The Neural Substrates of Non-Conscious Working Memory

Fredrik Bergström
In memory of my friend, lover, and the most beautiful mind I have known. Forever remembered, cherished, and admired.

Aida Alipour (1981 - 2016)

Mina vina omd (för-) bli
“nothing worth having comes easily”

Puss
Aida
# Table of Contents

Table of Contents................................................................................................................ iii  
Acknowledgements .............................................................................................................. v  
List of papers........................................................................................................................... vi  
Abstract................................................................................................................................ vii  
Introduction .......................................................................................................................... 1  
Background ............................................................................................................................ 2  
  Building blocks of cognition .............................................................................................. 2  
    Memory .............................................................................................................................. 2  
    Perception & action ......................................................................................................... 4  
    Attention .......................................................................................................................... 5  
  Consciousness .................................................................................................................... 7  
    The scientific study of consciousness ........................................................................... 8  
    Non-conscious cognition .............................................................................................. 8  
    Neural correlates of consciousness .............................................................................. 10  
Attention and conscious experience .................................................................................. 12  
Memory and conscious experience .................................................................................... 14  
  Priming .............................................................................................................................. 15  
  Episodic & semantic memory .......................................................................................... 17  
  Sensory memory .............................................................................................................. 19  
  Working memory ............................................................................................................. 21  
Aim ....................................................................................................................................... 27  
Materials and methods ......................................................................................................... 28  
  Participants ........................................................................................................................ 28  
  Inferring the absence of conscious experience ............................................................... 28  
  Trial procedures and stimuli ............................................................................................. 30  
    Study I ............................................................................................................................. 30  
    Study II ........................................................................................................................... 32  
    Study III ......................................................................................................................... 34  
  Functional magnetic resonance imaging (fMRI) .............................................................. 34  
  Univariate analysis and multi-voxel pattern analysis (MVPA) ......................................... 36  
Results .................................................................................................................................... 38  
  Study I ............................................................................................................................... 38  
  Study II ............................................................................................................................. 39  
  Study III ........................................................................................................................... 40  
Discussion ............................................................................................................................... 43  
  What kind of non-conscious memory is it? ...................................................................... 43  
    Working memory .......................................................................................................... 43  
    Iconic memory .............................................................................................................. 48  
    Priming ........................................................................................................................... 48  
    Hippocampi-based memory ......................................................................................... 49
Interim summary ........................................................................................................ 50
Critique against non-conscious memory ................................................................. 50
Is non-conscious memory weak and fragile? .......................................................... 53
Implications for working memory models ............................................................... 54
Implications for theories of consciousness ............................................................ 56
Practical implications ............................................................................................... 57
Synthesis .................................................................................................................. 58
Limitations and future directions ............................................................................. 60
Conclusions ............................................................................................................... 62
References ............................................................................................................... 63
Acknowledgements

First and foremost I want to thank my two and a half supervisors for their collective knowledge, guidance, and support, Johan Eriksson for his reflective and analytical approach, Lars Nyberg for his piercing and insightful comments, and Greger Orädd for his advice during the brief period as supervisor.

Secondly, I want to thank Benoni Edin for stimulating conversations and valuable book recommendations. One of those books was by Joaquín Fuster, who although minimal personal correspondence, has greatly influenced my conceptual understanding of the brain through his impressive body of work.

Thirdly I want to thank all my current and former colleagues for their support, especially Micael Andersson for technical support, Anders Lundqvist for fun with statistics, Göran Westling for building me stereoscopes, Mikael Stiernstedt for tips and tricks with Illustrator and Photoshop, Karolina Kauppi, Sara Pudas, Linnea Karlsson Wirebring, Per Nordmark, Lenita Lindgren, Urban Ekman, Sara Stillesjö, Amar Awad for creating a stimulating environment and random hilarities, the MR staff and administrators for narrowing my focus of attention.

Last but not least, I want to thank my friends and family for their support and help in alleviating any social isolation, especially Mauritz Börjeson and Jagannath Tammeleth for keeping in touch despite long distances and busy schedules. Special thanks to Aida Alipour for making life colorful.
List of papers

This doctorate dissertation is based on the following three studies:


Other papers

Abstract

**Background:** Despite our distinct impression to the contrary, we are only conscious of a fraction of all the neural activity underlying our thoughts and behavior. Most neural processes occur non-consciously, and in parallel with our conscious experience. However, it is still unclear what the limits of non-conscious processes are in terms of higher cognitive functions. Many recent studies have shown that increasingly more advanced functions can operate non-consciously, but non-conscious information is still thought to be fleeting and undetectable within 500 milliseconds. Here we used various techniques to render information non-conscious, together with functional magnetic resonance imaging (fMRI), to investigate if non-consciously presented information can be retained for several seconds, what the neural substrates of such retention are, and if it is consistent with working memory maintenance.

**Results:** In *Study I* we used an attentional blink paradigm to render stimuli (single letters) non-conscious, and a variable delay period (5 – 15 s) prior to memory test. It was found that non-conscious memory performance was above chance after all delay durations, and showed no signs of decline over time. Univariate fMRI analysis showed that the durable retention was associated with sustained BOLD signal change in the prefrontal cortex and cerebellum during the delay period. In *Study II* we used continuous flash suppression (CFS) to render stimuli (faces and tools) non-conscious, and a variable delay period (5 or 15 s) prior to memory test. The durable retention of up to 15 s was replicated, and it was found that stimuli identity and spatial position was retained until prospective use. In *Study III* we used CFS to render tools non-conscious, and a variable delay period (5 – 15 s) prior to memory test. It was found that memory performance was not better than chance. However, by using multi-voxel pattern analysis it was nonetheless possible to detect the presence vs. absence of non-conscious stimuli in the frontal cortex, and their spatial position (left vs. right) in the occipital cortex during the delay.

**Conclusions:** Overall these findings suggest that non-consciously presented information (identity and/or position) can be retained for several seconds, and is associated with BOLD signal in frontal and posterior regions. These findings are consistent with working memory maintenance of non-consciously presented information, and thereby constrain models of working memory and theories of consciousness.

**Keywords:** non-conscious, working memory, neural substrates, visual perception, consciousness, functional magnetic resonance imaging (fMRI)
**Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>attentional blink</td>
</tr>
<tr>
<td>BOLD</td>
<td>blood-oxygenation-level-dependent (signal)</td>
</tr>
<tr>
<td>CFS</td>
<td>continuous flash suppression</td>
</tr>
<tr>
<td>EEG</td>
<td>electroencephalography</td>
</tr>
<tr>
<td>fMRI</td>
<td>functional magnetic resonance imaging</td>
</tr>
<tr>
<td>GNW</td>
<td>global neuronal workspace (theory)</td>
</tr>
<tr>
<td>HOT</td>
<td>higher-order thought (theory)</td>
</tr>
<tr>
<td>ISI</td>
<td>inter stimulus interval</td>
</tr>
<tr>
<td>ITI</td>
<td>inter trial interval</td>
</tr>
<tr>
<td>LGN</td>
<td>lateral geniculate nucleus</td>
</tr>
<tr>
<td>MEG</td>
<td>magnetoencephalography</td>
</tr>
<tr>
<td>MVPA</td>
<td>multivoxel pattern analysis</td>
</tr>
<tr>
<td>PAS</td>
<td>perceptual awareness scale</td>
</tr>
<tr>
<td>tDCS</td>
<td>transcranial direct-current stimulation</td>
</tr>
<tr>
<td>TMS</td>
<td>transcranial magnetic stimulation</td>
</tr>
</tbody>
</table>
Introduction

We are memories. Our memories define who and what we are, that is, everything from what species we are to our individual personalities. As we spring into existence we have already inherited memories from our parents and ancestors all the way back to the beginning of life on earth. Our genetic inheritance interacts with our environment, and form new memories throughout our existence. The vast majority of the memories that make up our uniqueness as human beings is at any given time non-consciously (i.e., not consciously experienced), and latently stored as widely-distributed and interconnected webs of neurons. Together they constitute the grey and squishy bundle of nerve-cells commonly known as the brain.

Normally, subsets of these interconnected neurons will become active and fire signals between each other, which enables us to go about our everyday lives. Inexplicably, these activated networks of neurons also enable us to consciously experience the content of our current memories as we perceive our external and internal environment – sensory experiences, emotions, goals and motivations, sense of agency, etc. – as we navigate our way through the adventures of life. How our brain gives rise to our subjective experiences remains a mystery. Philosophers call this mystery “the explanatory gap”. This gap is what some philosophers and cognitive neuroscientists are trying to explain. One way to become informed about what is special with consciousness, is to get a better understanding of what is not special with it, by investigating the limits of non-conscious cognition.

There is a special kind of short-term memory that is intimately linked with conscious experience. It is called working memory – the temporary retention of information for prospective use – and is immensely important for us. Working memory is used when we, for example, perform mental arithmetic, are trying to solve difficult problems, or doing contingency planning for the future. Not surprisingly, working memory capacity is therefore a good predictor of general intelligence and success in life. Indeed, most of our advanced behavior is associated with conscious experience and working memory. However, most of our neural activity occurs non-consciously, and in parallel with our conscious experience. The limits of our non-conscious cognition in terms of higher cognitive functions remain unclear. Empirical studies have shown that increasingly more advanced cognitive functions can operate non-consciously, but it is still believed that non-consciously perceived information is fleeting and undetectable within 500 milliseconds. The questions we have tried to find answers to within the context of this dissertation are if non-conscious presented information can be retained for several seconds, what the neural substrates of such retention are, and if it is consistent with working memory maintenance.
Background

Building blocks of cognition
This section is meant to briefly define the functional properties of the brain that are relevant for this dissertation, and give a brief overview of the functional organization of the brain. Importantly, the cognitive functions will be defined at a psychological level, as well as anchored in neural terms.

Memory
Our memories define what and who we are, without them, we would literally not exist. Memory is therefore the most fundamental property of any organism, and forms the basis of all cognition. Indeed, the word cognition, which today is used as a blanket term for all cognitive functions, comes from the Latin word cognoscere meaning “to learn” or “to know”. Broadly defined, memory is an organism’s capacity to retain information about itself and its external environment, such as in tissue scarring, the immune system, or the brain. However, here we will be focusing on memories in the brain.

The science of memory has a long history dating back to the ancient Greek philosophers (Hergenhan, 2001; Radvansky, 2014). Plato thought humans had immortal souls that bathed in pure and complete knowledge, which only could be attained by introspection and reason, rendering sensory experience redundant. He therefore thought humans were born with all their memories. Aristotle, instead, thought humans were born as tabulae rasae (“blank slates”), and that memories were imprints (as on wax tablets) caused by sensory experience to later be recalled. Contrary to Plato, he therefore thought that we acquire all our memories through experience. Aristotle also postulated what were to become the bases of all learning theory, namely, the laws of association. These laws state that when we think of something we also tend to think of things that are: experienced along with it (the law of contiguity), similar to it (the law of similarity), and its opposite (the law of contrast). Additionally, he posited that associations generally become stronger the more they are experienced (the law of frequency), but can sometimes become strong after only one experience. Aristotle’s ideas on memory, as we shall see, were not all too far off, except about memories being stored in the heart instead of the brain.

Ebbinghaus (1885/1913) was the first experimental psychologist to systematically study memory and learning. He studied himself while trying to learn lists of nonsense syllables with varying list lengths, retention intervals, and learning conditions. From his meticulous efforts he was able to estimate how long it took, and how frequent practice was needed to learn certain information (the learning curve), the optimal spacing between practice sessions, the time it takes to forget information (the forgetting curve), that
overlearning could extend the forgetting curve, and that forgotten memories no longer consciously accessible nevertheless could have a non-conscious influence on behavior and learning.

In the brain, memories are stored in the structure of interconnected and distributed networks of neurons. New neural memories are formed by Hayek-Hebbian principles (“neurons that fire together wire together”) to create new or modify existing neural networks (Hayek, 1952; Hebb, 1949). These neural connections are formed throughout the neocortex with important modulatory input from subcortical limbic regions, such as the hippocampi and amygdalae. Thus, similarly to what Aristotle postulated, connections between neurons (i.e., memories) are formed by way of associations that become stronger with frequency. However, an emotional event can cause life-long associations by way of modulatory input from the amygdalae.

Contrary to the thoughts of Plato and Aristotle, organisms can be said to consist of innate memories (that we are born with), and acquired memories. Innate memories are called phyletic memories (“memory between species and organisms”), while memories formed by experience are called ontogenetic memory (“memory within the life span of an organism”). Phyletic memory consists of the genetically determined functional and structural layout of the brain (all things that are similar across individuals), and serves as the base on which ontogenetic memories grow (things that are different between individuals). Phyletic and ontogenetic memories are intertwined, and there is no way to clearly separate them. The evolutionarily older primary sensory and motor regions seem to be relatively more phyletic, while the relatively newer cortical regions of association seem to be more ontogenetic. However, even primary sensory cortices need some sensory experience during critical periods to develop functionally. In the broadest sense, we therefore consist entirely of memories (our ancestor’s and our own).

Memory, therefore, is a global property of the brain, and cannot be assigned to any specific part of the brain. However, memory can be differentiated based on its current state (active or inactive) over time. The difference between active and inactive memory states has been used to distinguish between short-term (active) and long-term (inactive) memory, as well as conscious (active) and non-conscious (inactive) states. Traditionally, short-term memory was therefore assumed to be an active conscious state, while latent long-term memory was assumed to be an inactive non-conscious state (until reactivated). Short-term memory (or retention) is henceforth used as an umbrella term to refer to any short-lived memory (i.e., sensory memory, working memory, and short-lived latent neural changes). Sensory memory is a brief high capacity memory, and working memory is a durable low capacity memory dependent on prospective use. Memory can also be spatially differentiated in terms of its specific content. That is, different memory content can be found in different areas of the brain. The spatial distinction
between content and the temporal distinction among states largely underlies
the categorization of memory (Fuster, 1995; Tulving, 1972). The various
categorizations of short-term and long-term memory will be more closely
reviewed below in the section on memory and conscious experience.

**Perception & action**

Simplified, the frontal cortex can be said to contain motor memory, while the
posterior parts of the cortex contain perceptual memory, but the two are
intertwined and the distinction is merely conceptual (Fuster, 1995, 2003a,
2015). In the posterior cortex, perceptual knowledge is organized
hierarchically with concrete unimodal elementary sensory features in the
primary sensory cortices, and progressively more abstract polymodal sensory,
semantic, and conceptual knowledge towards the central regions of the
posterior cortex (Binder & Desai, 2011; Fuster, 1997, 2003a). Perception is not
a passive process, nor does it reflect reality as it really is. As suggested by
Berkeley in 1709, Helmholtz in 1925, and Hayek (1952) perception is the active
process of interpretation and categorization of sensory information guided by
memory (Fuster, 2003a; Palmer, 1999). Essentially, we perceive the world as
we remember it. That is, new sensory information reactivates (phylogenic and
ontogenetic) memory to process and interpret it. In the process, new sensory
information is stored as extensions to old perceptual memory, which in turn
will interpret incoming sensory information accordingly. To perceive,
therefore, is to remember or “re-cognize” (Fuster, 1995).

Motor knowledge follows a similar hierarchical gradient such as simple
movement-related information in the primary motor cortex, goal-related
information in premotor cortex, plans and contingencies in the prefrontal
cortex (Badre & D’Esposito, 2009; Fuster, 2015; Koechlin & Summerfield,
2007). It has been suggested that the prefrontal cortex is crucial for temporal
integration of the past and the future, which is an essential component of
working memory (Fuster & Bressler, 2015; Ingvar, 1985). However, the lowest
levels of the motor hierarchy are not in the cortex, but in the basal ganglia,
cerebellum, brain stem, and spinal cord, where well-learned sensory-motor
interactions and reflexive actions are stored. Action is the activation of
(phylogenic and ontogenetic) executive and/or simpler motor memory.

The perceptual and motor hierarchies are interconnected at all hierarchical
levels in the cortex, and together with subcortical modulation, form a
perception-action cycle (Fuster & Bressler, 2015; Fuster, 2003a, 2004). The
perception-action cycle makes the brain a causally deterministic self-guiding
system, without need of an executive homunculus or free will to govern it. It
is through the perception-action cycle an organism traverses its environment,
by gathering sensory information as input, transforming it to actionable
output, as it continuously learns and adapts to the environment. The
perception-action cycle is also crucial for sustaining recurrent activity of
internal representations without external input, for example during working memory. However, perception as well as action, relies on attention to be able to select what to perceive and what action to initiate. Indeed, perception and action is the interaction between memory and attention.

**Attention**

In our everyday lives, aspects of attention can be described by the “cocktail party effect” (Broadbent, 1958). That is, in a room full of voices it is possible to selectively focus on and understand one voice at a time (and switch as you please), while all other voices become garbled background noise. However, if one of the background voices addresses you by name you are likely to automatically switch focus to that voice. This scene captures many of the distinguishing features of attention. Throughout history there has been many influential cognitive models of attention born out of the cocktail party effect (Broadbent, 1958; Driver, 2001; Neisser, 1967; Treisman & Gelade, 1980; Treisman, 1960). However, I prefer the following one, which is stated more closely to neurophysiology, but is nonetheless consistent with neuropsychological findings.

**Attention** is another fundamental property of the nervous system, and can be defined as the **selective allocation of neural resources**. In neural terms this translates to excitation of some neural networks and inhibition of other competing or irrelevant networks (Desimone & Duncan, 1995; Franconeri, Alvarez, & Cavanagh, 2013; Fuster, 2003a). The efficient allocation of attentional or neural resources is crucial for normal functioning because of its limited capacity (Marois & Ivanoff, 2005). The **focus (or foci) of attention** is the neural networks with most resources allocated to them, that is, most neural activity at a specific point in time. The information inside the focus of attention is usually consciously experienced. Outside the focus of attention there is less neural activity, that is, less attended networks, as well as, unattended networks that are (more or less) completely inhibited. Less attended information can be consciously or non-consciously processed, while unattended information is not processed at all (Chun & Marois, 2002). **Exogenous attention** is when the allocation of attentional resources are caused by external stimuli, usually something salient (e.g., a loud bang or one’s name) that reflexively captures ones attention. **Endogenous attention** refers to when attentional resources are allocated by internal processes such as an intention or will to achieve a goal or task. However, endogenous and exogenous attention does not capture all aspects of attention. Other subcortical structures involving motivation, reward, and emotion also modulates attention in the cortex (Awh et al., 2012; Fuster, 2003a; Petersen & Posner, 2012).

Sustained attention is likely driven by recurrent neural activity at some stage in the perception–action cycle. Attention does not only operate on
perception of external stimuli, but can also be directed internally (Nobre, Coull, Maquet, & Frith, 2004). When internally directed, sustained attention maintains activity in perceptual networks without the need for external stimuli. Examples of internally directed attention are working memory and mental imagery (e.g., visualizing or imagining something from memory). Working memory is the interaction between attention and memory, without external input, for its prospective use.

Other concepts that have been associated with attention are arousal, vigilance, and alertness (Petersen & Posner, 2012; Posner & Petersen, 1990). These properties are strongly associated with subcortical structures such as the reticular formation (Moruzzi & Horace, 1949) and the cholinergic system (Perry, Walker, Grace, & Perry, 1999). Although arousal clearly interacts with attention (Fuster, 1958) I prefer to separate them. Whereas attention concerns the allocation of neural resources, arousal concerns the available quantity of neural resources, rather than the allocation per se.

The conception of attention as neural competition, excitation and inhibition, among neural networks (Desimone & Duncan, 1995; Franconeri et al., 2013; Fuster, 2003a) has the merit of working equally well within all systems, and at all levels of the nervous system. For example, neural competition from on/off-centers in retinal ganglion cells to distributed networks in visual associative cortex, and distributed networks in the prefrontal cortex to reciprocal spinal innervation of extensor and flexor muscles in limbs (Fuster, 2003a). It is also consistent with the load theory of attention, which holds that attentional resources are allocated by an exogenous perceptual selection mechanisms, and an endogenous attentional control mechanism (Lavie, Hirst, de Fockert, & Viding, 2004; Lavie & Tsal, 1994; Lavie, 1995). Further empirical support comes from cell recordings in primates (Chelazzi, Miller, Duncan, & Desimone, 1993; Moran & Desimone, 1985), and human neuroimaging (Rees, Frith, & Lavie, 1997; Rees, Rees, Russell, & Frith, 2012).

Attention, like memory, is thus a global property that exists in the whole brain, not only in a specific brain region, and can be differentiated based on informational content. For example, lesions to the right parietal cortex can lead to unilateral neglect, that is, a dysfunction in spatial attention. Patients with spatial neglect do not attend to their left side in space, and can for example miss shaving the left side of their face or leave the food on the left side of the plate (Corbetta & Shulman, 2002, 2011). Discrete lesions in the visual cortex will make it impossible to attend to (and perceive or remember) that specific visual content.

When we apply the various attentional terms to the cocktail party context we get the following. We can endogenously focus attention between single voices as we please, and thus consciously experience and understand its content. Although the garbled noise of voices in the background is barely
noted, it is still consciously and/or non-consciously processed to some extent. When our name is non-consciously processed it can induce an exogenous shift to attend the voice that uttered your name, and thereby make it conscious.

**Consciousness**

It is still unclear whether consciousness is a function such as memory, perception, action, and attention, or if it is an epiphenomenon (i.e., without function - a fancy side effect of being). To make matters worse, there is no proper definition of consciousness that avoids circular reasoning (Block, 1995; Chalmers, 1996), and it is an ambiguous concept that can mean different things among laymen, philosophers, and scientists (Zeman, 2001). However, conscious experience is here assumed to be a property of the brain, and I will try to clarify what it refers to in this dissertation.

A common conceptual distinction is made between states or levels of consciousness, and the content of consciousness. The former refers to different global states of general arousal (e.g., coma, sleep, and wakefulness), while the latter refers to the subjective experience of being, that is, what it is like to experience something (e.g., the redness of red or the sensation of pain; Nagel, 1974). This is, however, merely a conceptual distinction, as the two are intimately linked (Hohwy, 2009). The content of consciousness varies, and can for example be perceptual (e.g., visual, auditory), executive (e.g., intentions, sense of agency), or metacognitive (i.e., cognitive states about other cognitive states) in nature. Traditionally, states or levels of consciousness have been depicted two-dimensionally as a function of the content of consciousness (Laureys, 2005). However, it has recently been suggested that global states of consciousness more aptly should be conceptualized as multi-dimensional states with content- and functionality-related dimensions (Bayne, Hohwy, & Owen, 2016). I tend to agree, and although a certain global state of arousal may be necessary to experience conscious content, it should not be conflated with it. In sum, the global state of arousal determines the amount of neural resources available, attention is the allocation of resources among competing networks, and conscious experience is the phenomenon that somehow ensues when networks receive enough neural resources.

This dissertation will mainly focus on the content of consciousness, and the use of conscious, consciousness, phenomenal, or experience will henceforth refer to when a subjective experience accompanies certain neural processes. Likewise, non-conscious will refer to the absence of subjective experiences (despite certain neural processes). The neural substrates of conscious experience are not known, and what the scientific study of consciousness is trying to find out.
The scientific study of consciousness

The scientific study of conscious experience began in the late 1800’s with Fechner’s psychophysics, Wundt’s and Titchener’s introspective approaches, and William James writings, but disappeared in the 1920’s as behaviorism rejected all things subjective and unobservable (Hergenhan, 2001; Revonsuo, 2010). It was not until the late 1900’s that consciousness began to be considered a respectable topic for scientific and philosophical research again (Baars, 1988; Chalmers, 1996; Crick & Koch, 1998; Nagel, 1974).

The scientific study of consciousness have since then mainly focused on finding the functional and neural correlates of consciousness (Crick & Koch, 2003; Koch, Massimini, Boly, & Tononi, 2016). The former tries to determine what functions, if any, are uniquely associated with consciousness, and to dissociate those functions (e.g., perception, action, attention, and memory) from conscious experience. The latter is concerned with finding the specific neural substrates underlying the conscious experience of certain content.

In both cases it has been common to use a simple subtraction logic when designing experiments. Such experiments usually have three conditions: (i) baseline trials without target stimulus, and therefore no conscious experience of the absent stimulus, (ii) trials with target stimulus, but without conscious experience of the stimulus because of some manipulation, and (iii) trials with target stimulus and conscious experience of the stimulus. Subtracting (ii) from (iii) will reveal correlates of conscious experience. Although successful, it has been pointed out that subtraction experiments nonetheless can be confounded by other functional properties that closely correlate with conscious experience, such as attentional control or the reporting of perceptual experience (Aru, Bachmann, Singer, & Melloni, 2011; Naghavi & Nyberg, 2005). Researchers are thus continuously trying to develop new ways to separate the neural correlates of conscious experience from possible confounds (e.g., no-report paradigms; Tsuchiya, Wilke, Frässle, & Lamme, 2015). Subtracting (i) from (ii) will reveal correlates of non-conscious processing. If certain functions and neuronal mechanisms are found to be associated with non-conscious processing, they can be inferred not to be neural substrates and functions unique to consciousness. It is therefore not only relevant to explore the functional and neural boundaries of non-conscious cognition for its own sake, but also to become informed about conscious experience.

Non-conscious cognition

The idea of non-conscious perception was perhaps first expressed by Leibniz in 1765, which he called petites perceptions (little perceptions). Leibniz thought these petites perceptions were too small to be experienced, but if enough of them were aggregated they would eventually pass a limen (i.e., threshold), and become conscious (Hergenhan, 2001; Schacter, 1987). The
idea of a limen proved to be very influential in experimental psychology as investigations into subliminal (under the threshold or simply non-conscious) perception began in the late 1800’s, and remains relevant today. The first experiments found support for non-conscious perception (Pierce & Jastrow, 1884; Sidis, 1898; Stroh, Shaw, & Washburn, 1908), and similar results followed until Eriksen (1960) criticized the use of subjective measures of conscious experience. The result was wide-spread skepticism about the existence of non-conscious perception. In the 1980’s several studies using objective measures and process-dissociation procedures found evidence of non-conscious semantic priming. However, most of these studies were methodologically flawed (Holender, 1986). After further methodological and technical advances there was a consensus about the existence of non-conscious lexical and orthographical processing in the 1990’s, and the controversy shifted from the existence towards the extent or limits of non-conscious cognition (for a historical review see Kouider & Dehaene, 2007). This was about the same time that serious efforts toward a scientific study of consciousness, and the search for neural correlates of consciousness began (Baars, 1988; Chalmers, 1996; Crick & Koch, 1998; Nagel, 1974). Since the 1990’s it has been shown that non-consciously presented information can be processed at all levels throughout the visual system (Dehaene & Changeux, 2011; Dehaene, Charles, King, & Marti, 2014; Rees, Kreiman, & Koch, 2002). This suggests that nowhere in the visual system is neural activity in itself sufficient for having visual experiences.

Even more interesting is that studies have continued to push the limits of non-conscious cognition, and found that non-conscious information can engage increasingly more complex and flexible executive functions (some associated with prefrontal and parietal cortex). This suggests that neural activity in the prefrontal and parietal cortices might also not be sufficient in itself for conscious experience. For example, it has been shown that the monetary value of non-consciously presented images of coins can be processed in reward related brain regions (Pessiglione et al., 2007) and modulate working memory performance (Zedelius, Veling, & Aarts, 2011). Non-consciously presented information can engage various cognitive control and inhibitory mechanisms, and related frontal cortical regions (Lau & Passingham, 2007; Reuss, Kiesel, & Kunde, 2015; Reuss, Kiesel, Kunde, & Hommel, 2011; van Gaal, Lamme, Fahrenfort, & Ridderinkhof, 2011; van Gaal, Ridderinkhof, Scholte, & Lamme, 2010). Error-monitoring has been shown to occur non-consciously by measuring post-error slowing in reaction times despite participants not being conscious of making the errors (Logan & Crump, 2010; Pavone, Marzi, & Girelli, 2009). Simpler arithmetic tasks such as adding two single digits (Ric & Muller, 2012), and adding or subtracting three single digits (Sklar et al., 2012) seems to be performed non-consciously. Indeed, it seems like most if not all cognitive processes can be engaged non-
consciously to some extent. The difference between conscious and non-conscious processes does not seem to be one of kind, but rather of degree. That is, non-conscious neural and behavioral effects tend to be small compared to its conscious counterpart, but to what extent that depends on conservative experimental conditions is unclear.

**Neural correlates of consciousness**

It is still somewhat unclear what the neural correlates of consciousness are, but there has been considerable progress, especially in the visual system. At the lowest level of the visual system, in the retinae, different wave-lengths of light are received from the external environment, and transformed into neural signals, which propagates through the visual system, and ultimately, somehow, causes the vivid visual experiences we have every day. Activity in the retinae does not seem to correlate with conscious perception. Patients without functioning retinae experience phosphenes (flashes of light) when transcranial magnetic stimulation (TMS) is used on the visual cortex (Koch, 2004). Furthermore, the retinal “blind spots” where nerve fibers pass through the retinae are not visually experienced as blind spots (Koch, 2004). The retinae are therefore not necessary for conscious perception. The next stop in the visual system is the lateral geniculate nucleus (LGN) in the thalamus, which is a relay station between the retinae and the primary visual cortex. The LGN is the earliest stage of visual processing that correlates with conscious perception. During binocular rivalry BOLD signal change in eye-specific regions increase when that eye’s input is conscious, and decrease when suppressed (Haynes, Deichmann, & Rees, 2005; Wunderlich, Schneider, & Kastner, 2005). The LGN does not seem to be necessary since TMS to the visual cortex still induce phosphenes. However, it is possible that TMS induced cortical activity is fed back to the LGN.

The LGN relays 90% of the visual nerve projections to the primary visual cortex, where finer details are processed, such as line orientation and length, motion direction, color, contrast, spatial frequency, and ocular dominance (Tong, 2003). Neural activity in the primary visual cortex is not sufficient for visual experience because there are several cases of non-conscious visual processing in the primary visual cortex (Haynes & Rees, 2005; He & MacLeod, 2001; Jiang, Zhou, & He, 2007; Sergent, Baillet, & Dehaene, 2005; Zou, He, & Zhang, 2016). Lesions in the primary visual cortex disrupt visual experience, and can lead to cases of blindsight, in which patients perform better than chance on forced-choice discriminations, without visual experience of the stimuli (Cowey, 2010; Leopold, 2012). In addition, patients with lesions in higher visual areas, but intact primary visual cortex, do not have visual experiences (Horton & Hoyt, 1991).

It is still a matter of debate whether the primary visual cortex is directly necessary for visual experience, or indirectly necessary as an information
gateway to higher visual regions (Crick & Koch, 1995; Silvanto, 2014; Tong, 2003). Neural activity in the primary visual cortex of primates does not seem to correlate with conscious perception, in contrast to activity in higher visual regions that do correlate with visual experience (Leopold, 2012; Logothetis, 1998). BOLD signal change in the human primary visual cortex sometimes correlate with conscious perception (Harrison & Tong, 2009; Lee, Blake, & Heeger, 2005; Polonsky, Blake, Braun, & Heeger, 2000; Ress & Heeger, 2003), but not when controlling for attention (Bahrami, Lavie, & Rees, 2007; Lee, Blake, & Heeger, 2007; Tse, Martinez-Conde, Schlegel, & Macknik, 2005; Watanabe et al., 2011). Other studies suggest that the primary visual cortex might be directly necessary for visual experience when information is fed back from higher visual regions (Koivisto, Mäntylä, & Silvanto, 2010; Silvanto, Cowey, Lavie, & Walsh, 2005). It is therefore still unclear in what capacity the primary visual cortex is necessary for visual experience.

Visual information is relayed from the primary visual cortex to higher visual regions where information, such as color, motion, spatial position, faces, and other aspects are processed in relatively specialized regions. (for review see Grill-Spector & Malach, 2004). The higher visual regions follow a ventral stream that is important for object recognition in the inferior temporal cortex (Kravitz, Saleem, Baker, Ungerleider, & Mishkin, 2012), and a dorsal stream toward the parietal cortex that processes spatial information in relation to visually guided action and navigation (Kravitz, Saleem, Baker, & Mishkin, 2011; Whitwell, Milner, & Goodale, 2014). Cell-recordings in primates, and human neuroimaging studies converge on the finding that activity/signal in specialized visual regions correlate with visual experience of the same content (Koch, Massimini, Boly, & Tononi, 2016; Logothetis, 1998; Lumer, Friston, & Rees, 1998; Moutoussis, Keliris, Kourtzi, & Logothetis, 2005; Rees, Kreiman, & Koch, 2002; Rees, 2007; Zeki, Watson, & Frackowiak, 1993). Lesions to specialized visual regions eliminates the possibility of visual experiences of such content (Barton, 2011). In addition, activity in higher visual cortex correlates with visual experience in the absence of external stimuli, in cases as visual imagery (Kreiman, Koch, & Fried, 2000; O’Craven & Kanwisher, 2000), hallucinations (Ffytche et al., 1998), and illusions (Zeki et al., 1993). For example, disrupting the region that processes motion with TMS will disrupt the illusion of motion (Ruzzoli et al., 2011). The higher visual cortex is therefore necessary for visual experience, but not sufficient. Because it is now known that non-conscious visual processing can occur throughout the higher visual regions (Dehaene et al., 1994; Dehaene & Changeux, 2011; Fang & He, 2005; Marois, Yi, & Chun, 2004; Moutoussis & Zeki, 2002).

Additionally, most previous studies consistently found that neural activity and BOLD signal change in prefrontal and parietal cortex correlated with conscious perception (Dehaene & Changeux, 2011; Naghavi & Nyberg, 2005; Rees et al., 2002; Rees, 2007; Tononi & Koch, 2008), which has been the
foundation of some prominent theories of consciousness (for more details see “implications for theories of consciousness” in the discussion). However, the latest efforts in trying to disentangle the neural correlates of consciousness from other closely related processes (Naghavi & Nyberg, 2005) have found that BOLD signal in the prefrontal cortex likely is associated with attention (Eriksson, Larsson, & Nyberg, 2008; Tse et al., 2005), and reporting the presence of visual experience (Frässle, Sommer, Jansen, Naber, & Einhauser, 2014) rather than visual experience per se. Frässle et al. (2014), elegantly, used the eye’s nystagmus reflex and pupil size as a proxy for explicit reports of conscious experience in a binocular rivalry paradigm. One of two competing stimuli (dark red or light green bars moving in opposite directions) were presented to each eye, the nystagmus reflex correlated with the motion, and the pupil size with the brightness, of the consciously experienced stimulus. Using fMRI they showed that the BOLD signal in frontal cortex did not correlate with conscious perception without explicit reports, and concluded that frontal cortex is related to introspection and action rather than sensory experience. It therefore seems like the frontal cortex might not be necessary for conscious sensory experience.

**Attention and conscious experience**

Attention and conscious experience are intimately linked, and it was previously thought to be the same function (William, 1890). However, it is now widely acknowledged that attention and conscious experience are two different properties of the brain. The debate now revolves around their relationship, and whether attention and conscious experience can be dissociated from each other. Attention is often used synonymously with endogenous attention, but I will separate attention from endogenous and exogenous attention. Here attention will refer to resource allocation by any mechanism. The latter two being different mechanisms for allocation of a shared neural resource. Some argue that endogenous attention is necessary, but not sufficient, for conscious experience (Cohen, Cavanagh, Chun, & Nakayama, 2012; Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006), while others argue that endogenous attention is neither necessary nor sufficient for conscious experience (Koch & Tsuchiya, 2007; Lamme, 2004).

It is clear that attention modulates visual experience. Covert exogenous attention affects the visual experience of stimuli in terms of enhanced contrast (Carrasco, Ling, & Read, 2004), spatial resolution and acuity (Gobell & Carrasco, 2005), and color (saturation, but not hue; Fuller & Carrasco, 2006). Similarly, covert endogenous attention can modulate the conscious experience of brightness (Tse, 2005), contrast (Liu, Abrams, & Carrasco, 2009), and spatial resolution (Abrams, Barbot, & Carrasco, 2010). However, it is still unclear whether endogenous and exogenous attention can have different effects on consciousness.
Exogenous attention can be engaged by non-conscious information. Non-consciously presented nude pictures of the opposite sex will automatically shift attention towards them, and for males also away, if the nude pictures are of the same sex (Jiang, Costello, Fang, Huang, & He, 2006). Endogenous attention to cues presented in blindsight patients’ “blind spot” speeds up their responses (Kentridge, Heywood, & Weiskrantz, 2004; Kentridge, Nijboer, & Heywood, 2008). It has also been shown that cuing exogenous spatial attention toward non-consciously presented stimuli can increase non-conscious priming effects (Marzouki, Grainger, & Theeuwes, 2007).

Conscious endogenous attention toward non-conscious Gabor gratings flanked by distractors (i.e., “crowding”) in the peripheral visual field does not necessarily make the orientation of the gratings lines conscious (He, Cavanagh, & Intriligator, 1996). Endogenous focal attention to conscious object-features automatically modulates non-conscious stimuli outside the focus of attention that shares the same features (Kanai, Tsuchiya, & Verstraten, 2006; Melcher, Papathomas, & Vidnyánszky, 2005). It has also been shown that endogenous attention and conscious experience can have opposite effects on the duration of afterimages (van Boxtel, Tsuchiya, & Koch, 2010). However, that conscious endogenous attention can modulate non-conscious stimuli without making them conscious, is not the same as non-conscious endogenous attention. Although attention was not of primary interest, cases of non-consciously engaged cognitive control/inhibition are examples of non-conscious endogenous motor attention (Lau & Passingham, 2007; Reuss et al., 2011; van Gaal et al., 2010). In such experiments the task is to perform one of two actions depending on which of them are cued. The non-consciously presented cue activate (attend) parts of the cognitive control system, and affects motor output if the non-conscious cue is incongruent with the performed task, and therefore constitutes non-conscious endogenous motor attention.

More controversially, Koch & Tsuchiya (2007) have argued that there are cases of conscious experience without endogenous attention, such as the gist of quickly presented scenes (Fei-Fei, Iyer, Koch, & Perona, 2007; Mack & Rock, 1998), and discrimination of stimuli in the periphery during dual tasks (Li, VanRullen, Koch, & Perona, 2002; Reddy, Reddy, & Koch, 2006). However, discrimination performance has been shown to decrease if endogenous attention is taxed enough when trying to determine the gist of scenes (Cohen, Alvarez, & Nakayama, 2011; Mack & Clarke, 2011; Marois et al., 2004; Yi, Woodman, Widders, Marois, & Chun, 2004), and other stimuli during dual tasks (Evans & Treisman, 2005; Joseph, Chun, & Nakayama, 1997; Landau & Bentin, 2008; Walker, Stafford, & Davis, 2008). Cohen et al., (2012) therefore argue that endogenous attention is necessary for consciousness, but that most tasks do not fully engage all attentional resources, and differences in task difficulty can explain why conscious
experience occur to varying extents while endogenous attention is directed at other tasks. In essence these experiments are in line with the load theory of attention (Lavie et al., 2004), and show that if there are attentional resources to spare, they can be exogenously allocated, and if enough is allocated it becomes conscious. Otherwise the information is non-consciously processed (Chun & Marois, 2002), or not processed at all. The studies also show that the focus of attention is not necessary for conscious experience, since information outside the focus can be experienced in some circumstances.

In sum, attention is necessary (but not sufficient) for, and can modulate, conscious and non-conscious processes. Exogenous and, importantly, endogenous attentional allocation can be engaged by non-conscious information. The fact that endogenous attention can be engaged by non-conscious information is a prerequisite for non-conscious working memory.

**Memory and conscious experience**
From Ebbinghaus’ work in 1885 to 1960 many believed there was only one unitary memory system. However, William James (1890) argued that there was two memory systems, which he called primary and secondary memory. According to James, the primary memory lasted only tens of seconds, while the secondary memory could last for a life time. It was not until 1960 that new findings began to support shorter memories (Palmer, 1999). Those shorter memories were later to be called sensory memory, and short-term or working memory instead of primary memory, in contrast to long-term memory instead of secondary memory. In the context of this dissertation I will be using short-term memory as an umbrella term for any memory that lasts a short time, and not synonymously with working memory.

Sensory memory and working memory have traditionally been conceived as inherently conscious, in contrast to long-term memory, which was thought to be non-conscious while not being actively recollected. Long-term memory can be divided into three epochs; encoding, storage, and retrieval. Studies of long-term memory have resulted in various categories. These categories have traditionally been based on whether the participant/patient could consciously recollect the encoding event during the retrieval epoch or not. The taxonomy based on conscious recollection contained two main categories: explicit and implicit memories. Explicit memory initially included episodic and semantic memory, while implicit memory includes priming, conditioning, and procedural memory (Squire & Zola-Morgan, 1988). However, Tulving (2002) has suggested that semantic memory more aptly should be considered an implicit memory since conscious recollection of the encoding event is not necessary. Although implicit memory is seemingly synonymous with non-conscious memory, an important distinction is made between the usage of implicit and non-conscious here forth. Contrary to explicit memory, implicit memory does not require conscious recollection during retrieval (Graf &
Implicit memory can therefore be consciously or non-consciously encoded. However, for what I call non-conscious memory, it is necessary that the information to be remembered is non-consciously encoded or perceived.

Not all these memory systems are relevant for this dissertation, and I will therefore only elaborate on a subset of them, namely, iconic memory, working memory, episodic/semantic memory, and priming. Although we intend to investigate non-conscious short-term memory, we need to consider the possibility that long-term memory mechanisms can underlie memory retention within short-term memory tasks as well.

**Priming**

Repetition priming (henceforth priming) is a well-studied phenomenon, and is regarded as an implicit memory (for historical review see Schacter, 1987). Priming is defined as a *change in task performance* (e.g., accuracy or reaction time) of a stimulus as a result of prior experience to an identical or similar stimulus. Priming is indexed by comparing the difference between two conditions where (i) a stimulus (i.e., a prime) was presented prior to an identical (or similar) target stimulus, and (ii) where the prime and target were different. The difference between matching and mismatching prime and target is the so-called *priming effect*. As an implicit memory, priming is something we rarely notice outside of the laboratory, but most certainly something that happens most of the time. Priming can occur when information is consciously or non-consciously encoded. I will refer to the former case as consciously encoded priming, and the latter as non-consciously encoded, or simply non-conscious, priming. These two types of priming, as we shall see, differ significantly in memory strength, but it is primarily non-conscious priming that is of interest in this dissertation.

Priming is considered a separate kind of memory based on dissociations between priming and other memories, such as, episodic/semantic memory, in neurophysiological lesions (Graf, Squire, & Mandler, 1984; Warrington & Weiskrantz, 1974), as well as, psychological studies (for review see Tulving & Schacter, 1990). Cell-recordings in primates and human neuroimaging have found that priming is tightly linked to a phenomenon called repetition suppression (Henson, 2003; Schacter, Dobbins, & Schnyer, 2004). Repetition suppression is when repeated exposure to the same stimulus results in a decrease in neural activity or BOLD signal the second time the stimulus is processed. There are several suggested models that try to explain how priming (a facilitation in performance) can lead to a decrease in neural activity/BOLD signal (Grill-Spector, Henson, & Martin, 2006). However, there are also cases of repetition enhancement (Henson, 2000; Tartaglia, Mongillo, & Brunel, 2015). Complicating matters more, it has been argued that repetition suppression is not caused by priming, but a form of antipriming. Antipriming
is when the first stimulus (prime) is not followed by an identical target stimulus, but instead, a different stimulus with overlapping features (Marsolek, 2008). Antipriming is essentially the opposite of priming, that is, a decrease in performance. When comparing baseline (i.e., no or little priming) trials with primed and antiprimed trials there was no difference between baseline and primed, but an increased BOLD signal for antiprimed trials (Marsolek et al., 2010). It was therefore argued that repetition suppression results from comparing primed with antiprimed trials, since priming experiments usually lack baseline trials.

Repetition suppression has been used to map out the neural correlates of priming in the brain, and by using various material in priming experiments it has been found that different areas in the brain are involved with priming of different material (Henson, 2003; Schacter, Wig, & Stevens, 2007). For example, visual priming depends on the visual cortex (Koutstaal et al., 2001), semantic priming on the temporal cortex (Rossell, Price, & Nobre, 2003), affective priming on the amygdala (Dannlowski et al., 2007), and contextual priming on the hippocampi (Greene, Gross, Elsinger, & Rao, 2007). These material-dependent findings have been confirmed by dissociations between priming of different material, such that, lesions or atrophy (e.g., from dementia) in specific brain regions causes content-specific priming deficits, while sparing priming of other content (Schacter & Buckner, 1998).

Consciously encoded perceptual priming (e.g., visual images) are known for being very robust, and can exist many years after just a single exposure (Cave, 1997; Mitchell, 2006; van Turennout, Ellmore, & Martin, 2000). Perceptual priming is therefore mainly thought to be latent long-term changes in neural networks. However, the definition of priming is quite broad, and could equally well correspond to short-lived latent neural changes, residual or persistent neural activity. Indeed, (conscious) working memory tasks can elicit very strong priming effects (e.g., ~ 250 milliseconds) in reaction times that likely depend on persistent activity rather than latent neural changes (e.g., in Study II & III of the present dissertation), and popular models of semantic priming is posited to be residual activity because it only last for seconds (which qualifies as iconic memory, see below). Thus, priming seems to occur in almost any brain region depending on material, consist of any possible neural memory mechanism, and thus partly or wholly overlap with most other kinds of memory. It has been said that a “useful concept in science frequently is one whose definition not only makes very clear what it includes, but also what it excludes” (Tulving, p. 384, 1972). I therefore believe that priming, as currently defined, is too inclusive to be a useful concept as a specific kind of memory, and more useful as a general memory phenomenon.

Non-consciously encoded priming is somewhat weaker than its conscious counterpart. Historically, as mentioned earlier, it was assumed that non-conscious retention was the result of priming, and that these non-conscious
priming effects ceased to be detectable within 500 ms (Dehaene & Changeux, 2011; Draine & Greenwald, 1998; Greenwald, Draine, & Abrams, 1996; Mattler, 2005). However, there are examples of non-consciously encoded perceptual priming effects lasting for 15 - 20 min (Bar & Biederman, 1998, 1999), and neural repetition suppression effects without behavioral effects after 47 min (Gaillard et al., 2007). It is possible that the discrepancy in non-conscious priming longevity mirrors the difference between consciously encoded perceptual (years) and semantic (seconds) priming to some extent. Given that non-conscious perceptual priming can last minutes in some cases it is necessary to consider it as a potential explanation for short-term retention of several seconds as well.

**Episodic & semantic memory**

Episodic/semantic memory is something we make use of all the time in our everyday lives, and probably what we most commonly think of when we hear the word memory. Episodic memory is called “explicit” because we are consciously aware of the encoding event when we recall it. Semantic memory is “implicit” because it is not necessarily accompanied by conscious recollection of the encoding event, even if there might be conscious experience of knowing the semantic information. I will nonetheless treat them together here because they share many similarities.

Although episodic and semantic memory have several things in common, there is a clear distinction to be made between them (Tulving, 1972, 1983). Episodic memory refers to our memory of events that have been experienced personally, at a specific place, and at a specific point in time. Semantic memory refers to our knowledge of the world, such as, facts or language. They are, however, intimately linked in usage and cortical organization. Furthermore, they are highly dependent on subcortical structures in the medial temporal lobe, primarily the hippocampi structures. Lesions to the medial temporal lobes can cause complete episodic and severe semantic anterograde amnesia (i.e., the inability to form new long-term memories), and to some extent retrograde amnesia (i.e., the inability to recall previous long-term memories), but leaves intelligence, short-term, and implicit memory largely intact, as demonstrated by the famous amnesia patients H.M. (Corkin, 2002; Scoville & Milner, 1957), and K.C. (Rosenbaum et al., 2005). H.M.’s lesions caused partial retrograde amnesia in semantic and episodic memory, but K.C.’s lesions caused complete episodic retrograde amnesia, while sparing semantic memories. H.M.’s episodic retrograde amnesia only seemed to stretch 5 years back from the operation initially, but it might have fully degraded over time (Corkin, 2002; Rosenbaum et al., 2005). Interestingly, it was later shown that H.M., and K.C. had learned new semantic information, albeit extremely slowly across many repetitions, without medial temporal

Human neuroimaging studies have found that activity in the hippocampi during encoding predicts subsequent recollection performance (Danker, Tompary, & Davachi, 2016; Eldridge, Knowlton, Furmanski, Bookheimer, & Engel, 2000; Nyberg, McIntosh, Houle, Nilsson, & Tulving, 1996). The medial temporal structures are therefore clearly essential for normal episodic/semantic memory function, but so is the neocortex.

Episodic memories are stored in widely distributed networks across the neocortex that connects various sensory information, time, and place throughout the neocortex (Cabeza & Nyberg, 2000; Fuster, 1995, 2009; Habib, Nyberg, & Tulving, 2003; Nyberg et al., 1997; Nyberg, Habib, McIntosh, & Tulving, 2000; Nyberg, McIntosh, Cabeza, et al., 1996). Different parts of the neocortex therefore contribute to different aspects or contents of episodic memories. For example, visual information is stored in the visual cortex, auditory information in the auditory cortex, spatial information in the parietal cortex, and temporal information relies on the prefrontal cortex. Semantic memory is mainly stored left lateralized in the temporal, parietal, but also in the frontal cortex (Binder, Desai, Graves, & Conant, 2009; Binder & Desai, 2011; Cabeza & Nyberg, 2000; Fuster, 1995, 2009; Mårtensson et al., 2012). Severe lesions to specific cortical regions therefore renders specific content (e.g., faces, names, when or where events took place) of those specific regions unusable, but does not affect other aspects of episodic/semantic memory.

The widely distributed networks in the neocortex form their connections relatively slowly, and it is largely believed that the hippocampi, which form new connections relatively quickly, holds the cortical networks together until they become strong enough for independence. However, there is still considerable debate about the specifics of the interaction between the medial temporal lobe and neocortex (Nadel & Moscovitch, 1997; Squire, Kosslyn, Zola-Morgan, Haist, & Musen, 1992; Winocur & Moscovitch, 2011).

As previously mentioned, it was traditionally assumed that episodic (and by some) semantic memory always was consciously recollected, and that the hippocampi, which these structures heavily depended on, also were linked to consciousness (Squire & Zola-Morgan, 1988; Squire & Dede, 2015; Tulving, 2002). However, this view was challenged by a series of recent studies from Henke’s lab that showed hippocampi-based retention of non-consciously encoded semantic associations (Degonda et al., 2005; Duss et al., 2014; Duss, Oggier, Reber, & Henke, 2011; Henke, Mondadori, et al., 2003; Henke, Treyer, et al., 2003; Reber, Luechinger, Boesiger, & Henke, 2012; Reber & Henke, 2011). These findings are consistent with semantic memory being an implicit memory (Tulving, 2002).
For example, it was found that non-conscious semantic associative learning engaged the hippocampi, and interacted with conscious memory (Degonda et al., 2005). Reber et al. (2012) showed that BOLD signal change in the hippocampi during non-conscious encoding predicted memory performance 1 minute later. In a study with a patient group with hippocampi lesions (amnesiacs), and healthy controls, Duss et al. (2014) found that only healthy controls could retain word pair associations (semantic memory), but both groups could retain a single non-conscious word (semantic priming). These findings show that hippocampi structures are necessary for semantic associative learning with or without conscious experience. Taken together, these, alongside other findings, have been used to criticize the traditional long-term memory taxonomy based on conscious recollection (Henke, 2010).

Semantic memory can thus be encoded by non-consciously presented information. However, the same cannot be said for episodic memory. Partly because conscious recollection is a part of the definition of episodic memory, and it seems unlikely to have a conscious recollection of something non-consciously encoded. Although, if a non-conscious memory fulfills all criteria of episodic memory except conscious recollection, I am inclined to think that the definition of episodic memory needs a revision. As far as I know, nothing like a non-conscious episodic memory has been shown to exist. However, that does not exclude its existence, and I will therefore use non-conscious hippocampi-based memory to refer to non-conscious semantic memory and/or memory of “episodic” information. Non-conscious hippocampi-based memory therefore needs to be considered as a potential explanation to short-term retention.

Sensory memory
Experiments prior to 1960 usually presented participants with brief (100 milliseconds) arrays of up to 20 letters, and found that participants only could remember a few letters (full-report paradigms). However, Sperling (1960) later added a tone after an array of three rows with four letters each presented for 50 milliseconds. The tone was presented after stimuli offset, and used to cue which row the participants had to report letters from (high, medium, low). Sperling found that participants could correctly recall about four letters from any cued row (partial-report paradigm). His findings show that it was possible to recall much more than a few letters, but only for a brief time, otherwise the memory had faded about the time it took to report four letters. Furthermore, it was found that this sensory memory had a capacity up to 16 out of 18 letters, but gradually decayed rapidly to an asymptote of four letters within 500 ms to 2 s after stimuli offset (depending on brightness), and that if other stimuli were subsequently presented at the same spatial position the memory would be overwritten (Averbach & Coriell, 1961; Averbach & Sperling, 1961). This brief visual sensory memory was later named **iconic memory** (Neisser, 1967).
Iconic memory was defined as the retention of information commonly measured by the partial-report paradigms (Averbach & Sperling, 1961; Coltheart, 1980; Neisser, 1967).

Iconic memory was traditionally thought to consist of three components, visible persistence (i.e., persisting conscious experience), neural persistence (i.e., residual neural activity), and information persistence (i.e., iconic memory). It was furthermore assumed that these three components described the same phenomenon at different levels. However, it has aptly been argued, that visual persistence, which shows inverse relationships with stimuli duration and luminance (if the duration is brief), can be dissociated from iconic memory, which does not show inverse relationships (Coltheart, 1980). Coltheart thinks it is likely that photoreceptor persistence in the retina can explain the inverse relationship with luminance, and residual activity in the LGN might explain the inverse relationship with stimuli duration, but residual cortical activity is likely also necessary. It is less clear how the neural substrates of iconic memory differ from visible persistence, since both must rely on the visual system. Nevertheless, since visible persistence can be dissociated from iconic memory, it seems plausible that there can be such a thing as non-conscious iconic memory. However, if cued prior to its complete disappearance, the information can be maintained within the limited capacity of working memory until probed. It might be possible that techniques can be used to render iconic memory non-conscious from the beginning without losing all the information.

In a study Sligte, Scholte, & Lamme (2008) proposed the existence of a new fragile visual short-term memory based on a dissociation from iconic memory and working memory, which by some have been criticized for actually being working memory (Makovski, 2012). However, I will briefly explain why I think they mistakenly confused “iconic memory” with visible persistence, and “fragile memory” with iconic memory. Sligte et al. (2008) are essentially using a modified partial-report paradigm with a retro-cue after 1 s, which they argue is the reason why it is not iconic memory (because of iconic memory’s rapid decay). Firstly, however, they are by (Sperling’s, Neisser’s, and Coltheart’s) definition, measuring iconic memory (i.e., the informational persistence), thus any memory effect is by definition an iconic memory effect. Secondly, they have a black or grey background after stimuli offset; according to Averbach & Sperling (1961) a black context can make iconic memory persist for more than 2 seconds. Thirdly, there are other differences that could affect memory retention, such as stimuli duration (five times longer than Sperling’s duration), the delayed match-to-sample task (recognition instead of free recall), and less complex stimuli (orientation of rectangles instead of letters), could facilitate performance. Thirdly, they show that fragile memory is not affected by isoluminant (same luminance but different colors) stimuli and masks of uniform bright color, but is erased by masks with irrelevant stimuli.
at same spatial position, which is all consistent with iconic memory (Averbach & Coriell, 1961; Coltheart, 1980). Fourthly, their “iconic memory” is affected by isoluminant stimuli and bright masks of uniform color, which is to be expected of visible persistence (Coltheart, 1980). The novelty, I think, is that they, in their fourth experiment, show that “fragile memory” (i.e., iconic memory) can last up to 4 seconds (with gradually reducing capacity over time) in some circumstances. I will therefore interpret subsequent studies of “fragile memory” as showing that iconic memory performance correlates with BOLD signal change in visual cortex (V4; Sligte, Scholte, & Lamme, 2009), does not depend on dorsolateral prefrontal cortex (Sligte, Wokke, Tesselaar, Scholte, & Lamme, 2011), and is erased only if irrelevant masks are of same spatial location and object type (Pinto, Sligte, Shapiro, & Lamme, 2013).

It therefore seems like the neural substrate of iconic memory is limited to residual neural activity in higher visual cortical areas, but it is unclear how it differs from visible persistence. Another implication is that backward masking might erase visible persistence, but not necessarily iconic memory, which means that non-conscious iconic memory is plausible, and might even explain non-conscious semantic priming effects. It might be the case, but has yet to be investigated, that non-consciously presented information can be retained in a rapidly decaying high capacity iconic memory that can last for a few seconds.

**Working memory**
Working memory is something we use daily when we reason, make decisions, and perform tasks in general. It is for example used to keep a phone number in mind until the number have been dialed, when doing arithmetic in the head, or when planning a future decision. It is known for having a relatively small capacity limit (Brady, Konkle, & Alvarez, 2011; Keisuke Fukuda, Awh, & Vogel, 2010), and is a reliable indicator of intelligence (Fukuda, Vogel, & Mayr, 2010; Unsworth, Fukuda, Awh, & Vogel, 2014). There might not be one definition of working memory that all agree on, but I will use the following: Working memory is “...the temporary retention of information - sensory or other – for the performance of a prospective act to solve a problem or attain a goal” (Fuster, p. 144, 2015). That the retention is for a prospective action is crucial, as it is the defining feature that separates working memory from sensory memory. Neurophysiological investigations have revealed that working memory is an emergent property from the interaction between long-term (phylogenic and ontogenetic) memory and sustained endogenous attention (for reviews see Eriksson, Vogel, Lansner, Bergström, & Nyberg, 2015; Fuster, 1995; Postle, 2006; Sreenivasan, Curtis, & D’Esposito, 2014). That is, the temporary activation of neural networks until a prospective act is complete. Similarly to long-term memory and perception, different brain regions represent different aspects of working memory, and partly depend on the specific task and material used (for meta analyses see Fuster, 2009; Nee et al.,
It is well established that the prefrontal cortex is important for working memory from studies with cell-recordings in primates (Funahashi, Bruce, & Goldman-Rakic, 1989; Fuster & Alexander, 1971; Goldman-Rakic, 1995), human neuroimaging (Courtney, Ungerleider, Keil, & Haxby, 1997; Postle & D’Esposito, 1999), lesions (D’Esposito & Postle, 1999), electrical, or magnetic stimulations (Brunoni & Vanderhasselt, 2014; Feredoes, Heinen, Weiskopf, Ruff, & Driver, 2011).

The role of the prefrontal cortex seems to be maintaining task-related information (Pochon et al., 2001; Rowe et al., 2007; Sakai & Passingham, 2003), and protect the remembered information from distractions (Feredoes et al., 2011; Sakai, Rowe, & Passingham, 2002). The prefrontal cortex seems to be involved no matter what kind of information that is being retained, but with slight variations in cortical location (Nee et al., 2013; Wager & Smith, 2003). For example, the frontal eye fields (also involved in coordinating eye movement) are involved in maintenance of visuospatial information (Courtney, Petit, Maisog, Ungerleider, & Haxby, 1998), while the left inferior frontal cortex (also involved in speech production) is involved in maintenance of verbal information (Paulesu, Frith, & Frackowiak, 1993), but the most anterior prefrontal cortex seems less dependent of material.

The posterior cortex is also important, but is much more dependent on the information held in working memory. That is, the visual cortex will be active when visual information is being maintained, the auditory cortex for sound, parietal cortex for spatial information, etc. (Druzgal & D’Esposito, 2001; Emrich, Riggall, Larocque, & Postle, 2013; Fuster & Jervey, 1981; Fuster, 2009; Harrison & Tong, 2009; Miller & Desimone, 1994; Miyashita & Chang, 1988; Postle, Stern, Rosen, & Corkin, 2000; Pratte & Tong, 2014). Locations of lesions therefore differentially affect working memory depending on location/content (Owen et al., 1996; Pisella, Berberovic, & Mattingley, 2004). Additionally, the parietal cortex seems involved in more executive aspects of working memory (Collette et al., 2005; Koenigs, Barby, Postle, & Grafman, 2009), and have been associated with working memory capacity (Riggall & Postle, 2012), but the latter could also reflect spatial strategies to simultaneously maintain many items.

These prefrontal and posterior brain areas seem to causally depend on each other for a normally functioning working memory. For example, cortical cooling of either prefrontal or stimulus-specific regions in primates cause a decrease in neural activity in the other region, together with a drop in working memory performance (Fuster, Bauer, & Jervey, 1985). Other brain areas relevant for working memory are the hippocampi (Libby, Hannula, & Ranganath, 2014; Ranganath & D’Esposito, 2001), basal ganglia (Cools & D’Esposito, 2011; Ekman et al., 2012; Wager & Smith, 2003), and cerebellum (Desmond, Chen, & Shieh, 2005; Desmond & Fiez, 1998; Stoodley & Schmahmann, 2009).
It should be noted that all the neurophysiological findings mentioned above report persistent neural activity during working memory maintenance. However, recent computational models and empirical findings suggest that the “persistent” part of the neural activity might partly or sometimes be an artifact created from averaging data over trials, and that short-lived latent neural changes instead might play an important role to maintain the memory trace between activity (Barak & Tsodyks, 2014; D’Esposito & Postle, 2015; Eriksson et al., 2015; Lundqvist, Herman, & Lansner, 2011; Mongillo, Barak, & Tsodyks, 2008; O’Reilly, Braver, & Cohen, 1999; Sprague, Ester, & Serences, 2016). These models might produce a more stable and economical working memory maintenance. By using trial-by-trial analyses, it has been argued that the intensive neural activity that occurs during perception encodes the information in short-lived latent (structural) changes, which is then retained by sporadic bursts of neural activity after stimulus offset, and again intensifies during the use of the information (Lundqvist et al., 2015).

Sprague et al. (2016) found that the fidelity of working memory content decreased when more information was maintained, but that a retro-cue indicating that only a subset of the maintained information was relevant retroactively increased the fidelity of the relevant information. In addition, the retroactive increase in fidelity correlated with memory performance. The recovery in fidelity is inconsistent with pure spiking models because they predict that lost fidelity cannot be recovered. It was therefore concluded that some latent neural change must have facilitated the recovery in fidelity/neural activity. Lewis-Peacock, Drysdale, Oberauer, & Postle (2012) used multivoxel pattern classification analysis of to decode the BOLD signal from prioritized and deprioritized information during a working memory task. Interestingly, they could only decode the prioritized, but not the deprioritized information. The deprioritized information was nonetheless remembered when later probed (on half of the trials). These results was later replicated with electroencephalography (EEG; LaRocque, Lewis-Peacock, Drysdale, Oberauer, & Postle, 2013). It might therefore be the case that working memory more accurately is an emergent property based on the interaction between long-term (phylogenetic and ontogenetic), short-lived (ontogenetic) memory, and sustained endogenous attention.

Briefly summarizing and integrating these neurophysiological findings about working memory with the framework espoused in the building blocks of cognition section leaves us with the following view. Presented stimuli are processed in stimulus-specific sensory areas when perceived. Stimulus-specific information is then retained by persistent neural activity caused by recurrent loops between stimulus-specific areas, and prefrontal action-related areas within the perception-action cycle. Persistent neural activity in posterior areas mainly relate to the information to be remembered, in the prefrontal cortex it relates to the maintenance of the prospective task (e.g., goal and
plan), and coordination of attentional resources (endogenous attentional control). The activity gradually shifts towards the primary motor cortex as the plan and motor action is implemented, and after task completion the neural activity subsides.

William James (1890) made a distinction between primary and secondary memory based on phenomenological grounds. Primary memory was consciously experienced, while secondary memory was not. The relationship between conscious experience and working memory carried over to Atkinson & Shiffrin’s (1968, 1971) model, which contained a sensory register (iconic memory), a short-term store that was conscious (working memory), and a long term store. Psychological evidence for multiple memory stores led Baddeley & Hitch (1974) to postulate a multi-store model, with a “central executive” that controls and coordinates information in three short-term stores (visuospatial sketchpad, phonological loop, and later episodic buffer; Baddeley, 1983, 2003, 2012). Baddeley’s (1992, 2000, 2003) model initially assumed that all working memory operations, input and output, was conscious, but later introduced an episodic buffer that was similar to Baars (1988, 2005) global workspace, and integrated information from other stores into conscious episodes. Baars & Franklin (2003) thought all active working memory operations were consciously experienced.

These early systems-based working memory models posited separate short-term and long-term systems or stores based on neurophysiological evidence that lesions can affect long-term, but not short-term memory, and vice versa (Scoville & Milner, 1957; Shallice & Warrington, 1970). However, later neurophysiological findings showed that the brain regions involved in perception, working memory, and long-term memory of a certain informational content overlap (Fuster, 2003; Lewis-Peacock & Postle, 2008). Contemporary (state-based) models therefore posit that the distinction between short-term and long-term memory depends on the state of the memory (active or latent). There are several state-based models that differ in the details about the number of states possible, their capacity, and neural substrates, but most seem to agree that information in the focus of attention is consciously experienced, and consists of persistent neural activity (Cowan, 1995; Fuster, 1995; Jonides et al., 2008; McElree, 2001; Oberauer, 2002).

Cowan (1995, 2008) and Fuster (1995, 2003a) conceives an intermediate state of an active portion of long-term memory outside the focus of attention, which is non-conscious, and presumably consist of less neural activity. McElree (2001; 2006) thinks there is only latent non-conscious memory with varying synaptic weights outside a very narrow focus of attention. Oberauer (2001, 2002, 2005) posits a very narrow focus of attention, a state of direct access or broad focus of attention (like Cowan’s focus of attention), and active long-term memory. However, Oberauer’s model is neutral about conscious experience. Support for these postulated differences in states comes from
different psychological experiments, and neuroimaging findings are still indecisive as there has been support for both McElree’s (Oztekin, Davachi, & McElree, 2010; Oztekin, McElree, Staresina, & Davachi, 2009), and Oberauer’s (Nee & Jonides, 2011, 2013) model.

Despite the historically intimate relationship with consciousness, current working memory models leave some theoretical room for non-conscious states even if that aspect seldom is in focus. Furthermore, empirical studies have shown that non-consciously presented information can engage endogenous attention in the form of cognitive control (Lau & Passingham, 2007; Reuss et al., 2015, 2011; van Gaal et al., 2010). Thus it has been shown that executive functions needed for working memory maintenance can operate on non-consciously presented information. It is thus plausible that working memory can operate on non-consciously presented information as well.

Hassin, Bargh, Engell, & McCulloch (2009) have made a case for implicit working memory, that is, non-conscious working memory operations on consciously presented information. They sequentially presented five disks (full or empty) in a grid matrix. The five disks could be in (i) a predetermined sequence/pattern, (ii) a “broken pattern” where the first four disks followed the predetermined pattern, but the last disk did not, and (iii) “control patterns” where the first three disks were randomly placed, and the last two disks followed the pattern. Participants did not know about the pattern, and was supposed to decide if the disks were full or empty. As predicted they found that reaction times were faster for predetermined patterns compared to broken patterns, and concluded that it was implicit working memory. However, it is difficult to know if this really was implicit working memory or implicit sequence learning based on basal ganglia (or both). The reaction times are also difficult to interpret since a closer look shows that reaction times for determined patterns was the same as for control patterns, but the broken pattern was slower than the other two patterns. Thus, there was no facilitated reaction time as predicted, only a slower reaction time from the broken patterns.

There have been two early attempts at exploring working memory maintenance of non-consciously presented information (Bergström, 2011, Bachelor’s thesis, not peer-reviewed; Soto, Mäntylä, & Silvanto, 2011). I used an attentional blink paradigm similar to Study I (see method section) to render target stimuli (single letters) non-conscious followed by a variable delay period (150 milliseconds, 3 seconds, and 7 seconds), and found that participants performed better than chance after all delay durations, challenging the assumption that non-conscious retention is extremely short-lived and fickle.

Soto, Mäntylä, & Silvanto (2011) used a backward masking paradigm to render a target stimuli (Gabor gratings) with different orientations non-
conscious. After a short delay (2 or 5 s with distractors) a test cue appeared. The test cue was also a Gabor grating, but with a different orientation than the target stimulus. The participants had to guess if the target stimuli orientation was clock-wise or counter clock-wise in respect to the test cue. They found above-chance performance despite no reported visual experience. These findings were interpreted as non-conscious working memory because the test cue’s orientation never matched the target stimulus (excluding repetition priming), and since the distractors did not erase the non-conscious memory trace (excluding iconic memory). Their findings were compelling and consistent with (Bergström, 2011). Together these findings provided interesting results for continued investigations into the existence of non-conscious short-term retention.
Aim

The overall aim of the empirical studies of this dissertation was to investigate short-term retention of non-consciously presented information to better understand its functional properties (i.e., longevity and content), neural substrate (using fMRI), and relate the findings to current models of memory (e.g., working memory), and theories of consciousness. More specifically, the aims of the three studies were to:

Study I
- Find support for durable retention (up to 15 seconds) based on subjective criterion of experience
- Estimate the longevity of such short-term retention
- Investigate the neural substrates of the retention

Study II
- Replicate longevity findings with new method to render stimuli non-conscious based on objective criterion of experience
- Investigate the specificity of the retained content (identity & position)

Study III
- Further investigate the neural substrates of the retention in relation to:
  - (i) prefrontal task-related cortical regions
  - (ii) posterior stimulus-related cortical regions
Materials and methods

This section will give an overview of the varying techniques and methodologies used in the three studies. It will also provide some of the reasoning and underlying assumptions behind using them.

Participants
For all studies healthy young participants were recruited from the Umeå University campus area. All participants had normal, or corrected to normal, vision, gave written informed consent, and was paid for their participation. All studies that used fMRI were approved by the ethics committee of the University Hospital of Northern Sweden. For more details concerning inclusion/exclusion see the appended articles.

Study I contained two experiments. Experiment I analyzed 24 participants. Experiment II consisted of two sessions (pre-fMRI and fMRI session), and analyzed 27 participants. Study II contained two experiments. Experiment I analyzed 15 participants, and Experiment II 16 participants. Study III consisted of one experiment with two sessions. The pre-fMRI session analyzed 25 participants, and the fMRI session 26 participants.

Inferring the absence of conscious experience
There are several ways of measuring conscious experience, and infer non-conscious perception. These different measures can be categorized into two main approaches; subjective and objective approaches. Subjective approaches define conscious experience as reported conscious experience. If participants therefore report “no conscious experience” of the target stimuli, any subsequent behavioral or neuroimaging effects are attributed to non-conscious processes. The main advantage with using a subjective approach is that you are more directly measuring the phenomenon of interest, since conscious experience is subjective by nature. Another advantage is the possibility to index different degrees of conscious experience and content. However, subjective measures risk not being completely exhaustive, that is, they might not detect all relevant conscious experience, and can therefore overestimate non-conscious effects (Merikle & Joordens, 1997; Reingold & Merikle, 1988; Snodgrass & Shevrin, 2006).

The objective approach relies on objective task performance instead of subjective reports, and define conscious experience as above-chance performance on a direct (i.e., instructed) task (e.g., discrimination accuracy). The absence of conscious experience is therefore inferred when the direct task performance is at chance, and only then can indirect measures (e.g., reaction time or BOLD signal change) be attributed to non-conscious processes. The main advantage with objective measures is that any detectable non-conscious
effect is very unlikely to be contaminated with weak conscious experience. However, objective measures might not exclusively measure conscious effects, and therefore risk underestimating non-conscious effects (Lau, 2008; Reingold & Merikle, 1988).

Historically, subjective measures were first used to investigate non-conscious processing (Pierce & Jastrow, 1884; Sidis, 1898; Stroh et al., 1908), but later suffered severe criticism (Eriksen, 1960), which gave rise to the use of objective measures. However, after methodological advances (and the shift from behaviorism to cognitive science) it is now generally accepted that both are valid approaches to measure the presence or absence of conscious experience. Indeed, it has been argued that an objective approach can be seen as a more conservative version of the subjective approach since they both produce similar and reliable results (Dehaene & Changeux, 2011; Merikle, Smilek, & Eastwood, 2001). There is currently no consensus on which measure is the most exhaustive and exclusive (Boly et al., 2013), and they can be seen as complementary.

Sandberg, Timmermans, Overgaard, & Cleeremans (2010) compared two subjective and one objective measure of conscious experience. Firstly, the perceptual awareness scale (PAS), in which participants are instructed to report their perceptual experience of the target stimuli ranging from no experience to clear experience (Ramsøy & Overgaard, 2004). Secondly, confidence ratings, where participants are instructed to report how confident they were of seeing or guessing correct target stimulus ranging from no confidence to high confidence (Cheesman & Merikle, 1986). Thirdly, an objective measure called post-decision wagering where participants are instructed to bet small monetary sums on their task accuracy ranging from a smaller to a larger monetary sum (Persaud, McLeod, & Cowey, 2007). They found that the PAS was more exhaustive, and thereby more conservative, than confidence ratings and post-decision wagering (Sandberg et al., 2010). In addition, Sandberg, Del Pin, Bibby, & Overgaard (2014) found that the PAS seems to be more exclusive than exclusion tasks (Debner & Jacoby, 1994). Another study found that confidence ratings were more exhaustive than post-decision wagering because the latter was affected by loss aversion (Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008).

Going forward we chose to use the PAS to measure conscious experience. We did so for three reasons. Firstly, since we intended to investigate non-conscious short-term memory, a fragile phenomenon that most assumed did not exist, we thought it prudent to begin with a liberal approach. If the liberal approach would yield memory effects we could try to replicate it with more conservative measures. Secondly, we chose subjective measures because they best catch conscious experience. Thirdly, we chose the PAS because it seemed the most conservative and reliable of the subjective measures. After settling
on how to measure conscious experience, we had to decide what technique we would use to make the target stimuli non-conscious.

**Trial procedures and stimuli**

There are various techniques available to render information non-conscious. These techniques all have in common that they, in different ways and to varying degrees, manipulate attention to make the target stimulus seemingly invisible. They usually manipulate either exogenous or endogenous attention. Techniques that manipulate endogenous attention tend to lead to more extensive non-conscious processing (Kouider & Dehaene, 2007). However, there is no technique that is superior to all other, they all have different strengths and weaknesses (for a review see Kim & Blake, 2005), which makes them more or less suitable for different purposes.

**Study I**

For the purpose of investigating the retention of non-consciously perceived information we wanted a technique that maximized non-conscious perceptual processing, yet was reliable enough to work on several consecutive trials even when the participants were aware of the manipulation. Here the underlying assumption was that stronger non-conscious perceptual processing would translate into longer and more robust retention. To that end, we decided to use an *attentional blink* paradigm for Study I (Figure 1). The attentional blink phenomenon is known to exhibit relatively strong non-conscious effects (Kouider & Dehaene, 2007; Luck, Vogel, & Shapiro, 1996; Marois, Yi, & Chun, 2004; Sergent, Baillet, & Dehaene, 2005).

To achieve an attentional blink two targets are presented within succession in a rapid serial visual stream of irrelevant stimuli (distractors). When endogenous attention is used to identify the first target (T1) there is subsequently an attentional blink (or gap) such that the second target (T2) is rendered non-conscious (Duncan, Ward, & Shapiro, 1994; Marois, Chun, & Gore, 2000; Raymond, Shapiro, & Arnell, 1992). However, T2 will only become non-conscious when there is a time-lag in the order of a few hundred milliseconds between T1 and T2 (short lag). If the time-lag is too short or too long there will be a reduced attentional blink. For baseline comparisons we also included trials without T2 (replaced by another distractor).

Previous literature suggested that more cognitively demanding T1 tasks generate stronger attentional blink effects (Martens & Wyble, 2010). We therefore made some modifications to the traditional attentional blink paradigm, in which T1 usually is a simple identification task (e.g., even or uneven number). In our attentional blink paradigm, T1 was a simple math task (e.g., 5+3 in red color) to be solved immediately, and remembered until the end of the trial. The T1 math task enabled us to present T2 (a single letter; A, S, D, or F; flanked by two irrelevant digits) for longer durations than
traditional attentional blink paradigms without being consciously experienced. Here we assumed that longer presentation durations would lead to more durable retention of the non-consciously presented stimuli. However, since the attentional blink worked differently on different participants, the presentation durations were adjusted during the experiment based on how many T2 had been visually experienced. The algorithm decreased or increased the presentation duration so that the participant only would experience approximately 50% of the short-lag trials.

To estimate the longevity of non-conscious retention in Experiment I we had a variable (distractor-filled) delay of 5, 10, or 15 s between T2 presentation and T2 response (Figure 1A). From these data points our plan was to extrapolate a potential decline in performance over time, and thereby estimate when it would reach chance level (if it had not already done so within the three time points). In some trials we used long time lags between T1 and T2 to make T2 visible (figure 1B). These long-lag trials had different lag length such that T2 was presented early, intermediately, or late in the trial. The participants’ could therefore not know whether they had missed an early T2 or if it was yet to come, and therefore had to attend the rapid serial visual presentation until its end.

After the delay, participants were shown a screen with all four letters, during which time they reported the letter that had been presented previously that trial. If participants did not consciously experience T2 they were instructed to guess by using the first alternative that came to mind or their gut-feeling. Each finger was mapped to a specific letter, making it possible to create stimulus-response bindings (Henson, Eckstein, Waszak, Frings, & Horner, 2014), which might further boost retention. Next, the participants estimated their perceptual experience of T2 using a four-point perceptual awareness scale (PAS; 1. no experience, 2. vague experience, 3. almost clear experience, 4. clear experience; Sandberg et al., 2010). Lastly, the participants’ had to report the solution to T1.

Experiment II: Pre-fMRI session had some minor changes to the protocol. The PAS was reduced to a three-point scale (1. no experience, 2. vague experience, 3. clear or almost clear experience). The new three-point scale was judged to be easier to use without sacrificing the critical distinction between “no experience” and “vague experience”. We also increased the T1 math task difficulty slightly to facilitate stronger attentional blink effects. The pre-fMRI session was primarily used to screen for unsuitable participants (e.g., non-blinkers that are immune to the attentional blink; Martens & Wyble, 2010), and calibrate the individual presentation durations, prior to the fMRI session (though presentation durations were also adjusted continuously during the fMRI session). The fMRI session had some additional changes (figure 1C). Inter trial intervals (ITIs) and inter stimulus intervals (ISIs) were jittered to reduce correlations between components in the statistical model. The delay
period only had distractors for a short time after T2 to maintain the attentional blink phenomenon, but otherwise consisted of a dot to be passively viewed. Removing the distractors was thought to make the stimulus-representation less noisy in terms of BOLD signal. However, in hindsight, it might have been preferable to leave them in for methodological reasons, and to strongly activate the prefrontal cortex in its role to protect the memoranda against distraction (Feredoes et al., 2011).

Figure 1. Illustration depicting the procedure for experiment 1 and 2. A math task (T1) and a letter (T2) flanked by distracters were presented in a rapid serial visual presentation (RSVP). The solution to T1, and T2 identity, were held in memory during a variable delay until probed for responses. (A) Short- and (B) long-lag trials in experiment 1 and the pre-fMRI session. In (B) the variable delays before and after T2 presentation were adjusted such that T2 appeared early, in the middle, or at the end of the RSVP. (C) Illustration of a short-lag trial in the fMRI session.

**Study II**

For Study II we chose to use a different technique to render stimuli non-conscious, namely, continuous flash suppression (CFS; Tsuchiya & Koch, 2005). We changed from the attentional blink to CFS for several reasons. The main strength with CFS is its strong suppression. The suppression leads to fewer ambiguities as to whether something was experienced or not, and stimuli can be presented for longer durations (more than a minute; Tsuchiya & Koch, 2005). Its main weakness is that the suppression might be too strong, and therefore limit non-conscious effects. We reasoned that we might be able to offset some of the lost effect by using relatively long presentation durations (seconds instead of milliseconds), and that we could use an objective instead of subjective criterion of experience.
CFS is based on the phenomenon bistable perception. Bistable perception is when the conscious experience switches between two mutually exclusive perceptual states, as is the case when perceiving ambiguous figures or during binocular rivalry when two completely different static images are simultaneously presented to each eye.

During CFS the visual input from the left side of the monitor was presented in isolation to the left eye, and vice versa for the right eye, by using a mirror stereoscope. A non-salient target stimulus is presented to the non-dominant eye, while colorful moving geometrical figures named mondrians (after the Dutch painter Piet Mondrian) are presented to the dominant eye. Because the mondrians are much more salient and dynamic, they will consequently capture exogenous attention, and thereby suppress the target stimulus from consciousness. In addition to the presentation condition with suppressed (non-conscious) target stimuli, there were also baseline trials without target stimuli where mondrians suppressed a blank grey background, and conscious trials where the target stimulus was presented to both eyes (without mondrians). We used CFS in combination with a common short-term memory task called delayed match-to-sample (DMS) task (figure 2). In the DMS task participants are shown a target stimulus to be remembered for a short delay-period until they are shown a probe. The probe can either match the memorandum or not, and the participants were instructed to make a match/non-match discrimination. A detection task followed where participants reported if there was any stimulus at all or not, and, similarly to Study I, a three-point PAS.

In Experiment I the participants were shown emotional faces in different spatial quadrants to be remembered for 5 or 15 s delays (figure 2A). The face sample either matched in face identity, emotion, and spatial position, or none of them. However, the participants were only instructed to remember the spatial position. We reasoned that spatial position had the highest likelihood of being non-consciously retained, but matching emotion and identity might facilitate retention, making it easier to detect any non-conscious effects.

Given the success of Experiment I (which could not separate the distinct information components), we intended to further probe the specificity of the information retained non-consciously in Experiment II. In Experiment II the participants were therefore shown tools in different spatial quadrants to be remembered for a 5 s delay (figure 2B). We decided to use tools instead of faces because their distinct shapes/features should facilitate better within-category discrimination, and previous studies indicate that tools might be processed to a larger extent during CFS (Almeida, Mahon, Nakayama, & Caramazza, 2008; Fang & He, 2005). The participants were instructed to remember the identity and position. The probe either matched identity and position, only identity, only position, or neither. By having a probe that could match only one information component, we could later determine whether
one of the components was driving the memory performance. Lastly, we decided to take away the detection task to make the trials shorter since there was no differential effect between the discrimination and detection task in Experiment I.

Study III
For Study III we used a very similar experimental protocol to Experiment II in Study II (figure 2B). Here we wanted to investigate the neural correlates underlying the non-conscious retention in Study II. The only differences were that we added the detection task, and the fMRI session had a delay duration ranging from 5 – 15 s and jittered ITI (figure 2).

Functional magnetic resonance imaging (fMRI)
Magnetic resonance imaging (MRI) and functional magnetic resonance imaging (fMRI) are relatively new neuroimaging techniques. After the basis for all modern MRI, nuclear magnetic resonance, was discovered in the 1940s (Bloch, 1946; Purcell, Torrey, & Pound, 1946) many non-biological MRI studies followed. It was not until the 1970s that the first biological MR image
was produced (Lauterbur, 1973), and after considerable improvement in MR image acquisition speed (Mansfield, 1977) that MRI became popular to image the anatomical structure of the brain. The birth of fMRI occurred in the 1990s when it was discovered that changes in blood oxygenation could be measured with MRI (Ogawa, Lee, Kay, & Tank, 1990; Ogawa et al., 1992). Today fMRI is one of the most popular and widely used techniques to investigate brain function. Its popularity stems mainly from being a non-invasive neuroimaging technique with high spatial resolution.

MRI makes use of very strong magnetic fields to create images of biological tissue. It is now common for MR scanners to have static magnetic fields of 3 Tesla or higher. For comparison, the magnets used to lift cars at scrap yards have approximately 1 Tesla. The static magnetic field aligns the atomic nuclei in the body to the magnetic field in a state of equilibrium (usually hydrogen atoms in water and fat because of its prevalence in the body). Radiofrequency coils then transmits electromagnetic waves into the body, which excites the atomic nuclei. The excited atomic nuclei absorb the electromagnetic energy from the radiofrequency coils causing some of them to change from a low-energy state to a high-energy state, and thereby change the equilibrium. When the radiofrequency coil stops transmitting electromagnetic waves, the atomic nuclei reverts to equilibrium (relaxation). As the atomic nuclei revert to low-energy states again they release the absorbed energy, which is received by the radiofrequency coils. The received electromagnetic energy is the raw MR signal. However, the raw MR signal has no spatial information. Gradient coils therefore superimpose magnetic gradients on the static magnetic field to modify its strength along the x, y, and z axis, making it possible to create three-dimensional images (for a primer on fMRI see Huettel, Song, & McCarthy, 2004).

By changing the parameters of excitation and relaxation, it is possible to relate the MR signal to different physiological properties, and thereby create different images for different purposes. The so called T1-weighted images are used to create high-resolution images of the brains anatomical structure. Because different brain tissues (grey matter, white matter, and cerebral spinal fluid) have different densities of hydrogen atoms, the corresponding MR signal becomes different. The difference in MR signal is then used to distinguish between different tissues when creating the structural images.

The so called T2*-weighted images are used to create low resolution images of the brains blood flow as an indirect measure of neural activity. More precisely, the T2*-weighted imaging measures the concentration of oxygenated (to deoxygenated) hemoglobin over time (which changes with blood flow), and is therefore referred to as the blood-oxygenation-level dependent (BOLD) signal (Kwong et al., 1992; Ogawa et al., 1990; Ogawa et al., 1992). The paramagnetic property of deoxygenated hemoglobin disrupts the MR signal. Thus, a low concentration of oxygen leads to a weak BOLD
signal, while a high concentration of oxygen leads to a high BOLD signal. When a task is performed, the corresponding neurons will fire, the firing neurons require energy in the form of glucose and oxygen, which is delivered through the blood stream, and gives rise to the BOLD signal.

However, the blood flow is relatively slow compared to neuronal firing. After an initial dip of 1 - 2 s (Buxton, 2001) the BOLD signal peaks 4 - 6 s after a brief stimulus presentation, and then returns to baseline after 12 -20 s ending with a late undershoot (Boynton, Engel, Glover, & Heeger, 1996; Poldrack, Mumford, & Nichols, 2011). The BOLD signal’s maximum observed amplitude is approximately 5% in primary sensory regions, and 0.1 - 0.5% for other regions (Logothetis, Pauls, Augath, Trinath, & Oeltermann, 2001).

Although the BOLD signal is an indirect measure of neuronal activity, and their exact relationship remains unclear, there is nonetheless a reliable correlation between the BOLD signal and neuronal activity, which mostly relates to the neurons synaptic (input) rather than spiking (output) activity (Goense & Logothetis, 2008; Logothetis et al., 2001; Logothetis & Wandell, 2004). There are, however, some caveats regarding the BOLD signal (Logothetis, 2008). Firstly, it is rather coarse, as such a typical fMRI voxel (smallest spatial unit) of 55 mm$^3$ contains about 5.5 million neurons. Secondly, the BOLD signal cannot distinguish between top-down and bottom-up signals. Thirdly, since excitatory and inhibitory neurons both cause regional blood flow, the BOLD signal likely conflates the two.

**Univariate analysis and multi-voxel pattern analysis (MVPA)**

After the fMRI data is collected it needs to go through a fair amount of preprocessing (slice timing correction, motion correction, normalization, and smoothing) prior to statistical analyses (for more on preprocessing see Huettel et al., 2004; Poldrack et al., 2011). The traditional way to analyze fMRI data is referred to as univariate (or “uni-voxel”) analysis, and essentially means that each voxel is analyzed independently from the other voxels. Instead, multivariate (or “multi-voxel”) analyses use the combined information from more than one voxel.

In short, the univariate analysis makes use of the general linear model (GLM) as a framework to do a multiple regression. The modelled BOLD signal time-series for each cognitive task are used to predict the collected fMRI data (the actual BOLD signal time series for all voxels) one voxel at the time. How good each modelled response fits the actual BOLD signal time-series in a voxel will then be calculated as parameter estimates (or beta values). For example, if a participant is randomly presented with visual or auditory stimuli over time, voxels in the visual cortex will have high beta values when related to the modelled time-series for the visual stimuli, and low beta values when related to the modelled time-series for the auditory stimuli, and vice versa for voxels in the auditory cortex. The beta values are converted to useful statistics, for
example, T values or Z values. These statistical values can then be used to infer whether the beta values of voxels are significantly different from zero, or beta values of another task (e.g., by subtracting the statistical values related to a visual task from an auditory task or a passive baseline task). Thus, making it possible to infer which voxels are associated with a certain task relative to another task (Lindquist, 2008; Smith, 2004).

There are many different kinds of multi-voxel pattern analyses (see Haxby, Connolly, & Guntupalli, 2014), but I will only focus on the particular one used in this dissertation (Study III), which uses machine learning algorithms that are trained to detect differences in fMRI patterns (amongst many voxels) between experimental conditions. Although a univariate analysis will be able to distinguish between two different visual stimuli presentations (e.g., faces vs. houses), it might not be able to distinguish between two different visual stimuli within the same category (e.g., two different faces). However, a multi-voxel pattern analysis might be able to make such a distinction when drawing on the BOLD signal pattern of all available voxels.

Firstly, one needs to select which voxels that should be used in the MVPA, which is called “feature selection”. The feature selection is usually a subset of all voxels in the brain based on a priori hypothesis or functional localizer. Selecting many irrelevant voxels will introduce unnecessary noise, and make it harder to detect pattern differences. For example, if the aim is to decode which face the participant is looking at, one might select voxels that are associated with face perception in general. Secondly, the data is organized such that the BOLD signal of the selected voxels corresponds to the right point in time when a certain face was presented. Thirdly, the chosen machine learning algorithm is then fed a portion of the data to train and learn to predict which face the participant is looking at based on the BOLD signal pattern of the voxels. Fourthly, once the algorithm is trained, it is fed new data for generalized testing, to determine if the trained algorithm can predict the correct face significantly better than chance. The classification performance is usually tested for significance with non-parametric permutation testing (for more on MVPA see Haynes & Rees, 2006; Haynes, 2015; Norman, Polyn, Detre, & Haxby, 2006; Tong & Pratte, 2012).
Results

Study I
The historically close relationship between working memory, long-term memory, and conscious experience has inspired theories of consciousness to posit that durable maintenance of information is a uniquely conscious function (Dehaene et al., 2014; Dehaene & Naccache, 2001). It was generally assumed that the retention of non-consciously presented information was caused by priming, and disappeared within 500 milliseconds (Dehaene & Changeux, 2011; Greenwald et al., 1996). However, we had reason to believe that non-consciously presented information could be retained for a few seconds (Bergström, 2011; Soto et al., 2011).

Overall, the experiments in Study I showed that non-consciously presented information could be retained up to 15 seconds without signs of decline. The long delay periods enabled us to isolate the sluggish BOLD signal specifically related to the delay period rather than stimuli presentations and discriminations. This non-conscious retention was associated with sustained BOLD signal change in the prefrontal cortex and cerebellum (Figure 3B). Contrary to expectations from (conscious) working memory maintenance, the sustained BOLD signal in the prefrontal cortex did not correlate with memory performance (Pessoa, Gutierrez, Bandettini, & Ungerleider, 2002; Wager, Spicer, Insler, & Smith, 2013).

The non-consciously presented letters were associated with BOLD signal change in the visual area thought to process letters (Heinzel et al., 2008; Park, Hebrank, Polk, & Park, 2012), and prefrontal and parietal regions previously related to letter perception (James & Gauthier, 2006). These areas largely overlapped with the more pervasive BOLD signal related to conscious letter perception (figure 3A). Surprisingly, there was no BOLD signal change related to the memory test of non-conscious information (figure 3C). Reaction times for memory tests of conscious information were quicker than non-conscious (and baseline) reaction times.

We lowered the statistical threshold (p ≤ .05) to see if there was any subthreshold hints of hippocampi involvement during presentation and memory test, but there were none1. This makes it less likely that hippocampi-based mechanism played a large role in the non-conscious retention.

In sum, we found that non-consciously presented information could be retained for much longer than previously assumed despite irrelevant distractors. This non-conscious retention was associated with sustained BOLD signal change in prefrontal and cerebellar regions. These initial findings are consistent with working memory, and can have implications on working memory models and theories of consciousness.

1 The hippocampi control analysis was not reported in the published article.
Figure 3. All working memory epochs with T2-seen > T2-absent trials in orange, T2-unseen > T2-absent trials in blue, and the overlap in purple (p ≤ .001, uncorrected for multiple comparisons, k = 20). The Y-axis: Beta values, X-axis: PAS = perceptual awareness scale, bar colors correspond to the PAS classification of T2-seen and T2-unseen trials, error bars: standard error of the mean. (A) Stimulus presentation (IFG = inferior frontal gyrus, SPL = superior parietal lobule, ITG = inferior temporal gyrus). (B) Delay period displayed at p ≤ .005 for illustrative purposes (PFC = prefrontal cortex, OFC = orbitofrontal cortex, Crus II in the cerebellum. (C) Response (PCG = postcentral gyrus, ITG = inferior temporal gyrus).

Study II

Given the results from Study I with a subjective criterion of conscious experience, we wanted to replicate those findings with a new method and material. We therefore changed technique from attentional blink to CFS. The CFS suppression leads to fewer ambiguities as to whether something was experienced or not, and stimuli can be presented for longer durations. CFS might also make it possible to use an objective criterion of experience. The purpose of Study II was to replicate the longevity findings from Study I with CFS, and to investigate what informational content that could be non-consciously retained.

Experiment I of Study II showed that spatial position (of emotional faces) could be retained up to 15 seconds, replicating the longevity of Study I with new material and CFS. In Experiment II we wanted to investigate the specificity of the retained information. That is, if tool identity and spatial position could be simultaneously retained rather than only one of them. The memory-test probe was therefore not limited to full match or non-match (as in Experiment I), but could also partially match the target stimuli in regard to identity or position. We could therefore analyze the hit rate, and false alarm
rates for partial matches, and non-matches separately (figure 4). It was found that the hit rate was larger than the other false alarm rates, and no difference between the false alarm rates. Had for example only spatial position been retained, the hit and spatial false alarm rates would have been equal, and larger than the others. These results show that it indeed was the conjunction of identity and position that was retained, and used during discrimination. The non-conscious effects found in Study II were on direct measures when reports indicated no experience, and therefore also based on subjective criteria. The reaction times did not show signs of non-conscious priming, and similarly to Study I the conscious reaction times were quicker than non-conscious (and baseline) reaction times.

In sum, we replicated Study I in terms of non-conscious retention longevity (15 seconds) with CFS and new material. Furthermore, we showed that the conjunction of stimuli identity and position can be non-consciously retained, and that reaction times showed no priming effects. These results support the notion that non-consciously presented information is more durable than previously assumed, and indicated that the experimental procedure was stable enough to use in the fMRI scanner.

![Figure 4](image.png)

Figure 4. The average proportion of non-conscious hits (sample matched identity and position), object-FA (sample matched object identity), spatial-FA (sample matched spatial position), and baseline-FA (sample did not match) with standard error for non-conscious trials. *p < 0.05, one-tailed.

**Study III**

A hallmark of working memory maintenance is persistent neural activity in prefrontal task-related, and posterior stimulus-related regions. In Study I we found sustained BOLD in prefrontal regions (and cerebellum), but not in stimulus-specific sensory regions. In Study III we therefore used the same paradigm as Study II with fMRI to see if we could extend our previous findings with sustained BOLD signal in regions related to stimuli identity and position.

In the pre-fMRI session the non-conscious memory performance (discrimination and detection tasks) was above chance after 5 second delay
periods. Surprisingly, discrimination and detection performance dropped to chance for the same participants after a delay of 5 – 15 seconds in the fMRI session. Additionally, the univariate fMRI analysis did not detect any BOLD signal change during the delay period. However, multivoxel pattern classification analysis (MVPA) can be more sensitive than behavioral measures (Haynes & Rees, 2005) and univariate analysis (Emrich et al., 2013; Riggall & Postle, 2012). We therefore used MVPA to further investigate the trial epochs (presentation, delay, and discrimination). We used the conscious trial epochs to select relevant subsets of voxels to include in the MVPA for the same epoch. That is, a conscious delay-based subset to decode non-conscious delay-related information, etc. Although our main interest was in the delay period, we began with stimuli presentations for comparison. It should be easier to decode something during presentations because of the constant visual input.

The classification algorithms successfully decoded non-conscious stimuli presence vs. absence using a subset of voxels from the whole brain. Spatial information (left vs. right visual field) was decoded using a subset of voxels in the parietal and occipital cortex. Next we tried to do the same for the delay period. The algorithms failed to decode presence vs. absence from a subset of whole-brain voxels. However, they successfully decoded presence vs. absence when using a subset of voxels in the frontal cortex, and spatial position (left vs right) from a subset of voxels in the occipital cortex. To control that the successful decoding of BOLD signal during the delay period was not caused by residual BOLD signal from the presentation, we used the same voxels that successfully decoded BOLD signal during the delay period, and tried to decode BOLD signal during the presentation epoch. Critically, the algorithms failed to decode presence vs. absence, and spatial position when using the same voxels as the delay period during the presentation epoch. This suggests that the successfully decoded BOLD signal during delay is related to memory, and not perception.

The univariate analysis detected BOLD signal change in the right anterior insula, and right inferior frontal cortex during the non-conscious discrimination tests. In addition, there was more pronounced BOLD signal change when the memory test probe was not identical (non-match) to the memoranda compared to when the probe and memoranda was identical (match). Non-matching probes also reveal higher BOLD signal change relative to baseline trials. These findings were similar for both conscious and non-conscious trials, and partially overlapped (figure 5).

Control analyses showed that non-conscious univariate and MVPA results cannot be explained by participants’ accidentally mislabeled non-conscious trials where they “experienced something” (PAS > 1) as “no experience” (PAS = 1). We determined that by dividing baseline trials in two groups, and contaminated one group with conscious trials. The amount of contaminated
trials was based on the proportion of conscious trials that were accidentally reported as “no experience” (1 trial per participant). Critically, neither univariate analysis nor MVPA could detect the contamination. In a similar control analysis as Study I (p ≤ .05), we found BOLD signal change in the right hippocampus during stimuli presentation, but not discrimination².

In sum, we found sustained stimulus-unspecific BOLD signal in frontal cortex, and stimulus-specific spatial information in the occipital cortex with an objective criterion of experience (memory performance at chance). These findings are in line with Study I & II, and show strong support for working memory maintenance of non-consciously presented information.

Figure 5. p ≤ .001, uncorrected for multiple comparisons, k = 20. **A.** BOLD signal change during conscious (hot colors) and non-conscious (cold colors) trials for memory probe recognition when the probe did not match the sample compared with sample-probe match (non-match > match). Overlap (purple) is evident in medial frontal (upper) and middle temporal (lower) regions. **B.** BOLD signal change during conscious (hot) and non-conscious (cold) trials for non-matching probes compared with baseline trials. Overlap (purple) is evident in middle frontal and supramarginal gyrus.

²The hippocampi control analysis was not reported in the manuscript.
Discussion

The purpose of this dissertation was to investigate the functional properties and neural substrates of non-conscious short-term retention, in order to determine what kind of memory it is, and possible implications to models of memory and theories of consciousness. Studies I – III show that non-consciously presented information can be retained for up to 15 seconds, sometimes with distractors, across different techniques (attentional blink and CFS) and material (letters, spatial position, and the conjunction of tools and position) to be remembered, and was based on neural and/or behavioral evidence using subjective and objective criterions of experience. In addition to sustained BOLD signal in the frontal cortex during the delay period, Study I found it in the cerebellum, and Study III in the occipital cortex (related to spatial position). I will here discuss these findings in relation to other literature, and possible theoretical and practical implications.

What kind of non-conscious memory is it?

Overall, the three studies corroborate and extend previous findings (Bergström, 2011; Soto et al., 2011), suggesting that non-consciously presented object identity and/or spatial position can be retained for several seconds, sometimes with distractors. However, this short-term retention could in principle be explained by many different kinds of memory, which here will be considered in turn.

Working memory

Conscious working memory maintenance is known to be relatively robust against distractions, and can be maintained for as long as the information is attended. Delay periods of several seconds are therefore not a problem as demonstrated by the high working memory performance in the conscious conditions of the studies in this dissertation. In addition to converging evidence from Study I & III showing that short-term retention of non-consciously presented information was associated with sustained BOLD signal in the frontal cortex, Study I found it in the cerebellum, and Study III found BOLD signal related to spatial position in the occipital cortex. It is possible that the absence of significant BOLD signal in visual regions in Study I can be explained by the relatively insensitive univariate analysis, and that a MVPA would have detected presence vs. absence or stimulus-specific features similar to Study III. Together, these findings are consistent with predictions based on (conscious) working memory, which commonly find task-related BOLD signal in the prefrontal cortex (Fuster, 2015; Pochon et al., 2001; Rowe et al., 2007; Sakai & Passingham, 2003), and stimulus-related BOLD signal in posterior regions during maintenance (Dru zinc & D’Esposito, 2001; Emrich et
Furthermore, the cerebellum is anatomically connected to the prefrontal cortex (Bostan, Dum, & Strick, 2013; Kelly & Strick, 2003), and have been related to verbal working memory (Desmond et al., 2005; Desmond & Fiez, 1998; Stoodley & Schmahmann, 2009). The findings of Study III are especially compelling because in addition to stimuli-unspecific information in the frontal cortex, there was also stimulus-specific spatial information in the occipital cortex - a hallmark of working memory - with an objective criterion of experience. Although it may seem unbelievable to decode non-conscious information from BOLD signal despite behavioral performance being at chance, it is in line with previous research. Haynes & Rees (2005) successfully decoded the orientation of fine lines in the primary visual cortex during non-conscious perception when discrimination was at chance. Furthermore, there is indirect behavioral support for non-conscious working memory maintenance when direct performance was at chance (Pan, Lin, Zhao, & Soto, 2013). Pan, Lin, Zhao, & Soto (2013) used backward masking to render faces non-conscious, and CFS during the delay period to suppress either an identical or different face, while timing how long it took for the suppressed face to break through suppression (and become conscious). Participants then had to guess if the face that broke through suppression was the same or different as the non-consciously presented face. Importantly, they found that faces that matched the non-consciously presented faces broke through suppression quicker compared to mismatching faces, but memory performance was not better than chance. Critically, a follow-up experiment showed that if there was no memory test at the end of the trial, there was no difference in break-through time between matching and mismatching trials. This study provides compelling evidence for non-conscious working memory for two reasons. First, they show that the non-conscious retention depends on prospective action, which is a crucial part of the definition for working memory. Secondly, they used an objective criterion for non-conscious memory (i.e., memory performance was at chance).

Showing that non-conscious short-term retention can depend on prospective use is an important contribution and complement, which lends weight to a working memory interpretation of Studies I - III. Studies I & III carry important contributions because unlike other studies, the BOLD signal can be related specifically to the delay period with relatively high spatial and content specificity. That is, they specifically focus on the actual retention processes related to the non-conscious information, which is crucial to determine if the retention state is active or latent.

Similarly, Dutta, Shah, Silvanto, & Soto (2014) used a similar masking paradigm as Soto et al. (2011), but with a 1.5 s delay and no distractors, to investigate its neural substrate with fMRI and transcranial direct current stimulation (tDCS). They found that BOLD signal change in prefrontal,
parietal, and temporal cortex correlated with memory performance, and that tDCS on the prefrontal cortex modulated performance. Although their findings are in line with Study I & III there are some caveats. Firstly, the sluggishness of the BOLD signal together with their short 1.5 s delay makes it impossible to separate the BOLD signal from presentation, delay, and discrimination. Secondly, the tDCS was applied for 15 min before the experiment, which similarly means that the modulated prefrontal involvement could be related to processes during any of the three trial epochs. Since it is impossible to know if their findings are related to encoding or retrieval rather than retention, it is impossible to conclude anything about how the information was retained, only that it was, and that the prefrontal cortex played a causal role.

King, Pescetelli, & Dehaene (2016; submitted, but not yet peer-reviewed) also used a similar masking paradigm as Soto et al. (2011), but with 800 milliseconds delay and no distractors. They used magnetoencephalography (MEG) to study neural activity during non-conscious retention. MEG has a much higher temporal (but lower spatial) resolution than fMRI, which permitted them to separate neural activity from presentation, delay, and discrimination despite short delay periods. Using MVPA (on whole-brain data) it was found that stimuli presence vs. absence could be decoded from presentation, delay, and discrimination, but stimulus-specific information (i.e., target angle) could only be decoded from presentation and discrimination, not the delay period.

The short delay period of 800 milliseconds opens up the possibility that neural activity represents residual activity from perception (iconic memory) rather than maintained activity as in working memory. The authors argue that the presence of non-conscious information during the delay indicates task-relevant maintenance. However, the target angle could not be decoded, and stimuli presence potentially contains target angle (task-relevant feature), spatial frequency and contrast level (task-irrelevant features). Since the task-irrelevant features could not be decoded from the delay period of consciously experienced stimuli, it could be argued that they are unlikely to persevere during non-conscious retention. However, after experiencing the stimuli participants likely focus attention on the relevant feature, and might thereby inhibit the irrelevant features more compared to when the stimuli was non-conscious. Consistent with Study I & III, the MEG pattern suggests that the non-consciously retained information is supported by global activity, not only by local activity in visual cortex, which is consistent with a working memory interpretation.

Studies I & III and King et al. (2016) seems to imply that non-consciously presented information can be maintained as sustained BOLD/neural signal. However, given the relatively weak signals and that BOLD and MEG signals are more sensitive to synaptic processing rather than neural spikes, it is
unclear how persistent the underlying neural spiking rate was, and the signal could reflect semi-persistent bursts of activity and/or metabolically demanding synaptic events rather than constant neural spiking, but nevertheless reflect short-term retention (Lundqvist et al., 2015; Mongillo et al., 2008; Shafi et al., 2007). This would mean that at its most extreme, the stimuli representation might be retained by sub-threshold membrane potentials without any neural spiking. However, for such sustained changes in membrane potentials there must be sustained input coming from somewhere else, which has to have neural spikes. In the case of Studies I & III, it would likely come in the form of sustained endogenous attention from the frontal cortex. For simplicity I will use persistent as in “persistent and/or on average persistent” to refer to all of these possible states ranging from non-spiking neural activity to constant spiking since my measures are on average BOLD signal, and because these active states will be considered together in contrast to latent states.

Studies supporting latent short-term retention have failed to find evidence of sustained BOLD/neural signal during successful (consciously presented) short-term retention of deprioritized information (LaRocque et al., 2013; Lewis-Peacock et al., 2012; Nelissen, Stokes, Nobre, & Rushworth, 2013). It has therefore been suggested that latent neural changes underlies deprioritized short-term retention. It remains difficult to know whether these deprioritized stimuli received unmeasurably small sporadic bursts of activity, very small amounts of continuous activity, or if they truly were completely inhibited (i.e., no neural activity at all). However, if the latter is assumed true, then it puts additional constraint on working memory models. Based on these findings it has been argued that latent neural changes underlies the working memory state called “activated long-term memory” (Larocque, Lewis-Peacock, & Postle, 2014).

It might seem surprising or contradictory that non-consciously presented information that never has been inside the focus of attention nevertheless can be maintained as persistent neural activity, while information previously conscious in the focus is retained as latent neural changes. How can these two different findings be reconciled within one working memory framework? A possible, but speculative, solution might be found in the definition of working memory – information retained for prospective use. I hypothesize that the persistent neural activity (in any form) is dependent on prospective action, while the latent neural change does not. A consequence of this line of reasoning is that any form of latent neural retention is not strictly working memory, but short-term memory, because it will decay in a time-interval dependent on other factors than prospective use.

It is currently hard to evaluate this hypothesis since the deprioritized stimuli in (LaRocque et al., 2013; Lewis-Peacock et al., 2012) was not currently necessary for a prospective action, but could become cued as necessary later
in the trial with a 50% probability. Similarly, Nelissen, Stokes, Nobre, & Rushworth (2013) also failed to decode the deprioritized stimuli, but in their case the stimuli never became relevant. In contradiction, Peters, Roelfsema, & Goebel (2012) were able to successfully decode unattended stimuli in a visual search task, but interestingly, they were always cued as relevant later in the trials. However, the interpretation gets slightly more complicated as the deprioritized stimuli appeared in the search stream on half of the trials, which could have biased decoding performance. Behaviorally, Oberauer’s state of activated long-term memory was initially based on the intrusion effect from information that no longer was task relevant, and never became relevant (Oberauer, 2001). In later studies the information was passive, and therefore not task relevant per se, but was still tested later on an unknown amount of trials (Oberauer, 2002, 2005). It seems clear that the context of how relevant the information outside the focus of attention is does affect its state as measured by behavioral studies (Zokaei, Ning, Manohar, Feredoes, & Husain, 2014). However, it is not possible to know the neural states during these behavioral studies.

Importantly, in Study I – III and other similar studies, the non-consciously presented stimuli was always relevant for prospective use, which might explain why non-conscious information outside the focus of attention seem to be maintained as persistent neural activity. The idea that persistent neural activity is dependent on prospective use might therefore reconcile the two conflicting cases where non-consciously perceived information can be maintained as persistent neural activity, and that consciously perceived but deprioritized information can be retained as short-lived latent neural changes. The length until deprioritized information might become relevant again and working memory load could also be factors that contribute to the neural state information is retained in. Study III had minimal working memory load, which might lend more neural resources to the non-conscious maintenance. However, Study I had a higher (but still small) load because the solution to the math task (T1) was simultaneously held in working memory consciously. It is interesting to note that many believe that attentional blink is the consequence of T1 being encoded into working memory (Bowman & Wyble, 2007). The letter (T2) is thus non-consciously encoded in working memory while participants are holding T1 in focus of attention while solving it. This suggests some level of parallel processing in terms of multiple prospective actions and encoding.

In sum, the findings of studies I – III and the current literature converge on the finding that non-consciously presented information can be maintained in working memory as persistent neural activity. However, we still need to consider the possibility of influences from other kinds of memory.
**Iconic memory**

Iconic memory can survive distractors in the same spatial location as long as they are of a different kind than the memoranda (Pinto et al., 2013), which was the case in Study I (letters vs. numbers). However, Soto et al. (2011) showed short-term retention despite conscious and non-conscious distractors of same type and spatial position as the memoranda, which should have overwritten iconic memory. Similarly, iconic memory is not dependent on prospective use (Pan et al., 2013). Although iconic memory is commonly thought to be a very brief (< 1 second) high capacity memory, it can last up to at least 4 seconds under certain conditions (with diminished capacity) when information is consciously presented (Averbach & Sperling, 1961; Sligte et al., 2008). In the name of prudence, I will assume that non-consciously presented information can last as long as consciously presented information. That means that iconic memory could have influenced the results of Experiment II in Study II, and the pre-fMRI session in Study III since the delay period was 5 seconds without distractors. Similarly, it could explain the retention after 1.5 seconds (Dutta et al., 2014) and 800 milliseconds (King et al., 2016). However, it seems unlikely that residual neural activity from non-consciously presented information would persist between 5 and 15 seconds. Even if it could, iconic memory would only predict BOLD signal/neural activity in visual regions during retention (Sligte et al., 2009), and is therefore inconsistent with the frontal involvement (Sligte et al., 2011) found in Study I & III and global oscillations in (King et al., 2016). Taken together, iconic memory is unlikely to have influenced the non-conscious short-term retention, with a cautious exception for the experiments with 5 second (Exp. II and pre-fMRI) and 1.5 second delay periods (Dutta et al., 2014).

**Priming**

Contrary to iconic memory, non-consciously encoded perceptual priming have been shown to last several minutes in some cases (Bar & Biederman, 1998, 1999; Gaillard et al., 2007). Similarly to iconic memory, non-conscious priming might be affected by distractors that share spatial position and object features at the same processing level, but is potentially quite robust against other distractors (Henson, 2003; Marsolek, 2008). The other studies used memory probes that never were identical to the memoranda, which make priming an unlikely explanation (Dutta et al., 2014; King et al., 2016; Soto et al., 2011). Priming can occur when stimuli are similar (not only identical) to the second repetition, usually in terms of semantic/conceptual relatedness or similar perceptual features in identification tasks. However, in these tasks they were to judge if the memorandum was clock-wise or counter clock-wise to the memory probe, thus the similarity or closeness to the probe cannot facilitate task performance.
In Study II we found that the conjunction of object identity and spatial position could be retained. Initially, we reasoned that such a feat would set working memory apart from priming. However, high-level visual object recognition has, contrary to common belief, been suggested to be position-dependent (Kravitz, Kriegeskorte, & Baker, 2010; Kravitz, Vinson, & Baker, 2008). If high-level visual object recognition automatically processes the object’s spatial position, then it should be possible for the conjunction of identity and position to be retained by non-conscious priming mechanisms. However, Study II and III did not find any signs of non-conscious priming effects in reaction times (Study I could not index priming because all four alternative letters were always shown during discrimination).

In Study III, there was higher BOLD signal change related to non-conscious discrimination when comparing memory probe non-match to match, and to baseline, which partly overlapped with the same comparison for conscious discriminations. These findings seems at odds with priming (consistent with the absence of priming indexed by reaction time) since there is no repetition suppression effect in the traditional sense (Henson, 2003). However, it is consistent with an antipriming interpretation of repetition suppression (Marsolek et al., 2010). It is not clear how to interpret the non-match findings in terms of priming. As mentioned previously, priming is too inclusive as a kind of memory, and can include latent structural neural changes, and residual activity, but I would also add persistent activity.

The conscious conditions in Study II & III shows strong priming effects in terms of reaction times (~ 250 milliseconds), and antipriming effects in BOLD signal (non-match > match). A facilitation of ~250 milliseconds seems to indicate more than only structural changes, since that is enough time for a sensory input to be sent from the retinae, processed in the cortex, and transformed to motor output. It therefore seems likelier that such strong priming effects mainly stems from active maintenance and preparation for response. If these effects are seen in a conscious working memory task, then finding them in the non-conscious condition cannot exclude working memory mechanisms by themselves. Certainly not when there are other indicators consistent with non-conscious working memory maintenance, which themselves are compatible with the antiprimed effects in BOLD signal.

Taken together, there is no indication that the non-conscious short-term retention is completely caused by non-conscious priming. However, there is a slight possibility that some priming mechanisms other than those overlapping with working memory mechanisms could have influenced the antiprimed repetition suppression effect to some extent.

**Hippocampi-based memory**

Non-consciously encoded hippocampi-based memory have been shown to last for minutes, with other stimuli presented between encoding and retrieval, and
should therefore not have a problem with distractors (Degonda et al., 2005; Reber et al., 2012). Contrary to findings in Study I – III, hippocampi-based long-term memory predicts the absence of delay-related BOLD/neural signal since it consists of latent (structural) neural changes. However, the presence of sustained BOLD signal during the delay period does not exclude influences from the hippocampi. We therefore investigated the presentation and discrimination epochs for signs of hippocampi-related encoding and/or retrieval respectively. In Study I there was no indication of hippocampi involvement during presentation or discrimination when lowering the threshold. In Study III, weak BOLD signal change appeared in the right hippocampus during stimuli-presentation. The absence of significant BOLD signal change in the hippocampi cannot exclude hippocampi-based influences (might be too weak to detect), but is likely not the whole story given the delay-related findings. Especially, when the medial temporal lobes including hippocampi are important for working memory tasks involving novel associations among items or features and long delay periods (Jonides et al., 2008; Ranganath & Blumenfeld, 2005; Ranganath & D’Esposito, 2001). It could therefore be the case that non-conscious hippocampi-based encoding of the conjunction of object identity and spatial position (for that specific trial) facilitates or necessitates the working memory maintenance of the associated features, and might be weakly active during the delay period as well. It can therefore not be assumed that there are no hippocampi-based influences, as latent neural changes or as working-memory related maintenance.

**Interim summary**
Overall, Studies I – III are consistent with the current literature on non-conscious short-term retention (Bergström, 2011; Dutta et al., 2014; King et al., 2016; Pan et al., 2013; Soto et al., 2011; Soto & Silvanto, 2014), which together makes a compelling case that non-consciously presented information can be maintained in working memory as persistent neural activity. It seems unlikely that iconic memory contributed to the short term retention. Although we cannot completely exclude influences from hippocampi-based memory and/or priming, we do not need to do so, because working memory, hippocampi-based memory, and priming are not mutually exclusive. Indeed, there are probably few if any memory tasks that isolate a single kind of memory (Tulving, 2002). Critically, however, these findings imply the existence of non-conscious working memory.

**Critique against non-conscious memory**
There has been some critique against the findings related to non-conscious working memory in Studies I & II and in (Dutta et al., 2014; Soto et al., 2011), but concerns any kind of non-conscious memory (Samaha, 2015; Stein, Kaiser, & Hesselmann, 2016). This criticism can be described as two main
concerns: (i) that subjective measures of experience might be biased towards underreporting, and (ii) that participants, after non-conscious perception, made a premature guess, and held the guess consciously in working memory. According to (i) and (ii), the findings can therefore be explained by conscious working memory rather than non-conscious working memory. In addition, Stein et al. (2016) directed more specific critique against the way d’ was used by (Dutta et al., 2014; Soto et al., 2011), for their reply see (Soto & Silvanto, 2016).

The possibility of bias when relying on subjective measures is well known, and applies to non-conscious cognition more generally. As described in the above section “Inferring the absence of conscious experience”, it is possible that participants report “no experience” when they in fact had a “vague experience” either systematically or accidentally. Subjective measures are liberal, and risk overestimating the non-conscious effect. However, objective measures are conservative, and run the opposite risk of underestimating or failing to find an existing non-conscious effects. Furthermore, both measures tend to converge on similar results, and might best be viewed as complementary.

We therefore reasoned that it was most sensible to start with subjective measures since a null effect likely meant that there was nothing to find, while a significant effect could be further investigated with more conservative measures. Studies I & II should therefore be interpreted with that caveat in mind rather than being dismissed altogether. However, Study III showed sustained BOLD signal despite subjective reports of “no experience”, objective direct measures (discrimination and detection) being at chance, and indirect measures (reaction times) not being different from each other or baseline. Additionally, control analyses in Study III showed that neither the univariate nor the MVPA results can be explained by accidental mislabeling of PAS. Study III, therefore, not only shows strong support for non-conscious working memory with an objective criterion of experience, but also strengthens the validity of studies I & II by partly replicating (and extending) their results of frontal involvement by maintenance of spatial information.

If, as Stein et al., (2016) suggests, participants had decided what to guess prematurely, and held the conscious guess in working memory during the delay, then reaction times should be similar for conscious and non-conscious trials. However, control analyses for Study I – III showed that discrimination reaction times were much faster for conscious compared to non-conscious trials. Furthermore, such conscious guesses would apply equally to baseline trials, and can therefore not explain the BOLD signal difference between non-conscious and baseline trials in Study I & III.

There is another concern that has been raised against non-conscious high-level processing in general. The concern is that non-conscious high-level features are driven by conscious experience of low-level features. For example,
maybe only semantic information can be processed non-consciously when color and/or shape are processed consciously. This line of argument goes for both subjective and objective criteria of experience, and depends on what processing level the subjective or objective task is measuring. For example, if the task is to discriminate between faces, it is possible that face discrimination performance is at chance, while still consciously experiencing the shape and spatial position of the faces. If the subjective or objective measures only measures the higher level it is impossible to know if the lower level is conscious or not. There is empirical support for this hypothesis with the attentional blink (Elliott, Baird, & Giesbrecht, 2016) and CFS (Gelbard-Sagiv, Faivre, Mudrik, & Koch, 2016).

It is not entirely clear that this is a criticism per se. It seems possible to argue that if the research question concerns a specific processing level, it does not matter if lower levels are experienced or not. However, I will address this issue here anyway. In Study I – III the participants were always instructed to use the PAS scale on all levels of the stimuli. That is, “no experience” was only to be used if there was no experience what so ever. If something undefined was experienced they were to use “vague experience”, and if they experienced enough to identify it “almost or clear experience” should be used.

However, using the PAS as instructed would arguably not be enough during the attentional blink paradigm (Study I), since the lowest level that sets letters apart from numbers is still pretty high. That is, if participants consciously experienced the letters’ low-level features, such as color and part of the shape, the features would be too similar to the distractors’ (numbers) low-level features for the participants to know the difference. It is therefore possible that the participants consciously perceived some low-level features (without knowing it), and that it enabled or facilitated the non-conscious processing of higher-levels necessary to discriminate between letter vs. number, different letters, and maybe the retention. This phenomenon might even be the reason why paradigms like attentional blink and inattentional blindness show relatively strong non-conscious effects. The CFS paradigms are less susceptible to this kind of criticism assuming the PAS instructions were followed, because experience of low-level features should be easy to discern from the mondrians, and reported as vague experiences. Furthermore, in Study III the performance on the detection (stimuli presence vs. absence) task was also at chance. This means that the participants were not able to detect any low-level features by objective measures either. The non-conscious BOLD signal can therefore not be attributed to conscious low-level processing based on an objective criterion (detection performance being at chance).

Overall, the critique raised against a non-conscious working memory interpretation does not seem to be able to explain the results of the studies in this dissertation or similar studies.
Is non-conscious memory weak and fragile?
The non-conscious behavioral and neural effects were statistically significant but small, which raises a couple of questions. Why are the effects small? Do such small effects have any theoretical and/or practical relevance? What do the small effects tell us about non-conscious working memory? By showing small but statistically significant effects, we can infer that non-conscious working memory exists. This is itself an important theoretical implication, and as discussed in later sections, can have implications on models of working memory and theories of consciousness. The practical significance will also be discussed further in a later section. If the measured effect of non-conscious memory is the full extent of its influence on behavior it might seem rather insignificant. What possible significance can such a weak and fragile memory effect have in our everyday lives all things considered? I will briefly discuss whether the observed non-conscious memory effects are fragile, and whether the measured effects are the full extent of its influences on behavior.

The non-conscious effects were small, but replicable over experiments, except for the fMRI session in Study III, which dropped to chance level. Why the performance dropped to chance is unclear. The mirror stereoscope used for CFS was slightly different in the fMRI scanner, which might have caused a slight rotation of the stimuli, but I do not see how that could affect retention. The delay period was longer (5 – 15 seconds) in the fMRI scanner compared to 5 seconds in the pre-fMRI session, which could have been a factor, but considering the 15 second retention in Study II that should not have been a problem by itself. In addition, the fMRI scanner environment was different, participants had to lie down in darkness, tightly squeezed, with rhythmic noise, for a long time without much rest/or movement, which causes drowsiness and less than optimal focus. The latter (maybe combined with the former changes) might have contributed to the drop in performance.

Given these overall small non-conscious effects it is tempting to assume that only a few number of trials contain memory traces, and therefore produce small mean effects. However, it is not possible to know which memory epoch(s) that caused the small non-conscious effects, and the performance drop in Study III. The bottleneck could be weak perception, maintenance, and/or use of information. That is, maintenance could be really reliable without much loss in fidelity, but if the perceived representation is weak and noisy, that is what will be maintained. Similarly, if perception and maintenance are relatively robust, that will not show behaviorally unless this non-conscious information is actually put to use. Given that the BOLD signal during the delay period did not correlate with memory performance in Study I, and we found maintenance related BOLD signal despite memory performance being at chance in Study III, one possibility is that the use of information is a smaller bottleneck than maintenance. It could therefore be the case that non-conscious memory traces were present on most trials, but
we cannot know if that is the case. The fact that non-conscious effects are small compared to conscious effects are likely due to the techniques that render information non-conscious. Ideally, the relevant information would be presented consciously, but used non-consciously, for maximum effect. However, such circumstance creates difficulty in controlling that the information is not used consciously. A solution is to present information non-consciously, that is, through some manipulation the information processing is degraded to such an extent that it is not experienced. Therein lies the problem. The only way to reliably measure effects of non-conscious processing is to severely limit them, and therefore inherently underestimate them. In the studies of this dissertation we have tried to isolate the non-conscious memory effect under relatively conservative experimental conditions, and they should be interpreted within that context. The full extent of non-conscious working memory can therefore not be inferred from these findings alone, and could be much larger than measured.

There are clinical cases where non-conscious processing does not appear to be as severely limited as in experiments with healthy participants. Blindsight patients have damage to their primary visual cortex, and report that they are blind in the affected visual field. They can nevertheless perform very well at certain discrimination tasks where the information is presented within their blind visual field. For some discrimination tasks the performance can be close to 100% correct when 50% is chance, while being lower in other tasks (Cowey, 2010). It is therefore possible that non-conscious working memory (and non-conscious cognition in general) has a much greater impact on our behavior than we are able to reliably measure in most experiments.

**Implications for working memory models**

I will here discuss the implications of non-conscious working memory maintenance as persistent neural activity on contemporary (state-based) working memory models. Based on all the data espoused through this dissertation, there are, somewhat simplified, four possible memory states: (i) latent (non-conscious) long-term memory with varying memory strength, (ii) latent (non-conscious) short-term memory with varying memory strength, (iii) non-conscious residual activity, (iv) conscious residual activity, (v) non-conscious persistent activity, and (vi) conscious persistent activity. The allocation of neural resources (attention) is what determines if a memory is in a latent or active state. What causes an active memory to become conscious is still a matter of debate, but the amount of attention (i.e., neural activity) is likely an important part in the neocortex. Whether a memory is active for prospective use is what distinguishes working memory from other active memory such as sensory memory (table 1).
Table 1. Taxonomy of memory states

<table>
<thead>
<tr>
<th>Index</th>
<th>Memory</th>
<th>Neural substrates</th>
<th>Conscious</th>
<th>For prospective use</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>LTM</td>
<td>Latent change</td>
<td>No</td>
<td>no</td>
</tr>
<tr>
<td>(ii)</td>
<td>STM</td>
<td>Latent change</td>
<td>No</td>
<td>no</td>
</tr>
<tr>
<td>(iii)</td>
<td>SM</td>
<td>Residual activity</td>
<td>No</td>
<td>no</td>
</tr>
<tr>
<td>(iv)</td>
<td>SM</td>
<td>Residual activity</td>
<td>Yes</td>
<td>no</td>
</tr>
<tr>
<td>(v)</td>
<td>WM</td>
<td>Persistent activity</td>
<td>No</td>
<td>yes</td>
</tr>
<tr>
<td>(vi)</td>
<td>WM</td>
<td>Persistent activity</td>
<td>Yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

Note: LTM = long-term memory, STM = short-term memory, SM = sensory memory, WM = working memory.

The existence of the (i) and (vi) state is not controversial, and therefore taken for granted here. Studies I – III, together with a growing amount of support (Bergström, 2011; Dutta et al., 2014; King et al., 2016; Pan et al., 2013; Soto et al., 2011; Soto & Silvanto, 2014), suggests that the (v) state exists, and other studies suggest that the (ii) state exists (LaRocque et al., 2013; Larocque et al., 2014; Lewis-Peacock et al., 2012; Nelissen et al., 2013). Given these constraints, a model of working memory should consist of at least one state resembling (v) and another resembling (vi). That is not to say that (i) and (ii) is not crucial for working memory. Indeed, working memory is the interaction between sustained attention and (i & ii).

As previously stated, most working memory models assume that the most privileged state information can be in, is the focus of attention, where information is maintained by persistent neural activity, and is consciously experienced. McElree's (2001; 2006) 1-state model posits a conscious focus of attention, and anything outside the focus is non-conscious latent long-term memory with varying strengths depending on how recent it was activated. It therefore assumes that conscious experience is necessary for working memory. Such 1-state models are inconsistent with findings of non-conscious persistent activity because they do not allow for the (v) state. In contrast, 2-state models usually allow for a non-conscious state as the focus of attention only captures a subset of all working memory content. Cowan's model (1995, 2001, 2008) may be compatible with these constraints because it assumes a non-conscious portion of activated long-term memory in addition to the conscious focus of attention. However, the activated long-term memory is usually thought to decay if not attended again, and thus resembles (iii), but it is unclear if such attendance can be non-conscious as in (v). For clarity, it might be reasonable to distinguish these two cases as two separate states. Fuster's model (1995, 2003a, 2015) is consistent with these restraints, as it explicitly postulates that conscious experience is not necessary for working memory, and therefore predicts states like (v).
Oberauer’s 3-state model is difficult to evaluate in terms of conscious states because it remains neutral on that matter. However, I will try to make reasonable inferences about the different states. The focus of attention is presumably conscious, but it is less clear if what he calls the “state of direct access” or more recently “the broad focus of attention” is conscious or not. Since this broad focus is supposed to be similar to Cowan’s focus of attention, it seems reasonable to assume that this state also is conscious. Oberauer also posits an activated long-term memory that likely is non-conscious, but arguably, might not depend on prospective use (Oberauer, 2001), and might consist of latent synaptic changes (LaRocque et al., 2013; Larocque et al., 2014; Lewis-Peacock et al., 2012; Nelissen et al., 2013). It therefore seems like the activated long term memory state is more like (ii) than (iii), which makes this model inconsistent with findings supporting the existence of (v). However, as mentioned earlier, it is difficult to evaluate because the deprioritized information are sometimes never relevant and other times only relevant on a proportion of the trials. A more rigorous evaluation of how the probability of prospective use of deprioritized information affects the psychological and neural state of such information will surely clarify the ambiguity, and may even lead to more working memory states.

Implications for theories of consciousness
Historically, working memory and conscious experience has been tightly linked. Investigations into the neural correlates of consciousness showed that neural activity/BOLD signal in higher visual regions, and the prefrontal and parietal cortex correlated with conscious perception. The frontoparietal regions involved in conscious perception were also involved in other functions such as working memory and attention (Naghavi & Nyberg, 2005). These findings have inspired prominent theories of consciousness such as the Global Neuronal Workspace (GNW; Dehaene & Changeux, 2011; Dehaene et al., 2014; Dehaene & Naccache, 2001), and the Higher-Order Thought (HOT; Lau & Rosenthal, 2011a, 2011b) theories. According to the GNW theory, non-conscious information is processed in local modules and is relatively short-lived (< 500 milliseconds). For that information to become consciously experienced and maintained in working memory, the information needs to be globally broadcast via the frontoparietal network.

However, the findings of Studies I - III seem to be inconsistent with the GNW theory. Firstly, that non-consciously presented information can be retained for longer than assumed (up to 15 seconds). Secondly, that this non-conscious short-term retention seems to be supported by relatively global persistent fronto-occipital (Study III) and fronto-cerebellar (Study I) neural activity. Thirdly, that conscious experience is not necessary for working memory maintenance. Fourthly, that non-consciously presented information can partly activate the cognitive control network during non-conscious
discrimination in Study III, which adds to previous findings (Lau & Passingham, 2007; van Gaal et al., 2010). The relatively global (and presumably some local) recurrent activity between frontal and posterior regions also seems to contradict theories that postulate recurrent activity in sensory regions to be sufficient for conscious experience (Lamme & Roelfsema, 2000; Lamme, 2006).

According to the HOT theory, for some information to become conscious there must be a second higher-order thought (in frontoparietal regions) about a first-order state (in sensory regions). The HOT theory does not seem to be affected much of the findings since it is neutral regarding non-conscious functions. Although, it might be hard to see how non-conscious working memory would not constitute some kind of HOT state of the maintained information. Furthermore, a non-conscious HOT about a non-conscious first-order state should cause a conscious experience (Lau & Rosenthal, 2011b), which clearly is not in the case of non-conscious working memory. However, HOT theorists would likely sidestep this concern by claiming that working memory is not the right kind of second-order state. First-order theories of consciousness are not affected as they generally posit that it is the level of activity in higher visual regions that determines if that regional content is experienced or not, and therefore allow non-conscious executive functions (Fuster, 1995, 2003a, 2015; Zeki, 2001, 2008).

**Practical implications**

What, if any, practical implications are there with the findings presented within this dissertation? A common question I receive when telling people about these findings is “so you have essentially shown that subliminal brainwashing is possible?” I therefore thought it relevant to shortly discuss if not too much about what practical implications there are, then at least what practical implications there are not. It seems many know about the infamous study by James Vicary from 1957. Vicary claimed to have flashed the words “Eat Popcorn” and “Drink Coke” non-consciously for 1/3 of a millisecond on the movie screen while showing the movie Picnic to an unsuspecting audience, and found an 18% and 58% increase in Coke and popcorn sales respectively. Although these claims were later revealed to be fabrications made to boost sales for Vicary’s failing marketing business, it seems to have had a lasting effect legally and in public opinion. Pratkanis (1992) critically reviewed the attempts to replicate Vicary’s claims, and concluded that subliminal persuasion is no more effective than placebo, and of no use in marketing. How is that reconciled with non-conscious memory?

Although perception and memory are necessary for persuasion, they are not sufficient for it. Showing that non-consciously perceived information can be maintained in working memory or any other memory does not imply that peoples’ opinions can be changed without them being aware of it. Non-
conscious persuasion is thus a different research question with little support (Pratkanis, 1992). However, there have been some recent attempts to non-consciously influence participants' preference towards specific brands. Such subliminal advertising has found small non-conscious effects, but only if the non-conscious information is goal-relevant, the participants are motivated to pursue the goal, and the brand is not already habitually consumed (Bustin, Jones, Hansenne, & Quoidbach, 2015; Karremans, Stroebe, & Claus, 2006; Strahan, Spencer, & Zanna, 2002; Verwijmeren, Karremans, Stroebe, & Wigboldus, 2011). For example, participants can be non-consciously influenced to get a slight preference for “Lipton Ice”, but only if they are thirsty (Karremans et al., 2006). It is unclear if this goal-dependent retention is in any way related to non-conscious working memory. It could also be related to something like stimulus-reward mapping, which has been shown to occur when non-consciously presented stimuli are temporally paired with water drops as rewards to thirsty participants (Seitz, Kim, & Watanabe, 2009).

Other studies have found that non-consciously presenting affective information (Blanchfield, Hardy, & Marcora, 2014), and positive age stereotypes for older adults (Levy, Pilver, Chung, & Slade, 2014) can affect physical performance. Small non-conscious effects could potentially have a significant impact for large businesses and professional athletes, but consciously presenting information might result in larger effects. If so, presenting non-conscious information seems rather pointless. I think it is likely that consciously presenting information will have a larger (positive or negative) effect. However, Levy et al. (2014) found that the non-conscious intervention had greater effect than the conscious intervention on physical performance. Perhaps, non-conscious information might have smaller effects, but can more easily by-pass endogenous modulation of perceived information.

**Synthesis**

When external sensory information becomes available it is perceptually processed by reactivation of related latent (phylogenetic and ontogenetic) memory. During normal conditions this perceptual reactivation is high enough to (somehow) cause a conscious experience of the sensory information (figure 6A). In addition to neural activity, this reactivation can cause short-term or long-term latent structural neural changes in the reactivated neural networks. When the external sensory information no longer is available there is nonetheless residual neural activity (i.e., sensory memory) for a few seconds (figure 6A). Any short-lived latent neural changes can last up to a minute, while long-lived latent neural changes can last for a life time, and can later be endogenously or exogenously reactivated. Initially the residual activity will be high enough to cause a residual conscious experience, but quickly drop below the threshold for conscious experience. A subset of the sensory memory can
be temporarily maintained for longer periods until prospective use (i.e., in working memory).

The information is maintained in working memory by persistent recurrent activity between task-related prefrontal and posterior sensory regions. If the persistent neural activity is high enough (in the focus of attention), it will be consciously experienced, but is otherwise non-consciously maintained for prospective use as below threshold activity outside the focus of attention (figure 6B). The persistent activity during working memory can also cause short-term or long-term latent changes.

![Diagram of conscious and non-conscious memory states](image)

**Figure 6.** Illustration of conscious and non-conscious memory states.

If another subset of the actively maintained (or latently retained) sensory information becomes prioritized, then neural resources are endogenously reallocated toward the newly relevant information. The deprioritized information will either be maintained non-consciously as persistent below threshold activity or latently retained neural changes depending on how relevant it is for prospective use. Similar shifts between conscious and non-conscious states can occur in the prefrontal cortex related to the main task and potential subtasks or contingencies. For example, when mentally performing difficult multi-step arithmetic, all the pertinent information will be maintained while solving the task. However, when focusing on one of the multiple steps you might not be conscious of information concerning the other steps until focus shifts.

Perceived information does not have to be consciously experienced in the focus of attention in order to be maintained in working memory. Non-consciously perceived information can be maintained from passively viewing suppressed information (Study II - III), and while actively engaged in another
working memory manipulation task (Study I). It might even be possible that focusing attention on non-consciously maintained information can boost the activity above the threshold for experience. Indeed, it has been suggested that retrospectively cuing endogenous attention to the position of a non-consciously presented stimuli can retroactively cause conscious experience of them (Sergent et al., 2012; Thibault, van den Berg, Cavanagh, & Sergent, 2016)

For example, sometimes when thinking intensively about something, and someone asks something else, you might be conscious of that person saying something but not of what they are saying. However, when you are finished with whatever you were focusing on, you can shift attention toward what was previously said, and retroactively experience and understand it. It is possible that such retroactive experience can occur because the information was maintained non-consciously below the threshold of experience (given that you intend to respond) until attention is shifted to it.

These dynamic shifts between working memory, short-term, and long-term memory states enable an extremely flexible use of information. Whether a state is conscious or non-conscious seems to depend on the amount of allocated attentional resources, and can dynamically shift depending on task demands. Non-conscious working memory states are likely important for the normal functioning of working memory as we know it. Most studies of working memory have not specifically investigated conscious or non-conscious working memory, and have therefore essentially investigated the conjunction of conscious and non-conscious working memory.

**Limitations and future directions**
The empirical studies of this dissertation, as is the case with most studies, come with some caveats to consider, and to be addressed in future research. I will here briefly mention these limitations, and suggest research questions for future studies. It is not possible to know the exact functions of the prefrontal (Study I & III) and cerebellar (Study I) involvement in relation to non-conscious short-term retention based on these studies. We can only infer their functions based on previous literature about the brain, and various kinds of memory mechanisms. Future research will have to address these issues by designing experiments to specifically tease apart the different functions of these, and possibly other, brain regions during non-conscious retention. These studies can, for example, use MVPA to classify different memory content from the different brain regions similarly to what we did in Study III, but for other stimuli features and cognitive processes. Is the neural activity in the frontal cortex truly task-related and/or related to endogenous attentional control of the retained stimuli as predicted by working memory? What is the role of cerebellum in Study I? Could it relate to stimulus-response mappings similar to what is seen during conditioning? The role of cerebellum is not even clear when it comes to conscious working memory.
In our studies we do not know if the observed memory performance and BOLD signal are dependent on prospective use, which is necessary for it to be conceptualized as working memory. That is, if there was one condition with prospective action, and one without, would the memory and/or MVPA performance drop to chance level in the condition without prospective action? Although it has been previously shown that non-conscious retention can be dependent on prospective use in (Pan et al., 2013), it would be interesting to replicate such findings with neuroimaging.

It is important to keep in mind that the neuroimaging findings of Study I & III are correlations, and we can therefore not know anything about the causal mechanisms of these findings in relation to non-conscious short-term retention. Neither can we know if there is an interaction between fronto-cerebellar (Study I) and fronto-occipital (Study III) cortices, or if they only co-occur independently. Future studies can address the issue of causality by using TMS or patient groups with various lesions to determine the brain regions that are causally involved in non-conscious short-term retention (Feredoes et al., 2011). It is possible to use TMS to alternately disrupt the normal functioning of prefrontal and stimulus-specific cortical regions during non-conscious retention, similarly to what was done on primates with cortical cooling (Fuster et al., 1985). Would disrupting the neural activity cause a decrease in non-conscious memory performance, or cause the MVPA classification performance to drop to chance level in the region not disrupted by TMS? Another possibility is to investigate functional interactions by using functional connectivity analysis to see if voxels in the prefrontal cortex correlate with voxels in other regions of the brain during retention.

Another caveat is that the MVPA in Study III is not based on variance, and is therefore considered a fixed effects analysis instead of a random effects analysis. Fixed effects analyses are not as generalizable as random effects analyses, and can only infer that the observed effects are significant for the group that is tested, while random effects analyses can infer that the effects apply to others (not tested) in the same population.

There are of course many other questions, not linked to limitations in the studies of this dissertation, which should be pursued in future research. It is not known how the conscious and non-conscious mechanisms of working memory interact with each other. For example, if the same stimuli with the same features (e.g., identity and position) are non-consciously presented in two conditions, but the conscious instructions differs such that in one condition only the identity is to be remembered, and in the other only the position is to be remembered, will the instructed information be preferentially retained? In the opposite direction, to what extent can conscious working memory and decision-making make use of non-consciously presented information?
It would also be interesting to follow up on the different states of deprioritized information in working memory. For example, during which conditions is it possible to use MVPA to decode maintained information? Here one could manipulate the length of delay durations or vary the probability that the deprioritized information needs to be used later between 0% and 100% in the same experiment. Another unknown factor is the capacity of this non-conscious short-term retention. We know that conscious working memory has a very limited capacity, but it is not clear if the non-conscious retention share this limited capacity, has a little more, or perhaps a lot more capacity? Similarly, the idea of non-conscious iconic memory has not been investigated as far as I know. Is non-consciously presented information briefly retained with much higher capacity before a subset can be maintained for a more durable time? Conscious working memory is also known for its flexible use of information. To what extent can non-conscious information be manipulated for prospective use? How does non-conscious working memory relate to other kinds of implicit learning based on the basal ganglia?

In sum, the conclusions based on the studies of this dissertation about the existence of non-conscious working memory should be seen as tentative, not conclusive. As briefly mentioned in this section, these studies have presented us with fertile ground, ripe not with limitations, but with opportunity for future research.

Conclusions
Overall the findings of this dissertation show that non-consciously presented information (identity and/or position) can be retained for several seconds, sometimes with distractors, and is associated with sustained BOLD signal in frontal and posterior regions. Study I found sustained BOLD signal change in the prefrontal cortex and cerebellum, and Study III found sustained BOLD signal in the frontal cortex, and stimulus-specific BOLD signal in the occipital cortex. The sustained BOLD signal likely represents persistent neural activity of some kind, and is consistent with working memory maintenance of the non-conscious information. The non-conscious working memory effects are small in the relatively conservative experimental conditions of these studies, but do not necessarily mean that non-consciously maintained information does not have larger behavioral relevance in our everyday lives. The current results add constraints on models of working memory and theories of consciousness. Working memory models need to accommodate the possibility that information can be non-consciously maintained as persistent neural activity for prospective use, and not only as residual activity or latent neural changes. The current results also contradict the GNW theory’s assumptions that consciousness is necessary for durable retention, global recurrent activity, and working memory.
References


doi:10.1038/363345a0


working memory. Nature, 386(6625), 608–11. doi:10.1038/386608a0


Elliott, J. C., Baird, B., & Giesbrecht, B. (2016). Consciousness isn’t all-or-


James, K. H., & Gauthier, I. (2006). Letter processing automatically recruits
a sensory-motor brain network. *Neuropsychologia, 44*(14), 2937–49. doi:10.1016/j.neuropsychologia.2006.06.026


Kravitz, D. J., Kriegeskorte, N., & Baker, C. I. (2010). High-level visual object representations are constrained by position. *Cerebral Cortex*,
20(December), 2916–2925. doi:10.1093/cercor/bhq042


Verwijmeren, T., Karremans, J. C., Stroebe, W., & Wigboldus, D. H. J.


