. Crystalized intelligence refers to one's overall degree of factual knowledge, whereas fluid intelligence referrers to one's "ability to reason and to solve new problems independently of previously acquired knowledge" (Jaeggi, Buschkuehl, Jonides & Perrig, 2008).

Carruthers goes on to claim that "if there were an amodal global workspace for reflection in addition to a sensory-based working-memory system, then this should show up as an independent source of variance in people's intellectual performance" (Carruthers, 2015, p. 127). In other words, if non-sensory (i.e., amodal) working memory exists, then part of the variability in peoples' fluid *g* should be attributable to differences in the functioning of the non-sensory component of working memory. According to Carruthers, this prediction is testable: we can contrast the correlations between (i) purely sensory working memory tasks and fluid *g*, and (ii) more abstract working memory tasks and fluid *g*. If non-sensory working memory exists, then performance on the abstract tasks should have a *stronger* correlation with fluid *g*, since such tasks would capture the variance in *both* the sensory and non-sensory components of working memory (see Carruthers, 2015, p. 130). Conversely, the correlation between performance on purely sensory tasks should be weaker because "tests of purely sensory working memory would fail to include any measure of the variance in amodal thinking abilities that would (by hypothesis) account for a large proportion of our flexible general intelligence" (Carruthers, 2017a, p. 241).

Carruthers argues that the above prediction is false, however. He points to five studies that have examined the correlation between working memory and fluid *g* (Unsworth & Spillers, 2010; Burgess, Gray, Conway & Braver, 2011; Redick, Unsworth, Kelly & Engle, 2012; Shipstead, Redick, Hicks & Engle, 2012; Shipstead, Lindsey,

Marshall & Engle, 2014). These studies all included a "symmetry-span task", which Carruthers' takes to be a measure of *sensory* working memory. In the symmetry-span task, participants had to remember the positions of a number of sequentially-presented colored squares, while also assessing the symmetry of interspersed pixel patterns (Figure 12). Additionally, the studies also included other "operation span tasks", which Carruthers takes to be more abstract in nature. In these tasks, participants had to "recall lists of words, letters, or numbers while [simultaneously] undertaking a secondary mathematical task" (Carruthers, 2015, p. 130). The crucial finding in these studies—for our purposes—was that performance on the symmetry-span task correlated with fluid *g* *just as highly* as performance on the seemingly more abstract tasks. According to Carruthers, this finding puts pressure on the notion of non-sensory working memory. The fact that the symmetry-span task is an equally-good predictor of fluid *g* (when compared to more abstract tasks) suggests that it is *sensory-based* working memory that is driving the correlation between working memory and fluid *g* (Carruthers, 2015, p. 131). Non-sensory working memory appears to make no contribution to fluid *g*, and thus (allegedly) doesn't exist.

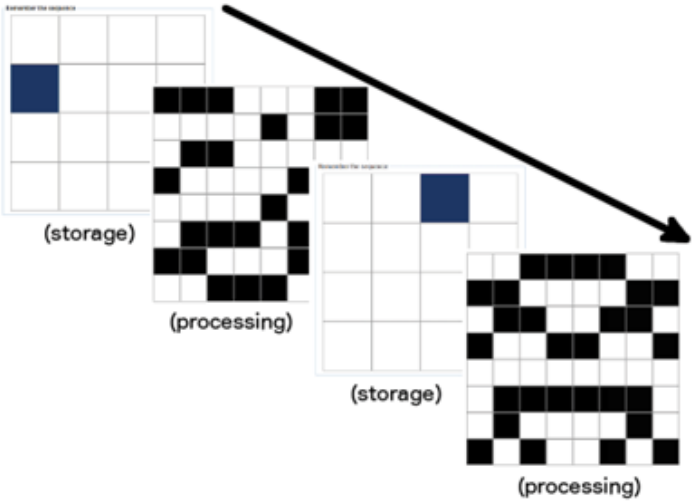

Figure 12: A depiction of the symmetry-span task. Participants are required to remember the position of the sequentially-presented blue squares, while also assessing the symmetry of interspersed pixel patterns (symmetrical vs nonsymmetrical). At the end of the trial (not shown), participants then have to report the positions of the remembered squares. From Stone & Towse (2015, p 4). Copyright © 2007, Stone & Towse. Reprinted under the terms of the Creative Commons Attribution License (CC BY).

There are at least two problems with this objection, however. First of all, the "sensory" task that Carruthers appeals to—the symmetry span task—is not unambiguously sensory in nature. As mentioned earlier in the context of aphantasia, information about spatial position (such as the location of blocks in a grid) could hypothetically be stored in an abstract spatial code, perhaps similar to a Cartesian coordinate system. Looking at figure 12, for instance, it is possible that participants remembered the location of the blue squares by retaining a set of abstract coordinates—e.g., [(1,3), (3,4)]—instead of a detailed visual image. For this reason, we can't assume that the symmetry span task is tapping into *sensory* working memory, rather than some

other form of non-sensory spatial working memory. This renders Carruthers' argument inconclusive, however. The above studies fail to show that sensory working memory tasks and abstract working memory tasks are equally correlated with fluid *g*, because *they do not include a pure test of sensory working memory*.[34]

Second, the prediction that Carruthers' attributes to the multi-level model—that there should be "decreasing overlap between working memory and fluid *g* when one moves from tasks that are more abstract in nature to those that are purely sensory" (Carruthers, 2015, p. 130)—is also questionable. It is widely believed that the correlation between working memory and fluid *g* is mediated by *attention*. That is, working memory and fluid *g* are related because they both rely on the same underlying attentional-control mechanisms (Conway et al., 2003). If this is the case, however, we might not expect to find much of a difference in the level of correlation (with fluid *g*) across sensory and abstract working memory tasks, since both types of tasks make use of similar attentional circuitry. Carruthers briefly considers this criticism. He insists, however, that sensory and non-sensory forms of attention "would surely involve distinct neural pathways",

---

[34] In a footnote, Carruthers suggests that "[e]ven if symmetry-span tests are not fully nonconceptual, they are certainly *less* conceptual in nature than the operation-span, reading-span, and memory-span tests, which use words, letters, or numbers" (Carruthers, 2015, p. 131). I find this remark somewhat puzzling, as it is not immediately clear what it means for representations to be *more* or *less* conceptual. Additionally, the fact that a representation is non-conceptual (or "less" conceptual) does not necessarily tell us anything about its status regarding the sensory/non-sensory distinction. There may well be non-conceptual forms of amodal representation in the brain (such as analog magnitude representations (Brannon, 2006)), and there may also be sensory-laden concepts (Prinz, 2002).

and thus that there should still be *some* variance in fluid *g* solely driven by non-sensory working memory (if such a thing exists) (Carruthers, 2015, p. 126). This is perhaps a fair point. But if we assume that the attentional-control mechanisms involved in sensory and non-sensory working memory are *largely overlapping*, then this variance might turn out to be exceedingly small—and might therefore be difficult to detect in experimental contexts. As such, it is not surprising (even from the perspective of those who *endorse* non-sensory working memory) that sensory and non-sensory working memory often appear to be correlated with fluid *g* to the same degree.

In sum, then, neither of Carruthers' additional objections are conclusive. In both cases, there are plausible responses that can be made on behalf on the multi-level model. Given that we have independent reasons to believe in the existence of non-sensory working memory—as outlined in section 4.3—I conclude that the multi-level model is preferable to Carruthers' sensory-based account.

4.5 Conclusion

In this chapter, I argued against Peter Carruthers' sensory-based account of working memory. I began by outlining Carruthers' sensory-based account and his main argument in its favor—the targets of attention argument. This argument was shown to be unconvincing, however. Carruthers' argument relies on the premise that attention only targets mid-level sensory areas, but this premise turns out to be empirically suspect: there is in fact evidence for attentional modulations in non-sensory regions of

the cortex, such as the frontal cortex (Lau et al., 2004; Ester et al., 2016). Next, I went on to provide *positive* evidence for the existence of non-sensory working memory. As we saw, support for the existence of non-sensory working memory comes from several lines of evidence, including electrophysiological studies with monkeys (Nieder, 2012; Vergara et al., 2016), human neuroimaging studies (Lee et al., 2013) and studies of individuals with aphantasia (Jacobs et al., 2018). In light of these findings, I endorsed a more permissive account of working memory—dubbed the multi-level model—that allows for the maintenance of both sensory and non-sensory representations. Although Carruthers raises two objections against the multi-level model, both of these objections were shown to be inconclusive. Thus, the multi-level model remains a plausible alternative to Carruthers sensory-based account.

5. Conclusion

This dissertation examined the nature of working memory. I argued that working memory is best understood as a functionally-defined process, which exhibits four main characteristics: (i) information maintenance, (ii) information manipulation, (iii) capacity limitations, and (iv) distractor resistance. I also examined the relationship between working memory and other mental phenomena, including attention and consciousness. As we saw, working memory is indeed *related* to attention and consciousness, but it should not be identified with either one. Empirical and phenomenological data indicate that working memory representations can be maintained—at least temporarily—in the absence of both attention and conscious experience. Finally, I addressed the question of whether working memory is always *sensory-based*, as proposed by Peter Carruthers (2015). I argued that Carruthers' sensory-based view is incorrect: recent neuroscientific evidence indicates that working memory is not restricted to sensory-based representations, but rather can include *both* sensory and non-sensory forms of representation.

Appendix A

One view that deserves further attention—but which does not easily fit into the narrative of the third chapter—is Jesse Prinz's Attended Intermediate Level (AIR) theory of consciousness. Prinz's AIR theory of consciousness assigns a central role to working memory, however his understanding of the relationship between working memory and consciousness is quite different from my own. Specifically, Prinz's theory holds that:

> "Consciousness arises when and only when intermediate-level representations undergo changes that allow them to become *available* to working memory" (Prinz, 2012, p. 97; emphasis added)

The important aspect of Prinz's theory, for our purposes, concerns his claim that consciousness coincides with the *availability* of information to working memory. This seems to flip my view on its head: I have argued that working memory representations can remain accessible to consciousness without thereby being experienced, whereas Prinz's seems to hold that information is conscious solely in virtue of being made *available* to working memory in the first place. What is going on here? How did Prinz and I arrive and such radically different conclusions, given that we are both working with the same empirical data?

Part of the confusion, I think, has to do with Prinz's somewhat atypical use of the term "working memory". Specifically, Prinz seems to use the term "working memory" to refer exclusively to the executive component of the working memory system. He tells us that "it is entirely wrong to think about working memory as storing representations;

rather it works by sustaining activity in perceptual centers" (Prinz, 2012, p. 101-102). Elsewhere, he explicitly states that, "working memory is not a storehouse but a collection of 'executive' processes" (2012, p. 321). Prinz also takes the neural correlates of working memory to be predominantly frontally-based, including specifically the lateral prefrontal cortex (see Prinz, 2012, p. 98). This restrictive interpretation is at odds with how I (and most other researchers) conceptualize working memory. While I agree with Prinz that working memory often works by sustaining representations in posterior perceptual areas, I take this to show, not that working memory is a purely executive process, but rather that the neural basis of working memory *includes* storage-related activity in perceptual areas (see, e.g., Sreenivasan, Curtis & D'Esposito, 2014).

Once this clarification is made, it is easier to see how Prinz's view and my own relate to one another. In saying that representations are conscious when they become available to working memory, what Prinz *really means* is that representations reach consciousness when they are "broadcast" (or directly poised to be broadcast) to frontal executive centers (Prinz, 2012, p. 99). I am happy to accept this as a possible account of consciousness. However, on my view, working memory is not restricted to what is currently attended and immediately broadcastable. Rather, there are also working memory representations that are "accessible" in a less immediate sense, which *could* be broadcast to frontal areas if attention were turned towards them. Prinz might protest that we shouldn't call these representations "working memory representations", but this

a purely terminological quibble. And, in any case, I am in good company. A growing number of scientists recognize the existence of both (i) currently attended-and-conscious working memory representations, and (ii) unconscious-but-accessible representations outside the focus of attention (see Gilchrist & Cowan, 2010; LaRocque, Lewis-Peacock & Postle, 2014; Sergent, 2018).

References

Aizawa, K., & Gillett, C. (2009). The (multiple) realization of psychological and other properties in the sciences. *Mind & Language*, *24*(2), 181-208.

Albers, A.M., Kok, P., Toni, I., Dijkerman, H.C., & de Lange, F.P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology*, *23*(15), 1427-1431.

Alvarez, G.A., & Franconeri, S.L. (2007). How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision*, *7*(13), 1-10.

Atkinson, R.C., & Shiffrin, R.M. (1968). Human memory: A proposed system and its control processes. In *Psychology of learning and motivation* (Vol. 2, pp. 89-195). New York: Academic Press.

Arranz-Paraíso, S., & Serrano-Pedraza, I. (2018). Testing the link between visual suppression and intelligence. *PloS one*, *13*(7), e0200151.

Awh, E., & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends in Cognitive Sciences*, *5*(3), 119-126.

Awh, E., & Pashler, H. (2000). Evidence for split attentional foci. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 834.

Awh, E., Jonides, J., & Reuter-Lorenz, P.A. (1998). Rehearsal in spatial working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 780-790.

Awh, E., Vogel, E. K., & Oh, S.H. (2006). Interactions between attention and working memory. *Neuroscience*, *139*(1), 201-208.

Baars, B.J. (1997). *In the theater of consciousness: The workspace of the mind*. Oxford: Oxford University Press.

Baars, B.J. (2001). A biocognitive approach to the conscious core of immediate memory. *Behavioral and Brain Sciences*, *24*(1), 115-116.

Baars, B.J. (2002). The conscious access hypothesis: origins and recent evidence. *Trends in cognitive Sciences*, *6*(1), 47-52.

Baars, B.J. (2003). Working memory requires conscious processes, not vice versa. In N. Osaka (Ed.), *Neural basis of consciousness* (pp. 11-26). Amsterdam: John Benjamins Publishing.

Baars, B.J., & Franklin, S. (2003). How conscious experience and working memory interact. *Trends in Cognitive Sciences*, *7*(4), 166-172.

Baddeley, A.D. (1992). Working memory: The interface between memory and cognition. *Journal of Cognitive Neuroscience*, *4*(3), 281-288.

Baddeley, A.D. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829-839.

Baddeley, A.D. (2007). *Working memory, thought, and action*. Oxford: Oxford University Press.

Baddeley, A.D. (2010). Working memory. *Current Biology*, *20*(4), R136-R140.

Baddeley, A.D. (2014). *Essentials of human memory (classic edition)*. New York: Psychology Press.

Baddeley, A.D., & Hitch, G. (1974). Working memory. In G.A. Bower (Ed.), *Psychology of learning and motivation* (Vol. 8, pp. 47-89). New York: Academic Press.

Baddeley, A.D., Grant, S., Wight, E., & Thomson, N. (1975). Imagery and visual working memory. In P.M.A. Rabbitt & S. Dornic (Eds.) *Attention and performance V* (pp. 205-217. London: Academic Press.

Bartolomeo, P. (2008). The neural correlates of visual mental imagery: an ongoing debate. *Cortex*, *44*(2), 107-108.

Bayne, T., & Montague, M. (Eds.). (2011). *Cognitive phenomenology*. Oxford: Oxford University Press.

Beninger, M. (*forthcomming*). The case for unconscious working memory. *Philosophical Psycholohy*

Bird, A. (2018). The metaphysics of natural kinds. *Synthese*, *195*(4), 1397-1426.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227-247.

Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences*, *9*(2), 46-52.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, *30*(5-6), 481-499.

Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences*, *15*(12), 567-575.

Bor, D., & Seth, A.K. (2012). Consciousness and the prefrontal parietal network: insights from attention, working memory, and chunking. *Frontiers in Psychology*, *3*, 63.

Brady, T.F., Konkle, T., & Alvarez, G.A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5), 1-34.

Brannon, E. M. (2006). The representation of numerical magnitude. *Current Opinion in Neurobiology*, *16*(2), 222-229.

Brown, R. (2014). Consciousness doesn't overflow cognition. *Frontiers in Psychology*, *5*, 1399.

Buehler, D. (2018). The central executive system. *Synthese*, *195*(5), 1969-1991.

Burgess, G.C., Gray, J.R., Conway, A.R., & Braver, T.S. (2011). Neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span. *Journal of Experimental Psychology: General*, *140*(4), 674-692.

Camina, E., & Güell, F. (2017). The neuroanatomical, neurophysiological and psychological basis of memory: Current models and their origins. *Frontiers in Pharmacology*, *8*, 438.

Camos, V., Johnson, M., Loaiza, V., Portrat, S., Souza, A., & Vergauwe, E. (2018). What is attentional refreshing in working memory?. *Annals of the New York Academy of Sciences*, *1424*, 19–32.

Carruthers, P. (2013). Evolution of working memory. *Proceedings of the National Academy of Sciences*, *110*(Supplement 2), 10371-10378.

Carruthers, P. (2014). On central cognition. *Philosophical Studies*, *170*(1), 143-162.

Carruthers, P. (2015). *The centered mind: What the science of working memory shows us about the nature of human thought*. Oxford: Oxford University Press.

Carruthers, P. (2017a). The illusion of conscious thought. *Journal of Consciousness Studies*, *24*(9-10), 228-252.

Carruthers, P. (2017b). Block's overflow argument. *Pacific Philosophical Quarterly*, *98*, 65-70.

Chao, L.L., & Knight, R.T. (1998). Contribution of human prefrontal cortex to delay performance. *Journal of Cognitive Neuroscience*, *10*(2), 167-177.

Christophel, T.B., Iamshchinina, P., Yan, C., Allefeld, C., & Haynes, J.D. (2018). Cortical specialization for attended versus unattended working memory. *Nature Neuroscience*, *21*(4), 494-496.

Christophel, T.B., Klink, P.C., Spitzer, B., Roelfsema, P.R., & Haynes, J.D. (2017). The distributed nature of working memory. *Trends in Cognitive Sciences*, *21*(2), 111-124.

Chudnoff, E. (20016). Review of The Centered Mind: What the Science of Working Memory Shows Us about the Nature of Human Thought, by Peter Carruthers. *Notre Dame Philosophical Reviews*, 2016.07.17. URL = https://ndpr.nd.edu/news/the-centered-mind-what-the-science-of-working-memory-shows-us-about-the-nature-of-human-thought/

Chun, M.M. (2011). Visual working memory as visual attention sustained internally over time. *Neuropsychologia*, *49*(6), 1407-1409.

Chun, M.M., Golomb, J.D., & Turk-Browne, N.B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*, 73-101.

Cohen, M.A., & Dennett, D.C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*(8), 358-364.

Cohen, M.A., Cavanagh, P., Chun, M.M., & Nakayama, K. (2012). The attentional requirements of consciousness. *Trends in Cognitive Sciences*, *16*(8), 411-417.

Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P.C. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence*, *32*(3), 277-296.

Conway, A.R., Kane, M.J., & Engle, R.W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547-552.

Corbetta, M., & Shulman, G.L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201-215.

Cowan, N. (1999). An embedded-processes model of working memory. In A Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62-101). New York: Cambridge University Press.

Cowan, N. (2008a). What are the differences between long-term, short-term, and working memory?. *Progress in Brain Research*, *169*, 323-338.

Cowan, N. (2008b). Sensory memory. In H.L. Roediger (Ed.), *Cognitive psychology of memory, Vol. 2 of Learning and memory: A comprehensive reference* (pp. 23-32). Oxford: Elsevier.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why?. *Current Directions in Psychological Science*, *19*(1), 51-57.

Cowan, N. (2011). The focus of attention as observed in visual working memory tasks: Making sense of competing claims. *Neuropsychologia*, *49*(6), 1401-1406.

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*(4), 1158-1170.

Cowan, N., Elliott, E.M., Saults, J.S., Morey, C.C., Mattox, S., Hismjatullina, A., & Conway, A.R. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*(1), 42-100.

Craver, C.F. (2004). Dissociable realization and kind splitting. *Philosophy of Science*, *71*(5), 960-971.

Curtis, C.E. (2006). Prefrontal and parietal contributions to spatial working memory. *Neuroscience*, *139*(1), 173-180.

Curtis, C.E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*(9), 415-423.

D'Esposito, M. (2007). From cognitive to neural models of working memory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 761-772.

D'Esposito, M., & Postle, B.R. (1999). The dependence of span and delayed-response performance on prefrontal cortex. *Neuropsychologia*, *37*(11), 1303-1315.

D'Esposito, M., & Postle, B.R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology*, *66*, 115-142.

Dade, L., Zatorre, R., Evans, A., & Jones-Gottman, M. (2001). Working memory in another dimension: functional imaging of human olfactory working memory. *Neuroimage*, *14*(3), 650-660.

De Brigard, F. (2012). The role of attention in conscious recollection. *Frontiers in Psychology*, *3*, 29.

De Gardelle, V., Sackur, J., & Kouider, S. (2009). Perceptual illusions in brief visual presentations. *Consciousness and Cognition*, *18*(3), 569-577.

Dehaene, S. (2014) *Consciousness and the Brain*. New York: Viking Press.

Dehaene, S., & Changeux, J.P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200-227.

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

Dehaene, S., Changeux, J.P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, *10*(5), 204-211.

Dehaene, S., Naccache, L., Cohen, L., Bihan, D.L., Mangin, J.F., Poline, J.B., & Rivière, D. (2001) Cerebral mechanisms of word masking and unconscious repetition priming, *Nature Neuroscience*. *4*(7), 752-758.

Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition*, *110*(3), 412-431.

Driver, J., & Mattingley, J.B. (1998). Parietal neglect and visual awareness. *Nature Neuroscience*, *1*(1), 17-22.

Elliot, A.J., & Devine, P.G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology*, *67*(3), 382-394.

Eriksen, C.W., & James, J.D.S. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics*, *40*(4), 225-240.

Ester, E.F., Sutterer, D. W., Serences, J.T., & Awh, E. (2016). Feature-selective attentional modulations in human frontoparietal cortex. *Journal of Neuroscience*, *36*(31), 8188-8199.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Fine, D. R. (2012). A life with prosopagnosia. *Cognitive Neuropsychology*, *29*(5-6), 354-359.

Flanagan, O.J. (1992). *Consciousness reconsidered*. Cambridge, MA: MIT Press.

Fortney, M. (2018). The Centre and Periphery of Conscious Thought. *Journal of Consciousness Studies*, *25*(3-4), 112-136.

Fougnie, D. (2008). The relationship between attention and working memory. In N. B. Johansen (Ed.) *New research on short-term memory* (pp. 1-45). New York: Nova Science.

Fougnie, D., & Marois, R. (2007). Executive working memory load induces inattentional blindness. *Psychonomic Bulletin & Review*, *14*(1), 142-147.

Funahashi, S. (2001). Neuronal mechanisms of executive control by the prefrontal cortex. *Neuroscience Research*, *39*(2), 147-165.

Funahashi, S., & Kubota, K. (1994). Working memory and prefrontal cortex. *Neuroscience Research*, *21*(1), 1-11.

Funahashi, S., Bruce, C.J., & Goldman-Rakic, P.S. (1989). Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of neurophysiology*, *61*(2), 331-349.

Fuster, J.M., & Alexander, G.E. (1971). Neuron activity related to short-term memory. *Science*, *173*(3997), 652-654.

Garavan, H. (1998). Serial attention within working memory. *Memory & Cognition*, *26*(2), 263-276.

Gazzaley, A., & Nobre, A.C. (2012). Top-down modulation: bridging selective attention and working memory. *Trends in Cognitive Sciences*, *16*(2), 129-135.

Giesbrecht, B., Woldorff, M.G., Song, A.W., & Mangun, G.R. (2003). Neural mechanisms of top-down control during spatial and feature attention. *Neuroimage*, *19*(3), 496-512.

Gilchrist, A.L., & Cowan, N. (2010). Conscious and unconscious aspects of working memory. In I. Czigler & I. Winkle (Eds.), *Unconscious Memory Representations in Perception: Processes and Mechanisms in the Brain* (pp. 1-36). Amsterdam: John Benjamins.

Goldman-Rakic, P.S. (1992). Working memory and the mind. *Scientific American*, *267*(3), 110-117.

Goldman-Rakic, P.S. (1995). Cellular basis of working memory. *Neuron*, *14*(3), 477-485.

Gomez-Lavin, J. (2017). The centered mind: What the science of working memory shows us about the nature of human thought. *Philosophical Psychology*, *30*(5), 685-688.

Gomez-Lavin, J. (*In preparation*). Working memory is not a natural kind.

Harris, J.A., Miniussi, C., Harris, I.M., & Diamond, M.E. (2002). Transient storage of a tactile memory trace in primary somatosensory cortex. *Journal of Neuroscience*, *22*(19), 8720-8725.

Harrison, S.A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632-635.

Hitch, G.J., Hu, Y., Allen, R.J., & Baddeley, A.D. (2018). Competition for the focus of attention in visual working memory: perceptual recency versus executive control. *Annals of the New York Academy of Sciences*, *1424*, 64–75.

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

Jacobs, C., Schwarzkopf, D.S., & Silvanto, J. (2018). Visual working memory performance in aphantasia. *cortex*, *105*, 61-73.

Jaeggi, S.M., Buschkuehl, M., Jonides, J., & Perrig, W.J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*(19), 6829-6833.

James, W. (1890/1983). *The principles of psychology*. Cambridge, MA: Harvard University Press.

Jantz, T.K., Tomory, J.J., Merrick, C., Cooper, S., Gazzaley, A., & Morsella, E. (2014). Subjective aspects of working memory performance: Memoranda-related imagery. *Consciousness and Cognition*, *25*, 88-100.

Kemmerer, D. (2015). Are we ever aware of concepts? A critical question for the Global Neuronal Workspace, Integrated Information, and Attended Intermediate-Level Representation theories of consciousness. *Neuroscience of Consciousness*, *2015*(1), 1-10.

Kentridge, R.W., Nijboer, T.C., & Heywood, C.A. (2008). Attended but unseen: Visual attention is not sufficient for visual awareness. *Neuropsychologia*, *46*(3), 864-869.

Keogh, R., & Pearson, J. (2018). The blind mind: No sensory visual imagery in aphantasia. *Cortex*, *105*, 53-60.

Kintsch, W., Healy, A.F., Hegarty, M., Pennington, B.F., & Salthouse, T.A. (1999). Models of working memory: Eight questions and some general issues. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 412-441). New York: Cambridge University Press.

Kiyonaga, A., & Egner, T. (2013). Working memory as internal attention: toward an integrative account of internal and external selection processes. *Psychonomic Bulletin & Review*, *20*(2), 228-242.

Koivisto, M., Ruohola, M., Vahtera, A., Lehmusvuo, T., & Intaite, M. (2018). The effects of working memory load on visual awareness and its electrophysiological correlates. *Neuropsychologia*, *120*, 86-96.

Kouider, S., de Gardelle, V., Sackur, J., & Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends in Cognitive Sciences*, *14*(7), 301-307.

Kumar, S., Joseph, S., Gander, P.E., Barascud, N., Halpern, A.R., & Griffiths, T.D. (2016). A brain system for auditory working memory. *Journal of Neuroscience*, *36*(16), 4492-4505.

Kuo, B.C., Stokes, M.G., & Nobre, A.C. (2012). Attention modulates maintenance of representations in visual short-term memory. *Journal of Cognitive Neuroscience*, *24*(1), 51-60.

Lara, A.H., & Wallis, J.D. (2015). The role of prefrontal cortex in working memory: a mini review. *Frontiers in Systems Neuroscience*, *9*, 173.

LaRocque, J.J., Eichenbaum, A.S., Starrett, M.J., Rose, N.S., Emrich, S.M., & Postle, B.R. (2015). The short-and long-term fates of memory items retained outside the focus of attention. *Memory & Cognition*, *43*(3), 453-468.

LaRocque, J.J., Lewis-Peacock, J.A., & Postle, B.R. (2014). Multiple neural states of representation in short-term memory? It's a matter of attention. *Frontiers in Human Neuroscience*, 8(5), 1-14.

LaRocque, J.J., Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K., & Postle, B.R. (2013). Decoding attended information in short-term memory: an EEG study. *Journal of Cognitive Neuroscience*, *25*(1), 127-142.

LaRocque, J.J., Riggall, A.C., Emrich, S.M., & Postle, B.R. (2016). Within-category decoding of information in different attentional states in short-term memory. *Cerebral Cortex*, *27*(10), 4881-4890.

Lau, H.C., Rogers, R.D., Haggard, P., & Passingham, R.E. (2004). Attention to intention. *Science*, *303*(5661), 1208-1210.

Lee, S.H., & Baker, C.I. (2016). Multi-voxel decoding and the topography of maintained information during visual working memory. *Frontiers in Systems Neuroscience*, 10, 2.

Lee, S.H., Kravitz, D.J., & Baker, C.I. (2013). Goal-dependent dissociation of visual and prefrontal cortices during working memory. *Nature Neuroscience*, *16*(8), 997-999.

Lepsien, J., & Nobre, A.C. (2006a). Cognitive control of attention in the human brain: Insights from orienting attention to mental representations. *Brain Research*, *1105*(1), 20-31.

Lepsien, J., & Nobre, A.C. (2006b). Attentional modulation of object representations in working memory. *Cerebral Cortex*, *17*(9), 2072-2083.

Lepsien, J., Thornton, I., & Nobre, A.C. (2011). Modulation of working-memory maintenance by directed attention. *Neuropsychologia*, *49*(6), 1569-1577.

Levin, J. (2018). Functionalism. In N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2018 Edition)*. URL=<https://plato.stanford.edu/archives/fall2018/entries/functionalism/>.

Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K., & Postle, B.R. (2012). Neural evidence for a distinction between short-term memory and the focus of attention. *Journal of Cognitive Neuroscience*, *24*(1), 61-79.

Libet, B. (1999). Do we have free will?. *Journal of Consciousness Studies*, *6*(8-9), 47-57.

Logie, R.H. (1996). The seven ages of working memory. In J.T.E. Richardson, R.W. Engle, L. Hasher, R.H. Logie, E.R. Stoltzfus, & R.T. Zacks (Eds.), *Working memory and human cognition* (pp. 31-65). Oxford: Oxford University Press.

Luck, S.J., & Vogel, E.K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279-281.

Luck, S.J., & Vogel, E.K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(8), 391-400.

Lundqvist, M., Herman, P., & Miller, E.K. (2018). Working memory: delay activity, yes! persistent activity? Maybe not. *Journal of Neuroscience*, *38*(32), 7013-7019.

Lundqvist, M., Rose, J., Herman, P., Brincat, S.L., Buschman, T.J., & Miller, E.K. (2016). Gamma and beta bursts underlie working memory. *Neuron*, *90*(1), 152-164.

Ma, W.J., Husain, M., & Bays, P.M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347-356.

Majerus, S., Péters, F., Bouffier, M., Cowan, N., & Phillips, C. (2018). The dorsal attention network reflects both encoding load and top–down control during working memory. *Journal of Cognitive Neuroscience*, *30*(2), 144-159.

Marois, R. (2015). The brain mechanisms of working memory: an evolving story. In P. Jolicoeur, C. Lefebvre, & J. Martinez-Trujillo (Eds.), *Mechanisms of sensory working memory: Attention and performance XXV* (pp. 23-31). London: Academic Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: W.H. Freeman and Company.

McCarthy, G., Blamire, A.M., Puce, A., Nobre, A.C., Bloch, G., Hyder, F., ... & Shulman, R.G. (1994). Functional magnetic resonance imaging of human prefrontal cortex activation during a spatial working memory task. *Proceedings of the National Academy of Sciences*, *91*(18), 8690-8694.

McClelland, T., & Bayne, T. (2016). Concepts, contents, and consciousness. *Neuroscience of Consciousness*, *2016*(1), 1-9.

McMains, S.A., & Somers, D.C. (2004). Multiple spotlights of attentional selection in human visual cortex. *Neuron*, *42*(4), 677-686.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81-97.

Milner, B. (1966). Amnesia following operation on the temporal lobes. In O.L. Zangwill and C.W.M. Whitty (Eds.), *Amnesia* (pp. 109-133). London: Butterworth.

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. New York: Cambridge University Press.

Mole, C., Smithies, D., & Wu, W. (Eds.). (2011). *Attention: Philosophical and psychological essays*. Oxford: Oxford University Press.

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science*, *319*(5869), 1543-1546.

Moore, T., & Armstrong, K.M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, *421*(6921), 370-373.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, *11*(1), 56-60.

Müller, N.G., Bartelt, O.A., Donner, T.H., Villringer, A., & Brandt, S.A. (2003). A physiological correlate of the "zoom lens" of visual attention. *Journal of Neuroscience*, *23*(9), 3561-3565.

Myers, N.E., Stokes, M.G., & Nobre, A.C. (2017). Prioritizing information during working memory: beyond sustained internal attention. *Trends in Cognitive Sciences*, *21*(6), 449-461.

Naghavi, H.R., & Nyberg, L. (2005). Common fronto-parietal activity in attention, memory, and consciousness: shared demands on integration?. *Consciousness and Cognition*, *14*(2), 390-425.

Neander, K. (2017). *A mark of the mental: In defense of informational teleosemantics*. Cambridge, MA: MIT Press.

Nieder, A. (2012). Supramodal numerosity selectivity of neurons in primate prefrontal and posterior parietal cortices. *Proceedings of the National Academy of Sciences*, *109*(29), 11860-11865.

Nieder, A. (2016). The neuronal code for number. *Nature Reviews Neuroscience*, *17*(6), 366-382.

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

Norman, K.A., Polyn, S.M., Detre, G.J., & Haxby, J.V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424-430.

Oberauer, K. (2002). Access to information in working memory: exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 411-421.

Overgaard, M. (2018). Phenomenal consciousness and cognitive access. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170353.

Pasternak, T., Lui, L.L., & Spinelli, P.M. (2015). Unilateral prefrontal lesions impair memory-guided comparisons of contralateral visual motion. *Journal of Neuroscience*, *35*(18), 7095-7105.

Pearson, J., & Keogh, R. (2019). Redefining Visual Working Memory: A Cognitive-Strategy, Brain-Region Approach. *Current Directions in Psychological Science*, *28*(3), 266–273.

Persuh, M., LaRock, E., & Berger, J. (2018). Working memory and consciousness: The current state of play. *Frontiers in Human Neuroscience*, *12*, 78.

Petersen, S.E., & Posner, M.I. (2012). The attention system of the human brain: 20 years after. *Annual Review of Neuroscience*, *35*, 73-89.

Phillips, W.A. (1974). On the distinction between sensory storage and short-term visual memory. *Perception & Psychophysics*, *16*(2), 283-290.

Poldrack, R.A., & Yarkoni, T. (2016). From brain maps to cognitive ontologies: informatics and the search for mental structure. *Annual Review of Psychology*, *67*, 587-612.

Postle, B.R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, *139*(1), 23-38.

Postle, B.R. (2007). Activated long-term memory? The bases of representation in working memory. In N. Osaka, R.H. Logie, & M. D'Esposito, M. (Eds.), *The cognitive neuroscience of working memory* (pp. 333-349). Oxford: Oxford University Press.

Postle, B.R. (2015). The cognitive neuroscience of visual short-term memory. *Current Opinion in Behavioral Sciences*, *1*, 40-46.

Postle, B.R. (2017). Working memory functions of the prefrontal cortex. In *The prefrontal cortex as an executive, emotional, and social brain* (pp. 39-48). Tokyo: Springer Japan.

Prinz, J. (2002). *Furnishing the mind: Concepts and their perceptual basis*. Cambridge, MA: MIT Press.

Prinz, J. (2012). *The conscious brain*. Oxford: Oxford University Press.

Ptak, R. (2012). The frontoparietal attention network of the human brain: action, saliency, and a priority map of the environment. *The Neuroscientist*, *18*(5), 502-515.

Redick, T.S., Unsworth, N., Kelly, A.J., & Engle, R.W. (2012). Faster, smarter? Working memory capacity and perceptual speed in relation to fluid intelligence. *Journal of Cognitive Psychology*, *24*(7), 844-854.

Rey, G. (2013). We Are Not All 'Self-Blind': A Defense of a Modest Introspectionism. *Mind & Langua*ge, *28*(3), 259-285.

Reynolds, J.H., & Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, *27*, 611-647.

Riggall, A.C., & Postle, B.R. (2012). The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. *Journal of Neuroscience*, *32*(38), 12990-12998.

Rose, N.S., LaRocque, J.J., Riggall, A.C., Gosseries, O., Starrett, M.J., Meyering, E.E., & Postle, B.R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*(6316), 1136-1139.

Rossi, A.F., Bichot, N.P., Desimone, R., & Ungerleider, L.G. (2007). Top–down attentional deficits in macaques with lesions of lateral prefrontal cortex. *Journal of Neuroscience*, *27*(42), 11306-11314.

Rowe, J., Friston, K., Frackowiak, R., & Passingham, R. (2002). Attention to action: specific modulation of corticocortical interactions in humans. *Neuroimage*, *17*(2), 988-998.

Ruff, C.C. (2011). A systems-neuroscience view of attention. In C. Mole, D. Smithies, & W. Wu (Eds.), *Attention: Philosophical and psychological essays* (pp. 1-23). Oxford: Oxford University Press.

Ruff, C.C., Bestmann, S., Blankenburg, F., Bjoertomt, O., Josephs, O., Weiskopf, N., ... & Driver, J. (2007). Distinct causal influences of parietal versus frontal areas on human visual cortex: evidence from concurrent TMS–fMRI. *Cerebral Cortex*, *18*(4), 817-827.

Ruff, C.C., Blankenburg, F., Bjoertomt, O., Bestmann, S., Freeman, E., Haynes, J.D., ... & Driver, J. (2006). Concurrent TMS-fMRI and psychophysics reveal frontal influences on human retinotopic visual cortex. *Current Biology*, *16*(15), 1479-1488.

Sakai, K., Rowe, J.B., & Passingham, R.E. (2002). Active maintenance in prefrontal area 46 creates distractor-resistant memory. *Nature Neuroscience*, *5*(5), 479-484.

Serences, J.T., Ester, E.F., Vogel, E.K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. *Psychological Science*, *20*(2), 207-214.

Sergent, C. (2018). The offline stream of conscious representations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170349.

Sergent, C., Baillet, S., & Dehaene, S. (2005) Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience*, *8*(10), 1391-1400.

Shallice, T., & Warrington, E.K. (1970). Independent functioning of verbal memory stores: A neuropsychological study. *The Quarterly Journal of Experimental Psychology*, *22*(2), 261-273.

Shepard, R.N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701-703.

Shipstead, Z., Lindsey, D.R., Marshall, R.L., & Engle, R.W. (2014). The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *Journal of Memory and Language*, *72*, 116-141.

Shipstead, Z., Redick, T.S., Hicks, K.L., & Engle, R.W. (2012). The scope and control of attention as separate aspects of working memory. *Memory*, *20*(6), 608-628.

Siewert, C. (1998). *The significance of consciousness*. Princeton, NJ: Princeton University Press.

Sligte, I.G., Scholte, H.S., & Lamme, V.A. (2008). Are there multiple visual short-term memory stores?. *PLOS one*, *3*(2), e1699.

Smith, E.E., & Jonides, J. (1999). Storage and executive processes in the frontal lobes. *Science*, *283*(5408), 1657-1661.

Soto, D., & Silvanto, J. (2014). Reappraising the relationship between working memory and conscious awareness. *Trends in Cognitive Sciences*, *18*(10), 520-525.

Soto, D., & Silvanto, J. (2016). Is conscious awareness needed for all working memory processes?. *Neuroscience of Consciousness*, *2016*(1), niw009.

Soto, D., Mäntylä, T., & Silvanto, J. (2011). Working memory without consciousness. *Current Biology*, *21*(22), R912-R913.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological monographs: General and Applied*, *74*(11), 1-29.

Sperling, G. (1963). A model for visual memory tasks. *Human Factors*, *5*(1), 19-31.

Sreenivasan, K.K., Curtis, C.E., & D'Esposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in Cognitive Sciences*, *18*(2), 82-89.

Stein, T., Kaiser, D., & Hesselmann, G. (2016). Can working memory be non-conscious?. *Neuroscience of Consciousness*, *2016*(1), niv011.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*(3736), 652-654.

Stone, J. M., & Towse, J. (2015). A working memory test battery: Java-based collection of seven working memory tasks. *Journal of Open Research Software*, *3*.

Todd, J.J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, *428*(6984), 751-754.

Todd, J.J., Fougnie, D., & Marois, R. (2005). Visual short-term memory load suppresses temporo-parietal junction activity and induces inattentional blindness. *Psychological Science*, *16*(12), 965-972.

Trübutschek, D., Marti, S., & Dehaene, S. (2019). Temporal-order information can be maintained in non-conscious working memory. *Scientific Reports*, *9*(1), 6484.

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.

Trübutschek, D., Marti, S., Ueberschär, H., & Dehaene, S. (2018). Probing the limits of activity-silent non-conscious working memory. *bioRxiv*, 379537.

Tye, M. (1995). *Ten problems of consciousness: A representational theory of the phenomenal mind*. Cambridge, MA: MIT Press.

Unsworth, N., & Spillers, G.J. (2010). Working memory capacity: Attention control, secondary memory, or both? A direct test of the dual-component model. *Journal of Memory and Language*, *62*(4), 392-406.

Van Boxtel, J.J., Tsuchiya, N., & Koch, C. (2010). Consciousness and attention: on sufficiency and necessity. *Frontiers in Psychology*, *1*, 217.

Vergara, J., Rivera, N., Rossi-Pool, R., & Romo, R. (2016). A neural parametric code for storing information of more than one sensory modality in working memory. *Neuron*, *89*(1), 54-62.

Vetter, P., & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition*, *27*, 62-75.

Vossel, S., Geng, J.J., & Fink, G.R. (2014). Dorsal and ventral attention systems: distinct neural circuits but collaborative roles. *The Neuroscientist*, *20*(2), 150-159.

Watzl, S. (2011). The philosophical significance of attention. *Philosophy Compass*, *6*(10), 722-733.

Waugh, N.C., & Norman, D.A. (1965). Primary memory. *Psychological Review*, *72*(2), 89-104.

Wells, G.L., & Petty, R.E. (1980). The effects of overt head movements on persuasion: Compatibility and incompatibility of responses. *Basic and Applied Social Psychology*, *1*(3), 219-230.

Woldorff, M.G., Gallen, C.C., Hampson, S.A., Hillyard, S.A., Pantev, C., Sobel, D., & Bloom, F.E. (1993). Modulation of early sensory processing in human auditory cortex during auditory selective attention. *Proceedings of the National Academy of Sciences*, *90*(18), 8722-8726.

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

Wu, W. (2011). Attention as selection for action. In C. Mole, D. Smithies, & W. Wu (Eds.), *Attention: Philosophical and psychological essays* (pp. 97-116). Oxford: Oxford University Press.

Wu, W. (2014a). *Attention: New problems of philosophy series*. London: Routledge

Wu, W. (2014b). Being in the workspace, from a neural point of view: comments on Peter Carruthers,'On central cognition'. *Philosophical Studies*, *170*(1), 163-174.

Zeman, A.Z., Della Sala, S., Torrens, L.A., Gountouna, V.E., McGonigle, D.J., & Logie, R.H. (2010). Loss of imagery phenomenology with intact visuo-spatial task performance: A case of 'blind imagination'. *Neuropsychologia*, *48*(1), 145-155.

Biography

Max Beninger received his B.A. in philosophy from Queen's University in June of 2012. Max then went on to pursue a Ph.D. in philosophy at Duke University. During his time at Duke, Max was awarded the Bass Instructional Fellowship and the Boone Fellowship for Canadian Graduate Students. He received his Ph.D. in September, 2019.