

THE DYNAMICS OF VISUAL REPRESENTATIONS
UNDER TOP-DOWN MODULATION

Pim Mostert

DONDERS
SERIES

THE DYNAMICS OF VISUAL REPRESENTATIONS UNDER TOP-DOWN MODULATION

Pim Mostert

Colofon

The work described in this thesis was carried out at the Donders Institute for Brain, Cognition and Behaviour, Radboud University, with financial support from the Netherlands Organisation for Scientific Research (NWO Research Talent 406-13- 001) awarded to Pim Mostert.

A full digital copy of the thesis is available online at:

https://www.predictivebrainlab.com/files/theses/PhD-thesis_Pim-Mostert.pdf

ISBN

978-94-6284-172-7

Design/lay-out

Wil van de Kamp

Print

24-Proefschriften.nl

© Pim Mostert, 2018

No part of this thesis may be reproduced, stored in a retrieval system or transmitted in any form or by any means without permission from the author, or when appropriate, from the publishers of the publications.

THE DYNAMICS OF VISUAL REPRESENTATIONS UNDER TOP-DOWN MODULATION

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op donderdag 24 januari 2019
om 16.30 uur precies

door

Pim Mostert
geboren op 5 november 1989
te Groningen, Nederland

Promotor

Prof. Dr. F. P. de Lange

Co-promotor

Dr. P. Kok (University College London, Verenigd Koninkrijk)

Manuscriptcommissie

Prof. dr. M.A.J. van Gerven

Dr. H.A. Slagter (Universiteit van Amsterdam)

Prof. dr. R. Vogels (KU Leuven, België)

Table of contents

1	General introduction	7
2	Dissociating sensory from decision processes in human perceptual decision making	17
3	Prior expectations induce pre-stimulus sensory templates	47
4	Similar neural activity patterns evoked by expected and unexpected object images	81
5	Eye movement-related confounds in neural decoding of visual working memory representations	101
6	General discussion	129
	Appendix	141
	References	143
	Nederlandse samenvatting	155
	Acknowledgements	158
	List of publications	160
	Biography	161
	Donders Graduate School for Cognitive Neuroscience	162



1

General introduction

How does our brain give rise to subjective visual experience? During large parts of the day, our eyes are constantly bombarded with light. These light rays carry information about our immediate surroundings. The eyes transduce this information into electrochemical signals and send it to the brain. Here, the information is further processed, ultimately leading to a conscious percept of the external world. How is this information processed? What representations are used by the brain? These questions have been extensively investigated, but many open questions remain.

A particularly intriguing facet of visual processing is that the subjective percept does not always accurately follow the objective physical world. Think of dreams, hallucinations or imagery. These are examples of where an individual sees something that is not physically present. Its complement - not seeing something while it is physically present - is also possible. Take a look at Fig. 1. Unless you've seen this image before, you'll most likely see a random collection of blobs. However, once you are made aware of the presence of a Dalmatian dog, it can no longer be unseen. Whereas you first failed to see something, you now clearly do - even though the physical stimulation has remained identical. Other obvious examples of where physical reality and subjective awareness do not match are given by visual illusions. In Fig. 2, the three elephants appear to be wholly different in size, even though they are actually the same.



Figure 1. While initially this image may seem like a random collection of blobs, one can easily see a Dalmatian in it. Crucially, after having seen it, it cannot be unseen.

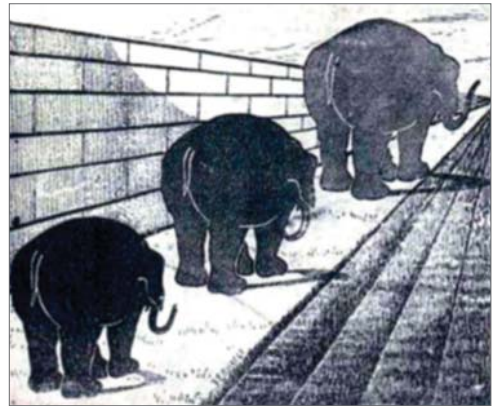


Figure 2. The top-right elephant appears to be largest, even though all three are of identical size. The converging lines from the wall and the road suggest that the top-right elephant is farthest away, and must therefore be biggest. Credits: www.moillusions.com.

In all of these cases, the visual percept is shaped by other factors than the purely bottom-up visual information as conveyed by incoming light. These factors may include voluntary effort (imagery), memory (Fig. 1) or spatial context (Fig. 2). I will collectively refer to such factors as internal top-down factors, because they originate from within the brain itself, as opposed to the bottom-up information transmitted by light entering the eyes.

While the influence of top-down factors on visual processing has been an active topic of study in recent years, relatively little attention has been paid to the temporal dynamics underlying this process. For instance, at which moment, relative to stimulus onset, do top-down modulations exert their influence? It is possible that incoming sensory information is modulated by top-down processes immediately upon arrival, or that it is first encoded in its original format after which it is gradually transformed. Another question is what happens to encoded visual information after the external source ceases. This is especially relevant when the information is still required for behavior after the visual stimulus has disappeared and therefore has to be retained over time. Is the sensory information maintained in the same format as first evoked by external stimulation, or is the relevant information stored in a different, perhaps more abstract format?

These questions are the central theme of this thesis. This thesis describes results from three particular domains: perceptual decision making, perceptual expectations and visual working memory (VWM). In particular, I examine the following issues:

- 1) What are the dynamics of the sensory representation?
- 2) How is this representation changed by internal top-down factors?
- 3) How do internally generated representations relate to bottom-up evoked representations in terms of their encoding format?

In parallel runs a methodological theme. The different experiments described in this thesis all make use of magnetoencephalography (MEG), which allows us to measure human neural signals with exquisite temporal resolution. Using this technique, I have applied time-resolved multivariate decoding methods that allow us to reveal the temporal unfolding of the information that is present in these neural signals. Finally, I make use of functional localizers, by which I can isolate sensory representations from the neural signal, and study how these are influenced by top-down factors. This particular combination of methods yields substantial benefits, allowing the main questions to be addressed from a new perspective.

Sensory representation

The sensory representation refers to the encoded visual information in the brain. This encoding is required in order for the information to be available to other neural processes. An influential model of decision making, for instance, posits that sensory information is encoded in sensory areas, from where it is read out by decision making areas in parietal and frontal areas (Gold and Shadlen, 2007; Heekeren et al., 2008). When a decision has been reached, an appropriate motor command is executed. Notably, the decision and motor processes are generic, in the sense that they are not linked to the specific sensory modality under investigation. Information stemming from auditory or tactile sources, for instance, may well be entered in the exact same decision process. Sensory representations, on the other hand, are linked to specific sensory transducers - the eyes, in the case of visual perception.

The format by which a sensory representation is encoded is determined by the collective tuning properties, or receptive fields, of the sensory neurons in question. A receptive field refers to a subspace of some feature space, in which stimulation will normally cause the respective neuron to fire. For example, neurons in the visual cortex possess localized spatial receptive fields, meaning that a particular neuron may respond to a stimulus presented in the upper-left visual field, but not to a stimulus presented in the upper-right corner. Neurons are also tuned to features other than spatial location. For instance, Hubel and Wiesel (1959, 1962) demonstrated that neurons in cat primary visual cortex respond vigorously to lines of a particular orientation, and an area known as the fusiform face area is particularly sensitive to visual presentation of faces (Kanwisher and Yovel, 2006). Thus, presentation of a stimulus leads to a characteristic pattern of activity that is determined by the collective tuning properties of the brain, in particular in sensory areas. It is this activity pattern that I refer to as the sensory representation, which encodes the visual information pertaining to the presented stimulus.

Operationally, I defined the sensory representation of a stimulus by means of a separate functional localizer task. In these blocks, the subjects are passively presented with visual stimuli while performing an unrelated task on a fixation dot. Importantly, the stimuli are not used for the task, thus rendering them irrelevant and relatively unattended. This approach allowed us to identify the sensory representations of the stimuli, whereby the neural activity is evoked in a bottom-up manner and while minimizing the influence of top-down factors.

Top-down influences on visual representations

The neural activity pattern evoked by a given stimulus is not fixed, and is shaped by factors other than the stimulus. In this thesis I will investigate the effect of top-down modulation in three specific domains: perceptual decision making, perceptual expectations and visual working memory.

Perceptual decision making

Perceptual decision making is the process by which a stimulus is transformed into a behavioral response about that stimulus, for instance whether the subject saw it or not. As mentioned above, an influential view is that the stimulus is encoded in sensory areas, from which it feeds into the decision process. However, there is now increasing evidence that this is not a one-way process, but that the sensory area is under constant top-down influence of the decision area. This insight came about with the observation that activity in sensory neurons correlated with the subject's eventual decision, even when the stimulus was identical (Ress and Heeger, 2003; Choe et al., 2014; Hesselmann et al., 2008a, 2008b; Nienborg and Cumming, 2009). More specifically, it was found that this correlation increased over time during a trial, i.e. as the formation of the decision progressed. On the other hand, the relative influence of the actual stimulus on the eventual decision decreased

over the course of the trial. These findings have been explained by arguing that decision areas impose their current best estimate of the external stimulus onto lower-level sensory areas via top-down feedback connections, resulting in a positive feedback loop that can drive the eventual decision solely on initial fluctuations in sensory evidence (Wimmer et al., 2015; Roelfsema and Nienborg, 2015; Haefner et al., 2016).

Perceptual expectations

Perceptual expectations refer to situations whereby visual features can be predicted from their context, which may include temporal and spatial cues. For instance, if a traffic light is green, one may expect to see a yellow light soon. The brain makes use of these statistics in order to form expectations about upcoming sensory information, thereby facilitating processing of that information (Bar, 2004, 2009). Perceptual expectations have a clear signature at the neural level: expected stimuli tend to evoke attenuated neural activity as compared to unexpected stimuli (Summerfield and de Lange, 2014). Moreover, expectation influences the informational content of the sensory representation, though there is some controversy regarding the nature of this change. While one study found an enhanced representation for expected stimuli (Kok et al., 2012), another found the opposite effect (Kumar et al., 2017). The reasons for this discrepancy are currently unclear and an active topic of investigation, possibly involving differences in neuroimaging method, species, and task- and attention-related differences.

Visual working memory

VWM is the process of keeping a transiently presented image online in one's mind such that it can guide behavior at a later time point, when the original image is no longer present. Although it was originally thought that the memorized item is encoded in persistent activity of prefrontal neurons (Curtis and D'Esposito, 2003), an important role has later been implicated for early visual cortex (Harrison and Tong, 2009; Serences et al., 2009; Albers et al., 2013). More specifically, it was found that keeping a visual stimulus online in working memory was associated with similar neural activity patterns in visual cortex as when that same stimulus was passively perceived. This inspired the view that high-fidelity representations of the memoranda are stored in respective sensory cortex, instantiated by top-down control from prefrontal neurons (Sreenivasan et al., 2014), though this remains topic of active debate (Xu, 2018; Scimeca et al., 2018; Gayet et al., 2018). In this thesis I investigate in particular the temporal aspect of this sensory instantiation of VWM items.

Multivariate decoding

The studies in this thesis rely heavily on the use of multivariate decoding analyses. Multivariate analyses - as opposed to univariate analyses - focus on patterns of activity and the correlations between features, rather than overall activity levels (Haxby et al., 2012; Tong and Pratte, 2012). This allows one to decode the information that is encoded

in the signal, for instance which stimulus a subject is perceiving (Kamitani and Tong, 2005).

I focused on multivariate decoding in combination with functional localizer data. As mentioned above, the functional localizer blocks define the sensory representations. By identifying these representations, I could trace the encoding of these representations in the experimental blocks. In this way, I was able to observe how encoding of the incoming visual information is modulated by a variety of cognitive or contextual factors.

Temporal dynamics

Most previous studies that specifically targeted the sensory representation were conducted using fMRI, because fMRI offers superior spatial resolution for use in multivariate decoding analyses. Over recent years however, electrophysiological methods such as MEG and EEG have become increasingly popular to be used with decoding analyses as well (Grootswagers et al., 2016; King and Dehaene, 2014). Owing to their high temporal resolution, these methods have opened up ways for new analyses and research questions. A particularly useful analysis method is that of temporal generalization, whereby the dynamics of the underlying neural code are revealed (King and Dehaene, 2014). Whether particular mental processes are accompanied by a dynamic or sustained neural code is currently subject of active debate, particularly in the field of working memory (Stokes, 2015).

Furthermore, time-resolved decoding analyses allow for further investigation of the results yielded by fMRI studies. For instance, the enhanced sensory representation for expected stimuli found by Kok et al. (2012) could be due to a change in response gain, but also due to a prolongation of the signal. Either scenario would show the same effect at the level of a temporally aggregating signal such as the BOLD signal. Moreover, a later study (Kok et al., 2014) found that when an expected stimulus was unexpectedly omitted, that its corresponding neural representation was nevertheless present. This suggests that the brain proactively activates a sensory template of the expected stimulus already before it is presented, but it is also possible that the brain actively instantiates a stimulus-specific surprise signal in order to convey the absence of the stimulus. Decoding analyses with high temporal fidelity allow for addressing these open issues.

Overview of this thesis

The four empirical chapters in this thesis each touch upon the three subthemes introduced above in order to collectively investigate the temporal profile of how top-down factors modulate sensory representations.

Chapter 2 describes a decision making experiment in which subjects are to detect a barely visible grating in a very briefly presented noise patch. This study addresses two

main questions. First, how is the sensory representation encoded when the subjective report does or does not match the actual stimulus? Second, computational models posit that a decision is reached by temporal integration of a decision variable that is driven by the strength of the sensory information. It is however unclear how such integration can occur when the stimulus information is only briefly present. Furthermore, the results demonstrate the added advantage of using a functional localizer. Whereas within-task decoding using cross-validation targets a potential wide range of underlying neural sources, between-task generalization using a carefully crafted functional localizer task allows for selective decoding of a specific signal of interest.

Chapters 3 and 4 describe perceptual expectation experiments. **Chapter 3** focuses on the temporal dynamics underlying perceptual expectations induced by a predictive cue. Specifically, does the brain instantiate a sensory template of an expected stimulus already before that stimulus is actually presented? In addition, we asked whether instantiation of such a template, and the quality thereof, is relevant for behavior. Among others, we found an effect of expectation on the sensory representation as probed by multivariate decoding, but did not see an effect in the event-related field (ERF). This highlights a benefit of including multivariate decoding analyses, as one might have incorrectly concluded that the expectation manipulation did not influence the neural signal had one only looked at the mean activity level as recorded in ERFs. **Chapter 4** focuses on a phenomenon known as expectation suppression, whereby expected stimuli evoke less neural activity than unexpected stimuli. The study was designed to address a number of open questions, but I surprisingly did not observe any expectation suppression. While this is a puzzling result considering the previous studies that did observe such an effect, I believe it is also an important one. By carefully considering a variety of factors at which our design differed from previous ones, this study helps map out the constraints and boundary conditions within which expectation suppression manifests itself. It thereby contributes to the development of contemporary theories about cortical function such as the predictive coding framework (Friston, 2015; Bogacz, 2017; Summerfield and de Lange, 2014).

Chapter 5 describes a combined VWM and mental imagery experiment. Previous studies have demonstrated that memorized and mental images share a neural code in early visual cortex with passively perceived gratings (Albers et al. 2013; Harrison and Tong, 2009). These experiments made use of fMRI which, due its low temporal resolution, precluded the researchers from investigating the temporal scale at which this mental image came about. Here I overcame this limitation by making use of the high temporal resolution of MEG, while probing the mental contents using multivariate decoding. However, I found that subjects made small but systematic eye movements in a way related to the item held in mind, which constituted a large confound in the neural recordings and the decoding results. This teaches an important lesson, namely that the internal maintenance or manipulation of visual information may be accompanied with small eye movements, even when subjects are instructed to fixate. Therefore, eye movement recordings should be obtained and inspected when employing such paradigms. Finally, it was demonstrated

that the functional localizer may provide a way to effectively counter these confounds.

Finally, in **Chapter 6** I summarize this thesis' findings and review how they contribute to our understanding of sensory encoding, top-down modulation thereof and associated underlying temporal dynamics. I conclude by discussing the thesis' implications for the current literature and provide directions for future research.



2

**Dissociating sensory from decision processes in
human perceptual decision making**

Abstract

A key question within systems neuroscience is how the brain translates physical stimulation into a behavioral response: perceptual decision making. To answer this question, it is important to dissociate the neural activity underlying the encoding of sensory information from the activity underlying the subsequent temporal integration into a decision variable. Here, we adopted a decoding approach to empirically assess this dissociation in human magnetoencephalography recordings. We used a functional localizer to identify the neural signature that reflects sensory-specific processes, and subsequently traced this signature while subjects were engaged in a perceptual decision making task. Our results revealed a temporal dissociation in which sensory processing was limited to an early time window and consistent with occipital areas, whereas decision-related processing became increasingly pronounced over time, and involved parietal and frontal areas. We found that the sensory processing accurately reflected the physical stimulus, irrespective of the eventual decision. Moreover, the sensory representation was stable and maintained over time when it was required for a subsequent decision, but unstable and variable over time when it was task-irrelevant. In contrast, decision-related activity displayed long-lasting sustained components. Together, our approach dissects neuro-anatomically and functionally distinct contributions to perceptual decisions.

This chapter has been published as:

Mostert, P., Kok, P., & de Lange, F. P. (2015) Dissociating sensory from decision processes in human perceptual decision making. *Scientific Reports* 5, 18253. doi: 10.1038/srep18253.

Introduction

A substantial part of cognitive neuroscience is devoted to the question of how the brain translates physical stimulation into behavioral decisions - an operation known as perceptual decision making (Gold and Shadlen, 2007; Heekeren et al., 2008). Theoretical frameworks posit that perceptual decisions arise from a sequence of functionally distinct processes (Ratcliff and McKoon, 2007). These frameworks distinguish the sensory process, where the physical stimulus is encoded into internal sensory evidence, from the decision process, that integrates this sensory evidence over time into a decision variable. A number of studies have revealed electrophysiological markers of these processes in humans, using a variety of paradigms (Philiastides and Sajda, 2006; Philiastides et al., 2006; Ratcliff et al., 2009; O'Connell et al., 2012; Wyart et al., 2012; de Lange et al., 2013; Kelly and O'Connell, 2013). Here we focus on the simplest of perceptual decision making tasks, stimulus detection, in which subjects are required to report the presence or absence of a stimulus in noise, and aimed to dissociate the neural activity underlying the sensory process from that underlying the decision process.

In this type of task, the behavioral response is typically used to post-hoc sort the data into the four stimulus/response-categories [hits, correct rejects (CRs), misses and false alarms (FAs)] in order to separate the underlying sensory and decision processes (Swets, 2014). By contrasting categories that differ on one dimension only (stimulus presence or behavioral report), one would expect to obtain the neural activity underlying the process that corresponds to that factor (e.g. Ress and Heeger, 2003; Ress et al., 2000; Lamme et al., 2002; Hulme et al., 2009; Choe et al., 2014). This approach suffers however from at least two conceptual problems.

First, the sensory and decision process are not fully dissociated. For example, if a stimulus is erroneously encoded during the sensory process, then an incorrect decision will likely follow. Thus the response factor targets not only the decision process, but also the sensory process. Moreover, the stimulus factor may target not only differences in the sensory process, but also the decision process. This is because, even when the final behavioral outcome is equal, the temporal integration during the decision process likely follows a deviating trajectory for incorrect decisions as compared to correct ones (Ratcliff and McKoon, 2007; Boldt and Yeung, 2015).

Second, as the post-hoc defined response factor is an observed variable, it is not under the experimental control of the researcher. As a result, any relation between response and neural activity may be confounded by third variables. For instance, perceptual decisions are modulated by ongoing fluctuations in neural activity or attention (Hesselmann et al., 2008a, 2008b, 2010; Monto et al., 2008; van Dijk et al., 2008). Thus, the response factor may target other processes, such as attention, that may subsequently modulate the sensory or decision process. Therefore, it is often unclear how to interpret differential neural activity revealed by this factor.

In summary, differential neural activity obtained by comparing the stimulus/response-categories is difficult to interpret, and conclusions derived from this approach require caution. Here, we adopted a different approach, which does not suffer from these limitations, to dissociate sensory from decision processes in human magnetoencephalography (MEG) recordings. We first identified, using a separate localizer task, the neural activity corresponding to the sensory process in absence of a decision. Then, we traced the neural signature of the sensory process while subjects performed a perceptual decision making task. We found that this between-task generalization method reliably identified the sensory representation in an early time window. Moreover, we observed that this sensory representation was stabilized and maintained over time, but only when required for a decision. In summary, our approach yields a new window onto the role of sensory processes during perceptual decision making.

Materials and Methods

Subjects

Twenty-four healthy human volunteers, recruited from the institute's subject pool, participated in the experiment and received either monetary compensation or study credits. The study was approved by the local ethics committee (CMO Arnhem-Nijmegen, Radboud University Medical Center) under the general ethics approval ("Imaging Human Cognition", CMO 2014/288), and the experiment was conducted in compliance with these guidelines. Written informed consent was obtained from each individual. Two subjects were excluded during preprocessing due to insufficient data quality (severe eye and muscle artifacts). The remaining twenty-two subjects (nine females, age 19-30 years) had normal or corrected-to-normal vision.

Stimuli

Stimulation consisted of visual noise presented on a gray (50% of maximum pixel intensity, luminance: 321 cd/m²) background, which could contain an embedded horizontal or vertical grating (Fig. 1B). Noise patches consisted of white noise that was subsequently smoothed with a Gaussian kernel ($SD = 0.05^\circ$). Gratings were sine waves with a spatial frequency of 1 cycle/ $^\circ$ and random phase. The gratings were embedded in the noise by averaging the two images, weighted according to a desired noise level (0%: full-contrast noise-free grating; 100%: pure noise). The pixel values of the resulting image were rescaled such that the minimum and maximum values mapped onto 0% and 100% of maximum pixel intensity, respectively. Finally, to obtain an annulus, all stimuli were masked with a radially oriented Gaussian mask ($SD = 2^\circ$) centered at a radius of 6° , resulting in an overall diameter of approximately 24° . Stimuli were generated and presented using MATLAB (The Mathworks, Inc., Natick, Massachusetts, United States) and the Psychophysics Toolbox extensions (Brainard, 1997).

Procedure and experimental design

Each subject participated in a behavioral practice session in order to become familiar with the experiment. The practice sessions were scheduled at most two days before the main experimental session. The main session involved three different types of blocks. After an initial staircase block (see below), subjects performed six sensory processing (localizer) blocks and six perceptual decision making blocks in alternating order, starting with a sensory processing block.

In the staircase block, we used the Quest staircase procedure (Watson and Pelli, 1983; as implemented in PsychToolbox) to estimate the individual noise level at which each subject correctly detects a grating embedded in noise in 70% of the cases. The procedure was similar to the perceptual decision making blocks (see figure 1), except that subjects only had to indicate the presence or absence of a grating and not its orientation. In addition, subjects received feedback on every trial. Only the trials in which a grating was actually presented were used to update the staircase. For the last eighteen subjects, convergence of the staircase was visually inspected at the end of the block and, if convergence was not yet achieved, more staircase trials were administered.

Each sensory processing block comprised 120 trials during which subjects were presented with a brief stimulus for 50 ms (Fig. 1A). In 50% of the trials, the stimulus was pure noise (referred to as noise trials) whereas the other 50% contained a grating embedded in the noise (referred to as grating trials). In half of the grating trials, the grating had a horizontal orientation whereas a vertical grating was present in the other half. The noise level of the grating trials was set to 90%. This value was chosen such that it was sufficiently high to be comparable to the stimuli presented during the perceptual decision making blocks, yet low enough to ensure clear visibility of the gratings (Fig. 1B). For the first seven subjects, the inter-trial interval was fixed to 950 ms, whereas for the remaining fifteen subjects this interval was randomly drawn from a uniform distribution between 850 and 1050 ms. As we were specifically interested in the neural signature of sensory processing in the absence of higher-level attentional and/or decision processes, subjects were required to perform a task at fixation to draw attention away from the stimuli. In 10% of the trials, and balanced across stimulus conditions, the fixation dot (diameter 0.2°) was absent during the 50 ms stimulus presentation. Subjects were instructed to report such a “blink” by pressing a button as quickly as possible. These “oddball” trials, as well as the non-oddball trials on which subjects erroneously pressed a button, were excluded from further analyses.

A perceptual decision making block comprised 80 trials. Each trial began with a 1000 ms fixation period, after which a stimulus was briefly presented for 50 ms (Fig. 1A). Again, this stimulus was either pure noise (in 50% of trials), or contained a grating (vertical grating in 25% and horizontal grating in 25% of the trials). Unlike the sensory processing blocks, the noise level was set to a much higher level, namely the individual threshold of perception as determined by the staircase procedure. Then, after 600 ms the letters

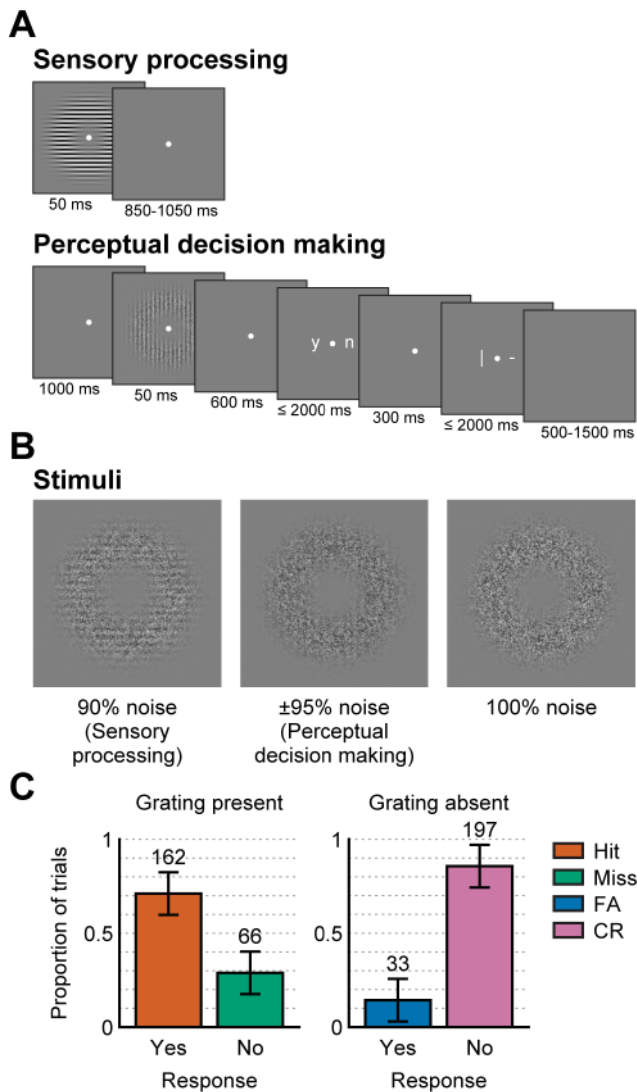


Figure 1. Experimental paradigm and behavioral results. **(A)** In the sensory processing blocks, the noise/grating stimuli were irrelevant and unattended. In the perceptual decision making blocks, a decision had to be made regarding the presence or absence of a grating. Grating visibility is enhanced for illustrative purposes. **(B)** Example stimuli. The noise level in perceptual decision making blocks was tailored to individual detection thresholds. **(C)** Average response proportions in perceptual decision making blocks, for grating present and grating absent trials, color-coded according to the four stimulus/response-categories. The numbers denote the average number of trials available in each of these categories. Error bars depict standard deviations.

‘y’ and ‘n’ (as abbreviations for “yes” and “no”, respectively) were displayed, centered around the fixation dot. Subjects reported their decision as to whether they had perceived a grating or not by pressing a button with either the left or the right hand, corresponding to the position of the letter that matches their decision. The position of the letters (‘y’ left and ‘n’ right, or ‘n’ left and ‘y’ right) was randomized across trials to orthogonalize perceptual decision and motor response preparation. Trials in which no button press was made were discarded from further analysis. After button press, or after 2000 ms in case of no button press, another fixation period of 300 ms followed, after which a second display with a horizontal and a vertical line was presented to inquire the subject’s decision regarding the orientation of the grating. The position of the lines were also randomized across trials. Subjects were instructed to, in trials where they perceived only noise, indicate the orientation to which they thought the noise was most similar. Finally, a blank screen was presented with an inter-trial interval drawn randomly from a uniform distribution between 500 and 1500 ms.

The two different grating orientations were included in the design because we had hypotheses regarding not only the sensory processing of stimulus presence versus absence, but also regarding the processing of orientation-specific signals. However, since we were not able to successfully decode the orientation of stimuli from the MEG signal, stimulus orientation was left out of further consideration.

Behavioral analysis

The observer’s sensitivity d' and criterion c were calculated as follows:

$$d' = Z(\text{Hit-rate}) - Z(\text{FA-rate}) \quad (1)$$

$$c = -\frac{1}{2} [Z(\text{Hit-rate}) + Z(\text{FA-rate})] \quad (2)$$

where $Z(\dots)$ is the inverse standard normal distribution.

MEG recording and preprocessing

Whole-head neural recordings were obtained using a 275-channel MEG system with axial gradiometers (VSM/CTF Systems, Coquitlam, BC, Canada) located in a magnetically shielded room. Throughout the experiment, head position was monitored online, and corrected if necessary, using three fiducial coils that were placed on the nasion and on earplugs in both ears. Behavioral responses were made using two MEG-compatible button boxes, one for each hand. Visual stimulation was projected from outside the magnetically shielded room, via mirrors onto a screen in front of the subject. Furthermore, both horizontal and vertical electrooculograms (EOGs), as well as an electrocardiogram (ECG) were recorded to facilitate removal of eye- and heart-related artifacts. All signals were sampled at a rate of 1200 Hz.

The data were preprocessed offline using FieldTrip (Oostenveld et al., 2010; www.fieldtriptoolbox.org). Notch filters were applied at 50, 100 and 150 Hz to remove line noise

and its harmonics. In order to identify artifacts, the variance (collapsed over channels and time) was calculated for each trial. Trials with large variances were subsequently selected for manual inspection and removed if they contained excessive and irregular artifacts. Independent component analysis was subsequently used to remove regular artifacts, such as heartbeats and eye blinks. Specifically, for each subject, the independent components were correlated to both EOGs and the ECG to identify potentially contaminating components, and these were subsequently inspected manually before removal. For covariance computation in the source localization, the remaining components were transformed back to sensor-space and subsequently low-pass filtered at a cut-off frequency of 30 Hz. For the decoding analysis however, the components were kept in component-space to ensure that its covariance matrix is of full rank as required for this analysis (see below). Finally, the data were baseline corrected on the interval of -200 to 0 ms relative to stimulus onset.

Decoding analysis

The general idea behind a neural decoding analysis is that it attempts to invert the encoding process. The encoding process determines the neural responses as a function of some parameter, for instance a physical stimulus or an experimental condition. A decoding analysis aims to invert this function in order to unveil the encoded parameter as a function of neural signals. The first step in achieving this is to estimate a forward model that describes the encoding process. For this, a data set is used in which the parameter is known and the corresponding neural signals are recorded empirically. Secondly, an inverse model is estimated on the basis of the forward model and subject to some criterion of optimality. We will refer to such an inverse model as a decoder, as it takes neural signals as input and produces an estimate of the encoded parameter as output.

Our method is largely based on linear discriminant analysis as described in Blankertz et al. (2011). Linear discriminant analysis attempts to find a linear transformation of the data, such that the resulting signal is optimally discriminative between two classes. First, we demeaned the data such that for each time point and for each feature, the average over trials equals zero. Features here refer to independent components (see *MEG recording and preprocessing*). Let $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ be column vectors of length F , where F is the number of features, that contain the neural responses in the training set for class 1 and 2, respectively, at some time point and averaged across trials. Then, the weights vector \mathbf{w} that optimally discriminates between classes on the basis of the features is given by (Blankertz et al., 2011):

$$\mathbf{w} = \tilde{\boldsymbol{\Sigma}}_C^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \quad (3)$$

where $\tilde{\boldsymbol{\Sigma}}_C$ is the common regularized covariance matrix (see below). Next, let \mathbf{X} be a matrix of size $F \times N$, where N refers to the number of trials, that contains the data to be decoded. The decoded signal \mathbf{y} is then obtained by:

$$\mathbf{y} = \mathbf{w}^T \mathbf{X} \quad (4)$$

where $(\dots)^T$ denotes the matrix transpose. In linear discriminant analysis, this signal is transformed into a discrete class membership by specifying a cut-off value for \mathbf{y} . Here, we deviated from standard linear discriminant analysis. Rather than assigning a discrete label to a trial, we were interested in a continuous measure of the degree to which a class is encoded in the neural signals. Thus, we did not apply a binary cut-off to the decoded signal. Furthermore, to make this signal comparable across time points, we added a normalization factor to the weights vector such that the mean difference in the decoded signal between classes equals a value of one. This is accomplished by modifying equation (3) into:

$$\mathbf{w} = \frac{\hat{\Sigma}_C^{-1}(\hat{\mu}_2 - \hat{\mu}_1)}{(\hat{\mu}_2 - \hat{\mu}_1)^T \hat{\Sigma}_C^{-1}(\hat{\mu}_2 - \hat{\mu}_1)} \quad (5)$$

in which the denominator is the normalization factor. Thus, equation (4) and (5) constitute our final decoder. We term the output of the decoder the “discriminant channel”, as it optimally discriminates between the two classes that it was trained on. Specifically, if we denote the mean discriminant channel amplitude in class 1 and 2 by $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$, respectively, then we expect $\bar{\mathbf{y}}_2 > \bar{\mathbf{y}}_1$ if there is information in the neural signals pertaining to the classes, whereas we expect $\bar{\mathbf{y}}_2 = \bar{\mathbf{y}}_1$ if no such information is available. In other words, the difference in the discriminant channel between classes is a measure of the discriminability of these classes on the basis of the neural signals.

The interpretation of the decoder’s output as a discriminant channel is further motivated by drawing an analogy to linearly constrained minimum variance (LCMV) spatial filters (van Veen et al., 1997), more commonly known as beamformers. A beamformer is an inverse model to estimate neural activity at the source level, given sensor-level activity. Beamformers are commonly used to extract time courses of activity of some source-level region of interest into so-called virtual channels, as if the activity in that region had been recorded directly. The forward model in this method, also known as the leadfield, describes how sensor-level activity varies as a function of source-level activity. The analogy is made by noting that equation (5) is equivalent to the calculation of beamformers, but whereas the forward models for beamformers are defined in a spatial sense, the forward model in our decoding method is defined in a discriminatory sense. It describes how activity at the sensor-level varies as a function of a discriminating parameter, namely the class, and is in fact the difference event-related field $\hat{\mu}_2 - \hat{\mu}_1$. If this difference event-related field is entered in the LCMV beamformer formula (van Veen et al., 1997) as the forward model, then equation (5) is the result. Thus, the output of our decoder is analogous to a virtual channel, but defined in a discriminatory rather than a spatial sense, hence the name discriminant channel. Contrary to spatially defined virtual channels, the discriminant channel may stem from a wide array of neural sources, and its output is the collective activity of this array. Finally, we point out that the normalization in equation (5) corresponds directly to the unit-gain constraint in LCMV beamformers (van Veen et al., 1997).

To facilitate comparison to other decoding methods, we included the results in the Supplementary Information normalized by their standard deviation at the individual level, yielding grand averaged Cohen's d effect sizes. Moreover, we also carried out an additional analysis in which we did set a cut-off value in order to assign a discrete class to each trial. This value was chosen as $\mathbf{w}^T(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)$, because this results in equal type I and type II error rates (Bandt et al., 2009). By assigning discrete labels to trials, the results can be expressed as the proportion of trials that are assigned to either of two classes.

An important element in the decoding analysis is that it takes advantage of correlations between features in order to suppress noise. Let the column vector \mathbf{x}_i of length F denote the data in trial i , and define the corresponding mean as:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \quad (6)$$

then the estimated covariance matrix is obtained by:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T \quad (7)$$

For optimal noise suppression, we improved this estimation by means of regularization by shrinkage, using the analytically determined optimal shrinkage parameter (for details, see Blankertz et al., 2011). This procedure was performed separately within each condition, and the resulting condition-specific regularized covariance matrices were subsequently averaged to obtain the common regularized covariance matrix $\tilde{\boldsymbol{\Sigma}}_c$.

The decoding analysis outlined above was performed in a time-resolved manner by applying it sequentially at each time point, in steps of 5 ms, resulting in an array of decoders. To improve the signal-to-noise ratio, the data were first averaged within a window of 29.2 ms centered around the time point of interest. The window length of 29.2 ms is based on an *a priori* chosen length of 30 ms, but minus one sample such that the window contained an odd number of samples for symmetric centering. Thus, the output of the time-resolved analysis is a one-dimensional time series of the amplitude of discriminant channel, for each individual trial.

It is imperative that the training data set is independent from the data set that is to be decoded in order to avoid “double dipping” (Kriegeskorte et al., 2009). Therefore, we adopted a leave-one-out approach whenever a trial was decoded that belonged to one of the classes on which the decoder was trained. Note that no such procedure is required when the decoded data set stems from a class different from those used in training, for instance when generalizing across conditions or across blocks.

To facilitate a neurophysiological interpretation of the decoded signals, we derived a

spatial projection of the discriminant channel. This spatial projection displays the coupling between the sensors and the discriminant channel and as such gives insight into which sensors contribute to the discrimination. The spatial projection \mathbf{a} is given by (Parra et al., 2002)

$$\mathbf{a} = \frac{\mathbf{X}\mathbf{y}^T}{\mathbf{y}\mathbf{y}^T} \quad (8)$$

and this was subsequently fed into the source localization (see *Source localization*).

The key aspect of the current study is that we used a functional localizer in order to extract the neural signature of sensory-specific processes, and subsequently traced this signature during perceptual decision making. This was done by training the decoders on the data from the sensory processing blocks and subsequently applying these decoders to the data from the perceptual decision making blocks. That is, the decoders were generalized across blocks. In addition to this generalization, we also trained and decoded within both the sensory processing and the perceptual decision making blocks separately. Within the sensory processing blocks, the decoders were trained to discriminate between noise trials and grating trials. Within the perceptual decision making blocks, the decoders were trained to discriminate between CRs and hits. This contrast was chosen, because these conditions correspond to accurate perceptual decision making and are therefore commonly used as a baseline to which inaccurate perceptual decisions (i.e. FAs and misses) are compared. These decoders were used to decode both CRs and hits themselves (using the leave-one-out approach), as well as the FAs and misses.

Finally, we implemented the temporal generalization method to elucidate the temporal organization of the neural processing stages that underlie sensory processing and perceptual decision making (King and Dehaene, 2014). When training a decoder on any specific time point, we simultaneously applied this decoder to all other time points. After averaging the obtained discriminant channel amplitude over trials, this results in a (training time) \times (decoding time) matrix per condition. Comparing these matrices for two specific conditions, by subtracting one from the other to obtain a difference matrix, provides insight into how discriminability between these conditions generalizes across time. For instance, a row in such a difference matrix, corresponding to some time point t_{train} , describes how well the two conditions can be discriminated over time on the basis of a decoder that is specifically trained, i.e. is optimally discriminative, at t_{train} . Conversely, a column, corresponding to some time point t_{decode} , gives insight into how well the two conditions can be discriminated at time point t_{decode} on the basis of the decoders trained on all other time points. As another example, consider the temporal generalization matrices obtained by generalizing from sensory processing to perceptual decision making, and consider in particular the average temporal generalization matrices for hits and CRs. Then, the difference between these matrices at the entry corresponding to some training time t_{train} , and decoding time t_{decode} is a measure of how well hits can be discriminated from CRs at t_{decode} , on the basis of the weights that was maximally discriminative between

noise and grating trials at t_{train} . The rationale behind the temporal generalization method is that the neural pattern identified by the decoder to be discriminative corresponds to some underlying neural process, and this pattern should be generated whenever that neural process is active. Thus, by testing for the presence of a particular neural pattern, one obtains an activity time course of the corresponding neural process. Note that the temporal generalization method described here is not to be confused with the between-block generalization described above.

Statistical testing

Statistical analyses were performed on subject-level temporal generalization matrices, averaged across trials. Contrasts between conditions were tested for statistical significance using permutation tests in conjunction with cluster-based correction for multiple comparisons (Maris and Oostenveld, 2007). Specifically, univariate paired t-tests were calculated for the entire matrix. Elements that passed a threshold value corresponding to a p-value of 0.05 (two-tailed) were marked, and neighboring marked elements were collected into separate negative and positive clusters. No specific constraint was set on the minimum number of marked elements in order to be considered a cluster - that is, the minimum number of required neighbors is 1. Elements were considered neighbors if they were directly adjacent, either cardinally or diagonally. Finally, the t-values within each cluster were summed and rectified, and these values were fed into the permutation framework as the test-statistic. Consequently, all tests were two-tailed. A cluster was considered significant when its p-value was below 0.05. The number of permutations per contrast was 10000.

Source localization

To substantiate the interpretation of the discriminant channels, we performed source localization on its spatial projection over time (see *Decoding analysis*). We did not *a priori* specify regions of interest and therefore made use of the minimum-norm estimation technique (Dale et al., 2000, as implemented in FieldTrip), rather than beamformers, because the former is the preferred technique when estimating distributed event-related source activity (Jensen and Hesse, 2010). Single-trial covariance estimates were calculated from the data during the baseline interval of -200 to 0 ms relative to stimulus-onset. These single-trial estimates were averaged within condition and subsequently averaged over conditions. Our source model included 8196 source locations, organized along a cortical mesh based on a template brain provided by FieldTrip. The source model was aligned with each subject's individual head position within the MEG system as determined by the three fiducial coils (see *MEG recording and preprocessing*). The covariance matrix was regularized, and the leadfields were depth-normalized and prewhitened prior to the source estimation.

The source estimation resulted in a dipole moment for each source location, over time.

This dipole moment is the change in orientation and amplitude that occurs with a unit change in the discriminant channel. Thus, if a particular source does not contribute to the discriminant channel, then the dipole moment for this location has amplitude zero. We therefore reduced the dipole moments to scalar values by taking the length of the vectors, resulting in a cortical map that represents each location's contribution to the discriminant component. However, as taking the length of a vector always results in a positive value, noise does not cancel out and therefore results in a positive bias. This bias is not uniform across source locations and as such interferes with the interpretability of the source localization. To counter this problem, we applied a permutation procedure to quantify the bias per source location and used this to normalize the source activity. Specifically, per subject and per time point of interest, trial labels were shuffled and subsequently fed into the decoding analysis and source estimation. This procedure was repeated 10000 times, resulting in a distribution of the noise for each source location. Then, noise-normalized activity at a source location is obtained by normalizing the observed activity according to the noise-distribution in that location:

$$z_r = \frac{r - m_0}{s_0} \quad (9)$$

where r is the observed source activity and m_0 and s_0 are the mean and the standard deviation, respectively, of the noise distribution. Finally, these noise-normalized activity maps were averaged across subjects.

Results

Behavioral results

During the perceptual decision task, subjects reported perceiving the grating when it was present on 71% of trials (SD = 11%, Fig. 1C), and falsely reported perceiving it when it was not present on 14% of trials (SD = 10%), leading to a sensitivity d' of 1.8 (SD = 0.56) and a criterion c of 0.3 (SD = 0.37), indicating a slightly conservative decision bias. The average noise level across subjects, obtained from the staircase, was 95.5% (SD = 0.5%).

In the sensory processing blocks, subjects detected the oddball trials in 86% (SD = 9.5%) of the cases, while falsely reporting an oddball in only 0.3% (SD = 0.3%) of the non-oddball trials. This above-chance, yet imperfect performance verified that subjects adhered to the instructions and were well able to do the task, while also showing that the task was sufficiently difficult to require attentional engagement.

Identifying sensory-specific neural processing

We first investigated whether we could reliably extract a neural signature that is specific to sensory processing. To this end, we focused on the neural responses during the

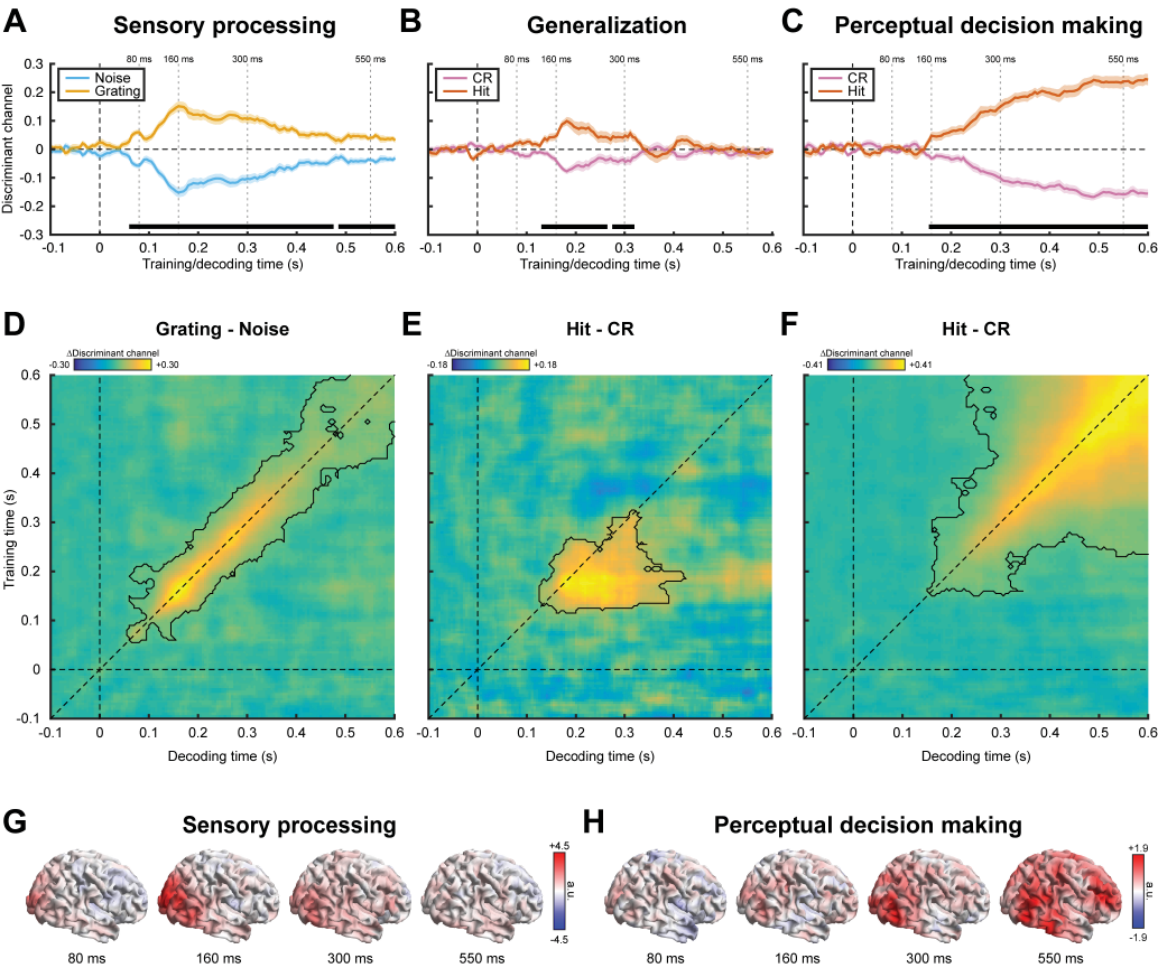


Figure 2. Decoding results for training and decoding within the functional localizer (**A**, **D**), for generalization from the functional localizer to correct perceptual decisions (**B**, **E**) and for training and decoding within correct perceptual decisions (**C**, **F**). (**A-C**), Average discriminant channel activity for noise and grating trials (**A**) and CRs and hits (**B-C**), at matched training and decoding time. The horizontal black bars mark time points that belong to a significant cluster as outlined in (**D-F**). Shaded areas depict the SEM. (**D-F**), Temporal generalization matrices. Note that the diagonals are identical to the differences between the curves in (**A-C**) and note that the color scales are variable across figure for optimal visualization. Significant clusters are demarcated by the contours. (**G-H**), Source level contributions to the discriminant channel trained on the functional localizer (**G**) and on perceptual decision making (**H**), at four specific time points.

sensory processing blocks. In these blocks, subjects were presented with noise stimuli, half of which contained an embedded grating. As these stimuli were not relevant for the task at hand, neural activity during these blocks is thus specific to sensory processing, uncontaminated by decision processes. Moreover, attention was drawn away from the stimuli towards a task at fixation, thereby minimizing potential effects of attentional modulation on the sensory processing. Although there is the theoretical possibility that the stimuli nevertheless drew attention in an automatic fashion, this bottom-up effect is presumably much weaker than the voluntary attentional engagement during perceptual decision making, and therefore most likely does not pose a problem for the present study.

We applied our decoding analysis to these data and extracted a discriminant channel that is maximally discriminative between noise and grating trials (Fig. 2A; see Supplementary Figures S1 and S2 for the same results expressed as classification accuracy and Cohen's d , respectively). The results show that these two conditions can be reliably discriminated on the basis of the MEG recordings as evidenced by a significant cluster ($p < 0.001$) that extends from 60 ms post stimulus onset throughout the rest of the segment. The discriminability peaks at 80 ms and 160 ms, after which it gradually decays back toward baseline.

Next, we calculated the temporal generalization matrix of the discriminant channel to probe the temporal organization of the neural processing stages underlying sensory processing. This matrix contains the activity of the discriminant channel over time, while trained on all other time points.

That means that the temporal generalization matrix provides insight into whether the conditions can be discriminated at some time point on the basis of the weights that are maximally discriminative at some other time point. The results show that, during sensory processing, the temporal generalization profile is largely located around the diagonal (Fig. 2D), indicating that a given discriminant channel only generalizes to temporally proximate neural signals. This is known as a "chain" profile (King and Dehaene, 2014) and suggests that distinct, sequential neural processes are involved in the encoding of gratings versus noise. In order to interpret what these neural processes are, we assessed the source-level projection of the discriminant channel (Fig. 2G). This projection depicts the contribution of neural sources to the discriminant channel, and thus provides insight into the brain areas involved in the encoding of the stimuli. The results show that contributions stem primarily from occipital areas at both discrimination peaks, but especially prominently for the later one.

To summarize, we found that we could reliably extract a neural signature over time that is characteristic of sensory processing, in absence of decision making. This neural signature is evident throughout most of the trial and peaks at 80 and 160 ms. It encapsulates distinct, sequential neural sources over time, and these are located primarily in occipital cortex, consistent with sensory processing.

Tracing sensory-specific processing during perceptual decision making

We proceeded to trace the sensory-specific neural signature while subjects were engaged in perceptual decision making. For this, we trained the decoders on the sensory processing blocks, serving as a functional localizer, to discriminate between noise and gratings and subsequently applied these decoders to the neural signals obtained during perceptual decision making.

We first focused on the trials in which subjects correctly reported the presence (hits) or absence (CRs) of a grating and asked whether these categories can be reliably distinguished on the basis of sensory processing. Fig. 2B depicts the discriminant channel activity for these two conditions. We found a significant cluster ($p = 0.006$), indicating that these categories are indeed different in terms of sensory processes. Interestingly, this cluster extended from 130 to 320 ms, whereas no differences between CRs and hits were found for earlier or later time points. In addition, the cluster exhibited a relative late onset as compared to the early onset of discriminability in the sensory processing blocks. This inability to discriminate between CRs and hits before 130 ms and after 320 ms is likely due to the threshold-visibility of the gratings.

Sensory processing during perceptual decision making was found to be qualitatively different from during the functional localizer, as revealed by the temporal generalization matrix (Fig. 2E). While discriminability between CRs and hits is significant along the diagonal, i.e. when training and decoding times are matched, a substantial portion of the temporal generalization profile is found below the diagonal. This elongated shape, known as a “sustained” profile (King and Dehaene, 2014) indicates that CRs and hits can be differentiated throughout an extended period of time, ranging from approximately 130 up to 400 ms, using the weights that discriminated between gratings and noise trials during the shorter interval of approximately 130 to 250 ms. In other words, the sensory representation of the stimulus, as defined during this relatively early period, is maintained over time.

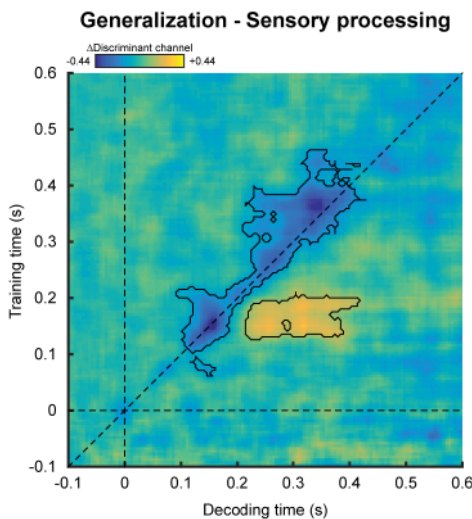


Figure 3. Comparison of the temporal generalization matrices depicting sensory processing during perceptual decision making (Fig. 2E) and depicting unattended sensory processing (Fig. 2D). Significant clusters are demarcated by the contours.

Direct comparison of the between-task generalization (Fig. 2E) to the temporal generalization matrix depicting sensory processing only (Fig. 2D) indeed revealed a significant positive cluster ($p = 0.04$) that extended from approximately 200 to 400 ms decoding time and 130 to 200 ms training time (Fig. 3). In addition, a significant negative cluster ($p = 0.003$) was found around the diagonal that extended from approximately 100 ms to 400 ms, showing that the diagonal decoding performance is worse during perceptual decision making than during the functional localizer. This difference is most likely due to the increased stimulus noise level during the perceptual decision making task, as compared to the localizer. In summary, these results suggest that although the overall sensory representation was weaker during perceptual decision making, the early part of this representation was stabilized and kept online in the visual system, but only when attended and/or required for a subsequent decision.

Tracing the neural signatures of correct perceptual decisions

We also examined whether we could discriminate between the different stimulus/response-categories within the perceptual decision-making task, by performing an analysis akin to more conventional approaches, in which we did not make use of the sensory processing blocks as a functional localizer but looked at the data obtained during perceptual decision making only.

We trained the decoders to discriminate between CRs and hits and calculated the average amplitude of the discriminant channel for these categories. We found that neural processes associated with these categories begin to diverge at 155 ms and, contrary to the generalization analysis in which we decoded sensory-specific processing only, this difference remained significant ($p < 0.001$) throughout the rest of the trial (Fig. 2C). As CRs differ from hits not only in terms of the presented stimulus, but also with respect to the behavioral decision, the later activity differences likely reflect differences in decision-related activity - activity to which the decoders trained on the sensory processing blocks were blind.

The interpretation of the later activity as representing processes distinct from sensory encoding is corroborated by the temporal generalization matrix. In addition to the rising discriminability along the diagonal, we observed long-lasting sustained activity throughout the interval of approximately 250 ms post-stimulus until the end of the trial, meaning that CRs could be differentiated from hits during this interval using the weights obtained from any other time point. Direct comparison of the between-task generalization matrix in this same interval. This “sustained” profile (King and Dehaene, 2014), that can be observed in addition to a chain profile along the diagonal, shows that at least some of the neural sources remain active throughout the rest of the trial. Earlier work in monkeys demonstrated that neurons encoding for the monkey’s decision remain active in a sustained manner when the behavioral report of the decision is postponed over a delay period (Roitman and Shadlen, 2002). Thus, our results are consistent with the interpretation that the later activity reflects

decision-related processing. Conversely, earlier discriminant channel activity prior to this sustained activity, up to approximately 250 ms, does not show this widespread temporal generalization, suggesting that different processes are at work in this time window, possibly transient sensory processes.

Finally, to substantiate these interpretations, we asked which brain areas contribute to the discrimination between CRs and hits. At both 160 ms and 300 ms, discriminant activity is observed selectively over occipital cortex. Interestingly, and in line with the interpretation that the later signal reflects a decision process, discriminant activity at 550 ms is much more widespread and encompasses parietal and frontal cortices, which are often implicated in decision making (Gold and Shadlen, 2007; Heekeren et al., 2008). It is also noteworthy that the contribution in occipital regions to the discriminant channel increases from 160 ms to 300 ms when a perceptual decision is made, whereas it decreases between these two time points in the functional localizer blocks. Indeed, this is in agreement with the observation that the sensory representation in the brain is enhanced and maintained over time when required for the task at hand.

Sensory processing during perceptual decision errors

So far we focused exclusively on correct perceptual decisions - i.e., CRs and hits, which differ on both stimulus and decision dimensions. We extended our previous between-block generalization analysis to the incorrect perceptual decisions and decoded FAs and misses, in addition to CRs and hits, using decoders that were trained on neural signals obtained during sensory processing (Fig. 4 A-D; see Supplementary Figures S3 and S4 for the same results expressed as classification proportions and Cohen's d , respectively). These categories were then compared to CRs and hits to assess how incorrect perceptual decisions deviate from correct ones with respect to sensory processing. This is interesting, because there has been considerable debate in the literature whether perceptual decision errors stem from faulty sensory encoding (Rees and Heeger, 2003; Jolij et al., 2011; Zhang et al., 2008) or are instead the result of noise in supramodal decision making (Ress et al., 2000; Hulme et al., 2009; Deco and Romo, 2008). Our approach using the between-task generalization is in a position to resolve this question, as it was specifically designed to probe sensory processing only.

We found that misses could be reliably discriminated from CRs ($p = 0.007$, Fig. 4B), and hits from FAs ($p = 0.016$, Fig. 4C). As was the case for the comparison of CRs to hits described above (Fig. 2E), the time period during which a discriminative signal was present ranged from approximately 150 ms to 300 ms. Furthermore, the temporal generalization matrix of these comparisons also displayed an elongated, below-diagonal profile. In contrast, no significant differentiation was obtained between the contrasts of FA versus CR (Fig. 4A) and hits versus misses (Fig. 4D). Although a trend may be discerned in the latter contrast, this did not constitute a significant cluster. In short, the decoders were only able to discriminate between conditions that are different with respect

to stimulus, but not between conditions that differ on decision. Therefore, given that these decoders were designed to target sensory processing only, these results show that the encoded sensory information accurately reflects the physical stimulus, even when an incorrect decision follows. These findings suggest that, in the current study, perceptual decision errors stemmed mainly from later decision-related processes, rather than from faulty sensory encoding.

We also subjected the perceptual decision errors to the within-task approach in which we trained on CRs and hits within perceptual decision making and used these weights to decode the perceptual decision errors (Fig. 4E-H). We found that FAs could be reliably separated from CRs during a relatively late time period (Fig. 4E; solid cluster: $p = 0.025$, dotted cluster: $p = 0.05$), but not during an earlier time period, consistent with the interpretation that the discriminant channel in the late period reflects decision processes. Interestingly, the situation is different when comparing hits to misses. Although these categories are also expected to differ only during the later decision period, we found that the significant cluster extended to earlier time points ($p < 0.001$, Fig. 4H).

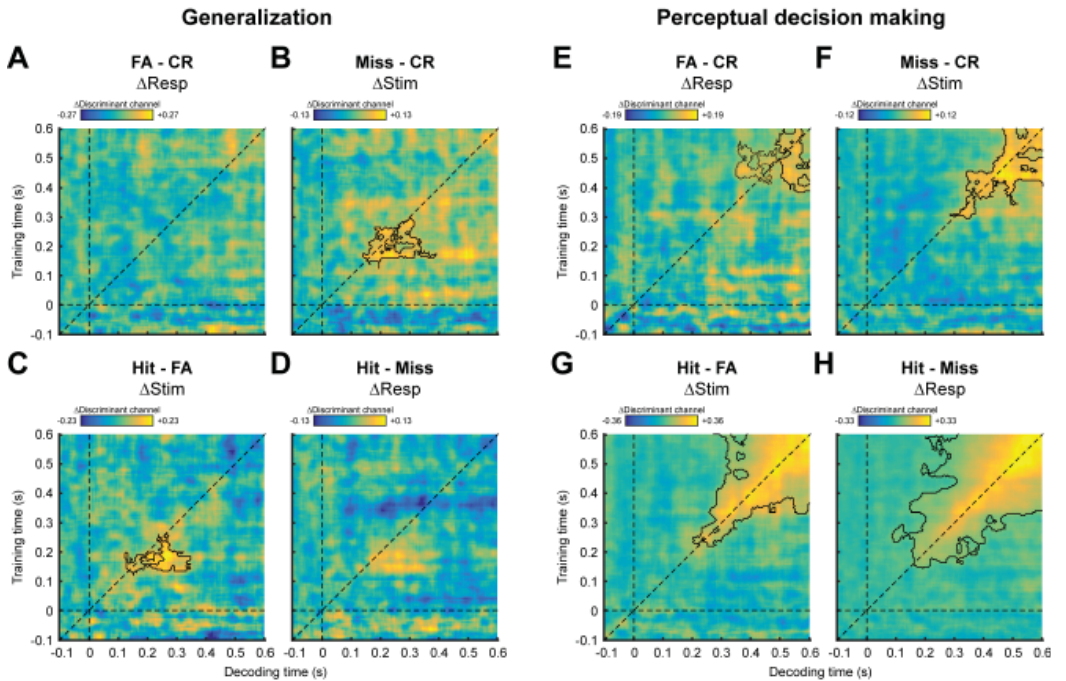


Figure 4. Temporal generalization matrices comparing correct to incorrect perceptual decisions, for generalization from the functional localizer to perceptual decision making (A-D) and for training and decoding within perceptual decision making (E-H). The subtitles emphasize the factor on which the contrasts vary. Note that the color scales are variable across figure for optimal visualization. Significant clusters are demarcated by the contours.

More discrepancies were found when comparing misses to CRs (Fig. 4F) and hits to FAs (Fig. 4G). As these categories differ only in terms of presented stimulus, they are expected to be discriminable solely during the early sensory period. This is however not what we found. Instead, these comparisons revealed significant clusters in late time windows (CR versus miss: $p = 0.013$; FA versus hit: $p < 0.001$), but not in early ones, suggesting that these categories differ in decision-related neural processing, despite identical behavioral response. These discrepancies highlight the difficulties in disentangling sensory from decision processes when only considering neural data obtained during perceptual decision making, without making use of a functional localizer, and we elaborate on this issue in the *Discussion*.

Finally, to consider the possibility that these discrepancies may have arisen from the fact that we trained on CRs versus hits instead of noise versus grating (as in the localizer), we conducted an additional analysis in which we trained on stimulus only, irrespective of the response, and used this decoder to compare the four categories (Supplementary Fig. S5). These analyses resulted in the same discrepancies mentioned above. FAs are significantly different from hits ($p < 0.001$), but not from CRs - in line with what one would expect, given that these decoders should be sensitive to differences in stimulus only. Again however, there were significant differences between misses and hits ($p = 0.004$), but not between CRs and misses. Moreover, all of these differences were confined to relatively late time windows ($> \pm 300$ ms), whereas no significant differences were found before that. These results are at odds with the results obtained from the between-task generalization, where we only found effects during early time windows (approximately 130 to 320 ms), and indeed highlight the utility of using a functional localizer.

Discussion

In the present study we sought to dissociate sensory processing from decision-related processing during perceptual decision making. We accomplished this using a novel approach where we employed a functional localizer task to identify the neural signature specific to sensory processing, and used this to trace the temporal trajectory of sensory encoding during perceptual decision making. Our results revealed a temporal dissociation between sensory- and decision-related neural activity. We found that that sensory information was encoded in neural signals during a relative early time window that extended from 130 to approximately 350 ms post stimulus, and that this encoded sensory information correctly reflected the physical stimulus even in the case of an incorrect decision. In contrast, we found a later, sustained decision-related neural process that extended from approximately 250 until at least 600 ms and become more pronounced over time, in agreement with previous reports (O'Connell et al., 2012; Kelly and O'Connell, 2013). Moreover, although gratings could be distinguished for a longer period of time in the functional localizer as compared to perceptual decision making, the sensory representation in the former was unstable and changing over time. During perceptual decision making on the other hand, the early sensory representation was stable for an extended period of time. Thus, these

results show that the early sensory representation was stabilized and maintained over time when required for a decision, but not when the stimulus was unattended and task-irrelevant.

A number of previous studies have also focused on disentangling sensory and decision processes (Poliakoff and Sajda, 2006; Poliakoff et al., 2006; Ratcliff et al., 2009; O'Connell et al., 2012; Wyart et al., 2012; de Lange et al., 2013; Kelly and O'Connell, 2013). For instance, Wyart et al. (2012) orthogonalized these processes by randomizing the decision-relevant information of the stimulus on a trial-to-trial basis, such that it did not correlate with the raw sensory information. This led to a temporal dissociation in which sensory-related signals preceded decision-related activity, similarly to our results. However, whereas these studies relied on external manipulation of the stimulus while assuming constant internal sensory processing, we instead kept the stimulus constant and capitalized on ongoing fluctuations in perception and/or decision making in order to extract decision-related activity. This paradigm is widely employed for a variety of purposes, for instance to uncover the neural correlates of consciousness (Crick and Koch, 1998; Salti et al., 2015; Schurger et al., 2015) or to extract internal perceptual templates (Gosselin and Schyns, 2003; Smith et al., 2012). However, as explained in the *Introduction*, using behavioral report as independent factor introduces interpretational limitations, as it is an observed variable and therefore not under experimental control. One commonly used way to facilitate interpretation is to delineate neural activity in the spatial domain (e.g. Ref. 15). For instance, activity in motion sensitive area MT is parametrically modulated by visual motion strength (Britten et al., 1992; Siegel et al., 2007) and may therefore be defined as encoding for sensory evidence. Similarly, activity in parietal areas has been found to exhibit characteristics of integration toward a decision boundary, with a rate proportional to the signal strength (O'Connell et al., 2012; Kelly and O'Connell, 2013; Roitman and Shadlen, 2002), and may therefore be defined as encoding for the decision process. However, dissociation on the basis of spatial location may be fallacious, because the activity of sensory neurons does not only reflect the stimulus, but can also reflect decision processes due to interactions between cognitive processes and sensory neurons (Nienborg and Cumming, 2009; Nienborg and Cumming, 2014; Nienborg et al., 2012). Therefore, merely recording from sensory areas may be insufficient to disentangle sensory processing from decision-related activity. In fact, these results point toward a more general, conceptual problem. Namely, that perceptual decisions are not the result of a sequential processing pipeline, but rather stem from complex, reciprocal interactions within a large network of areas (Wimmer et al., 2015). It is therefore conceptually challenging to unambiguously identify the neural signals underlying sensory processing when these signals are also used for decision making. This holds for all brain recording methods, as even the availability of superior spatial resolution, such as in single-cell recordings, does not resolve this problem.

We attempted to counter this problem by means of a separate functional localizer, which allowed us to unambiguously define the sensory-specific neural signature, in the

absence of decision making or attentional modulation. The use of a functional localizer is conceptually similar to a common practice in single-cell recordings, where the tuning properties of neurons are mapped in a separate session. Our approach extends this idea to human neuroimaging, and yielded two important results. First, we were able to address the following question: how does the brain reliably integrate sensory evidence if the stimulus is available for only a very limited amount of time? According to the sequential sampling framework, the decision variable is constructed by means of sequential sampling of the sensory evidence over time (Gold and Shadlen, 2007; Ratcliff and McKoon, 2007). This makes sense in the case where a stimulus is presented for a prolonged period of time, such as in the commonly used random-dot motion stimulus. However, it is not intuitive how sequential sampling should proceed after a brief stimulus has disappeared. Computational modeling work (Ratcliff and Rouder, 2000) suggests that sensory evidence has to remain available to the accumulator after stimulus offset in order to fit the observed data. Here we present a direct experimental demonstration of this assumption. Our results show that the sensory representation is maintained over time during perceptual decision making, but not when the stimulus is viewed while attention is directed away from it. Thus, it appears that the brain actively stabilizes the sensory representation, presumably by means of top-down mechanisms such as attention (Ress et al., 2000) or working memory (Ratcliff and Rouder, 2000; Harrison and Tong, 2009), when it is required for a subsequent decision.

Second, when comparing incorrect perceptual decisions (FAs and misses) to correct decisions (CRs and hits), we observed discrepancies between the decoded neural signals in the case where the decoders were trained on the perceptual decision making data and the case where the decoders were trained on the functional localizer. In the latter, we only found differences between the stimulus/response-categories that varied in terms of stimulus (CR versus miss; FA versus hit) and not between the categories in which physical stimulation was identical (CR versus FA; miss versus hit). Indeed, these differences were exclusive to a relatively early time window, as would be expected given that this early time period reflects sensory processing. However, different results were obtained when the decoders were trained on the neural signals recorded during correct perceptual decisions. In this case we did not observe early differences between CRs and misses, but we did find early differences between misses and hits. Thus, these results suggest that early neural activity reflects the eventual decision, rather than physical stimulation.

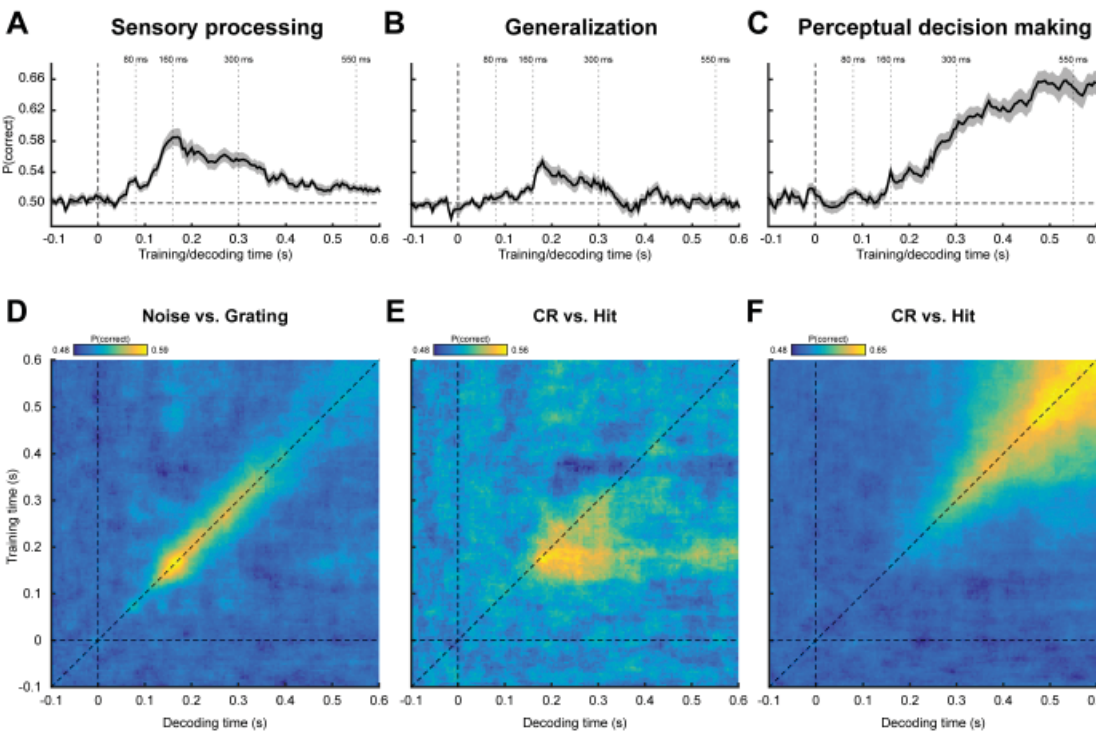
This paradox can be resolved by considering the data sets on which the decoders were trained in each case. In the case of the functional localizer, the conditions in the training data (noise versus grating) varied only with respect to physical stimulation. When training on the perceptual decision making data however, the conditions in the training data (CR and hit) diverged not only with respect to physical stimulation, but also with respect to behavioral decision. As behavioral decision is an observed variable, it is not under experimental control and its effect is therefore susceptible to confounding variables. One possible instantiation of such a confound would be trial-to-trial fluctuations in attention (Ress et al., 2000; van Dijk et al., 2008; Supèr et al., 2003), that modulate the likelihood

that an encoded grating is successfully transferred into the decision process. This would explain the discrepancy, because hits would be inherently accompanied with a different ongoing attentional state relative to misses, for else they would have been classified as a miss. No such bias would be present for the CRs, because no grating is encoded in these trials. Together with the threshold-visibility of the grating, we therefore suggest that the decoders that were discriminative between CRs and hits at early time points were primarily sensitive to these fluctuations, rather than to physical stimulation. As these fluctuations have an impact on the eventual behavioral decision, this provides an explanation for the decision-related activity at early time points when training on CRs versus hits. Indeed, it has recently been proposed that decision-related activity in sensory neurons may result from ongoing fluctuations in higher-level expectations about the upcoming stimulus, whose activity is projected back to lower-level sensory areas (Nienborg and Cumming, 2009; Pooresmaeili et al., 2014; Nienborg and Roelfsema, 2015).

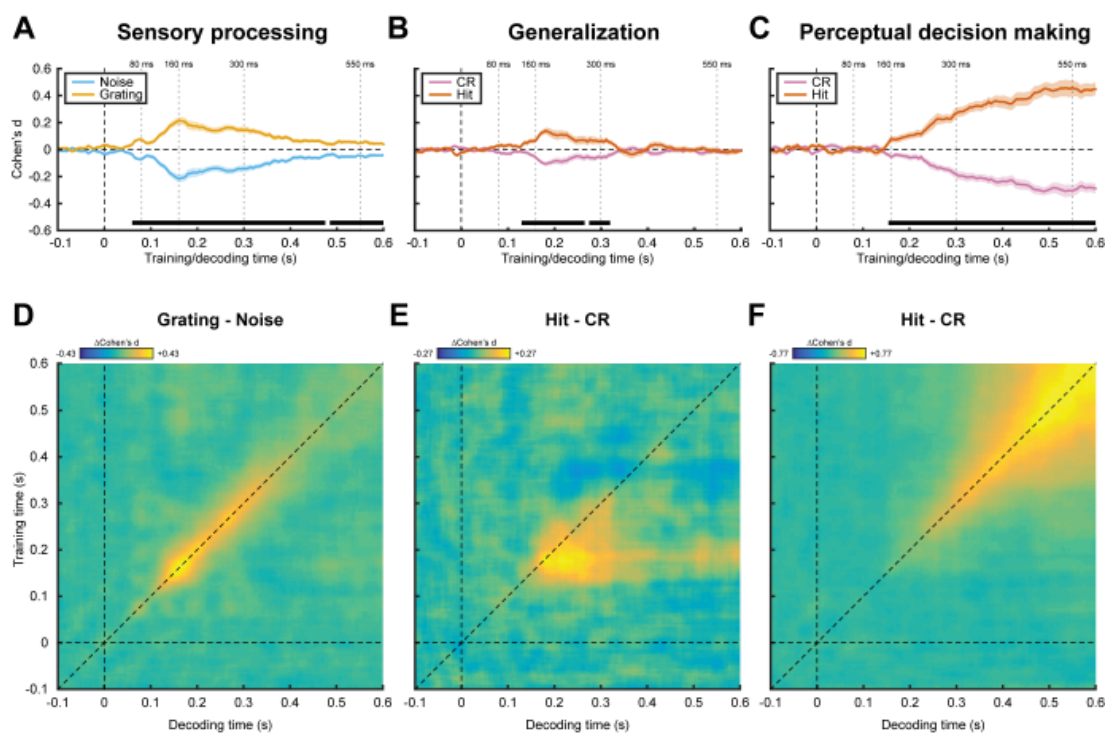
Finally, although we interpreted the later discriminant channel activity as reflecting the decision process, we nevertheless found differences in this time window between categories that are identical in terms of behavioral response. One explanation for this discrepancy may be that the integration of sensory evidence during incorrect decisions follows a deviating trajectory as compared to correct decisions (Ratcliff and McKoon, 2007; Boldt and Yeung, 2015). A second explanation is that subjective confidence in a perceptual decision is likely to be different between correct and incorrect decisions and, given that subjective confidence is manifested in electroencephalography signals (Boldt and Yeung, 2015), may therefore have given rise to the observed differential activity.

In conclusion, we presented an empirical dissociation between sensory processes and decision-related processes during human perceptual decision making. Our results are largely consistent with previous findings, but also provide new insights into the mechanisms underlying perceptual decision making. Our results suggest that a sensory representation is maintained when required for the task at hand and/or attended. Importantly, we also found that sensory processes accurately encode the physical stimulus during perceptual decision errors. We believe that our approach, as well as the insights obtained with it, make an important contribution to various fields of study, including that of perceptual decision making, the neural correlates of consciousness and visual cognition.

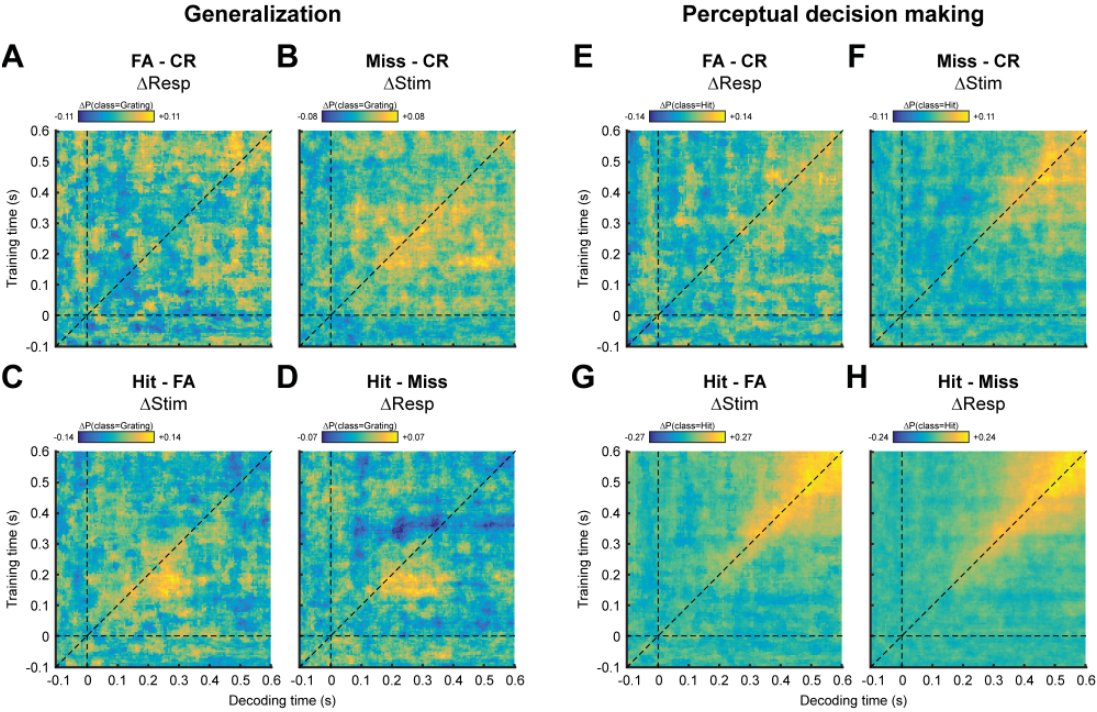
Supplementary information



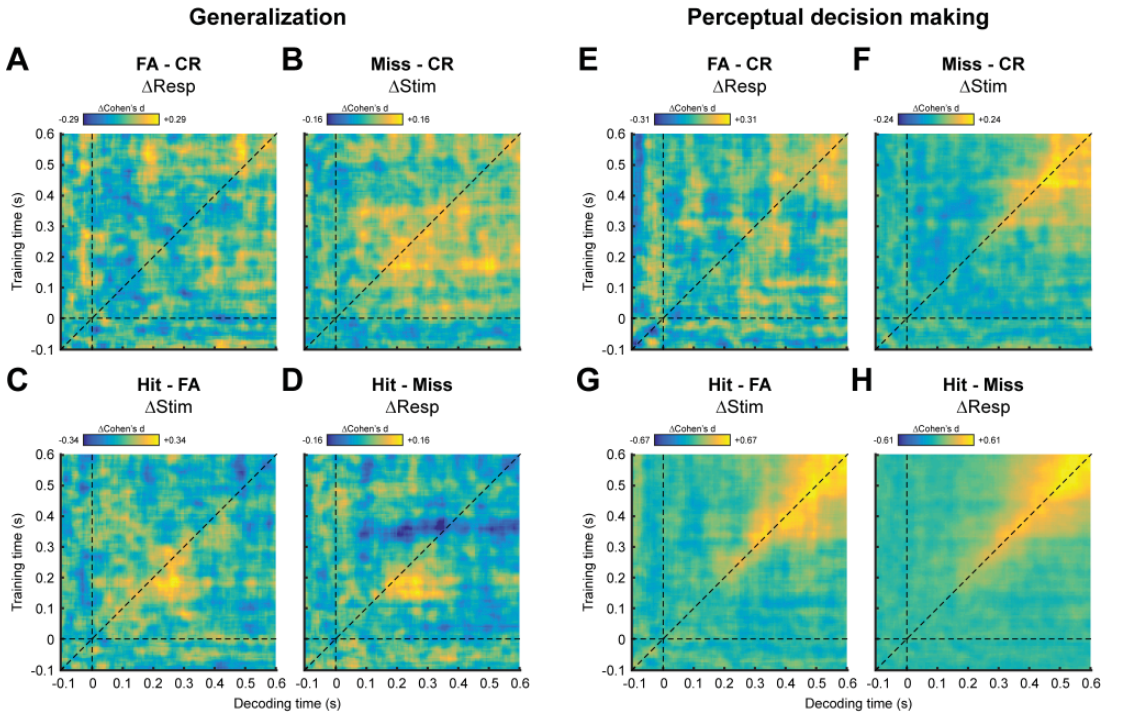
Supplementary Figure S1. The same results as presented in Figure 2A-F, but instead expressed as classification accuracy. For the between-task generalization, hits and CRs were considered correct when classified as gratings and noise, respectively. No statistical tests were conducted on these results.



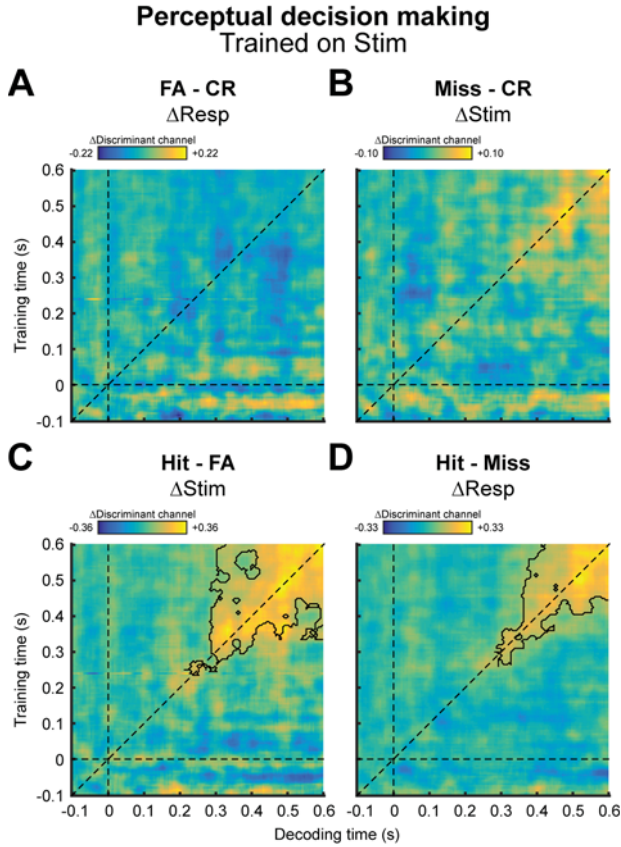
Supplementary Figure S2. The same results as presented in Figure 2A-F, but instead expressed as Cohen's d. No statistical tests were conducted on these results.



Supplementary Figure S3. The same results as presented in Figure 4, but instead expressed as the proportion of trials within each of the four stimulus/response-categories that were classified as grating (A-D) or as hit (E-H). No statistical tests were conducted on these results.



Supplementary Figure S4. The same results as presented in Figure 4, but instead expressed as Cohen's d . No statistical tests were conducted on these results.



Supplementary Figure S5. The results of an analysis akin to Figure 4, but instead the decoders were trained to discriminate between grating present versus absent, irrespective of the subject's decision (as opposed to training on CRs versus hits). However, as the numbers of trials are unequal across the four categories, the stimulus presence is correlated to decision. To counter this, we trained on the unweighted average of CRs and FAs (i.e. stimulus absent) versus the unweighted average of misses and hits (i.e. stimulus present). Specifically, in Eq. 5 (see *Methods*), this corresponds to $\hat{\mu}_1 = \frac{1}{2}(\hat{\mu}_{\text{CR}} + \hat{\mu}_{\text{FA}})$ and $\hat{\mu}_2 = \frac{1}{2}(\hat{\mu}_{\text{Miss}} + \hat{\mu}_{\text{Hit}})$. Similarly, the common covariance matrix was calculated as the unweighted average of the four individual covariance matrices within each category.



3

**Prior expectations induce
pre-stimulus sensory templates**

Abstract

Perception can be described as a process of inference, integrating bottom-up sensory inputs and top-down expectations. However, it is unclear how this process is neurally implemented. It has been proposed that expectations lead to pre-stimulus baseline increases in sensory neurons tuned to the expected stimulus, which in turn affects the processing of subsequent stimuli. Recent fMRI studies have revealed stimulus-specific patterns of activation in sensory cortex as a result of expectation, but this method lacks the temporal resolution necessary to distinguish pre- from post-stimulus processes. Here, we combined human MEG with multivariate decoding techniques to probe the representational content of neural signals in a time-resolved manner. We observed a representation of expected stimuli in the neural signal shortly before they were presented, demonstrating that expectations indeed induce a pre-activation of stimulus templates. The strength of these pre-stimulus expectation templates correlated with participants' behavioural improvement when the expected feature was task-relevant. These results suggest a mechanism for how predictive perception can be neurally implemented.

This chapter has been published as:

Kok, P., Mostert, P. & de Lange, F. P. (2017) Prior expectations induce pre-stimulus sensory templates. *Proceedings of the National Academy Sciences* 114(39), 10473-10478. doi: 10.1073/pnas.1705652114.

Introduction

Perception is heavily influenced by prior knowledge (von Helmholtz, 1866; Gregory, 1997; Kersten et al., 2004). Accordingly, many theories cast perception as a process of inference, integrating bottom-up sensory inputs and top-down expectations (Lee and Mumford, 2003; Friston, 2005; Summerfield and de Lange, 2014). However, it is unclear how this integration is neurally implemented. It has been proposed that prior expectations lead to baseline increases in sensory neurons tuned to the expected stimulus (Wyart et al., 2012; SanMiguel et al., 2013; Kok et al., 2014), which in turn leads to improved neural processing of matching stimuli (Bell et al., 2016; Kok et al., 2012). In other words, expectations may induce stimulus templates in sensory cortex, prior to the actual presentation of the stimulus. Alternatively, top-down influences in sensory cortex may exert their influence only after the bottom-up stimulus has been initially processed, and the integration of the two sources of information may become apparent only during later stages of sensory processing (Rao et al., 2012).

The evidence necessary to distinguish between these hypotheses has been lacking. fMRI studies have revealed stimulus-specific patterns of activation in sensory cortex as a result of expectation (Kok et al., 2014; Hindy et al., 2016), but this method lacks the temporal resolution necessary to distinguish pre- from post-stimulus periods. Here, we combined MEG with multivariate decoding techniques to probe the representational content of neural signals in a time-resolved manner (Cichy et al., 2014; King and Dehaene, 2014; Mostert et al., 2015; Myers et al., 2015). The experimental paradigm was virtually identical to the ones employed in our previous fMRI studies that studied how expectations modulate stimulus-specific patterns of activity in the primary visual cortex (Kok et al., 2012, 2014). We trained a forward model to decode the orientation of task-irrelevant gratings from the MEG signal (Brouwer and Heeger, 2009, 2011), and applied this decoder to trials in which participants expected a grating of a particular orientation to be presented. This analysis revealed a neural representation of the expected grating that resembled the neural signal evoked by an actually presented grating. This representation was present already shortly before stimulus presentation, demonstrating that expectations can indeed induce the pre-activation of stimulus templates.

Results

Participants ($N = 23$) were exposed to auditory cues that predicted the likely orientation (45° or 135°) of an upcoming grating stimulus (Fig. 1A, B). This grating was followed by a second grating that differed slightly from the first in terms of orientation and contrast. In separate runs of the MEG session, participants performed either an orientation or contrast discrimination task on the two gratings (see *Materials and Methods* for details).

Behavioural results

Participants were able to discriminate small differences in orientation ($3.9^\circ \pm 0.5^\circ$, accuracy = $74.0\% \pm 1.6\%$, mean \pm SEM) and contrast ($4.6\% \pm 0.3\%$, accuracy = $76.6\% \pm 1.5\%$) of the cued gratings. There was no significant difference between the two tasks in terms of either accuracy ($F_{1,22} = 3.38$, $p = 0.080$) or reaction time (mean RT = 633 ms vs. 608 ms, $F_{1,22} = 2.89$, $p = 0.10$). Overall, accuracy and reaction times were not influenced by whether the cued grating had the expected or the unexpected orientation (accuracy: $F_{1,22} = 0.21$, $p = 0.65$; RT: $F_{1,22} < 0.01$, $p = 0.93$), nor was there an interaction between task and expectation (accuracy: $F_{1,22} = 0.96$, $p = 0.34$; RT: $F_{1,22} = 0.09$, $p = 0.77$). Note that these discrimination tasks were orthogonal to the expectation manipulation, in the sense that the expectation cue provided no information about the likely correct choice.

During the grating localiser (Fig. 1C, see *Materials and Methods* for details), participants correctly detected $91.2\% \pm 1.6\%$ (mean \pm SEM) of fixation flickers, and incorrectly pressed the button on $0.2\% \pm 0.1\%$ of trials, suggesting that participants were successfully engaged by the fixation task.

MEG results – Localiser orientation decoding

As mentioned, participants were exposed to auditory cues that predicted the likely orientation of an upcoming grating stimulus. The question we wanted to answer was whether the expectations induced by these auditory cues would evoke templates of the visual stimuli prior to the presentation of the gratings. To be able to uncover such sensory templates, we trained a decoding model to reconstruct the orientation of (task-irrelevant) visual gratings (Fig. 1C) from the MEG signal, in a time-resolved manner. First, we found that this model was highly accurate at reconstructing the orientation of such gratings from the MEG signal (Fig. 2). Grating orientation could be decoded across an extended period of time (from 40 to 655 ms post-stimulus, $p < 0.001$, and from 685 to 730 ms, $p = 0.018$), peaking around 120-160 ms post-stimulus (Fig. 2C). Furthermore, in the period around 100 to 330 ms post-stimulus, orientation decoding generalised across time, meaning that a decoder trained on the evoked response at, for example, 120 ms post-stimulus could reconstruct the grating orientation represented in the evoked response around 300 ms, and vice versa (Fig. 2D). In other words, certain aspects of the representation of grating orientation were sustained over time.

MEG results – Expectation induces stimulus templates

Our main question pertained to the presence of visual grating templates induced by the auditory expectation cues during the main experiment. Therefore, we applied our model trained on task-irrelevant gratings to trials containing gratings that were either validly or invalidly predicted, respectively (Fig. 3A). In both conditions, the decoding model trained on task-irrelevant gratings succeeded in accurately reconstructing the orientation of the

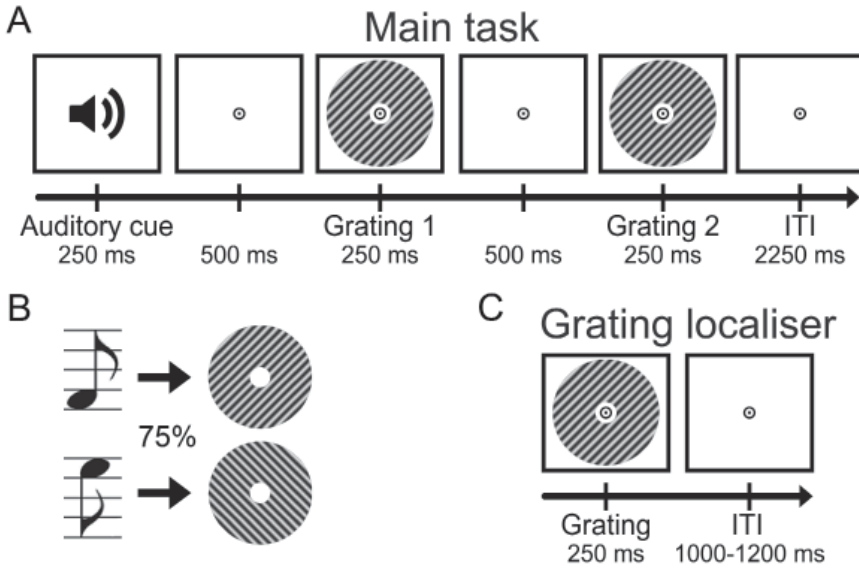


Figure 1. Experimental paradigm. **(A)** Each trial started with an auditory cue that predicted the orientation of the subsequent grating stimulus. This first grating was followed by a second one, which differed slightly from the first in terms of orientation and contrast. In separate runs, participants performed either an orientation or contrast discrimination task on the two gratings. **(B)** Throughout the experiment, two different tones were used as cues, each one predicting one of the two possible orientations (45° or 135°) with 75% validity. These contingencies were flipped halfway through the experiment. **(C)** In separate grating localiser runs, participants were exposed to task-irrelevant gratings while they performed a fixation dot dimming task.

gratings presented in the main experiment (valid expectation: cluster from training time 60 to 410 ms and decoding time 60 to 400 ms, $p < 0.001$, and from training time 205 to 325 ms and decoding time 400 to 495 ms, $p = 0.045$; invalid expectation: cluster from training time 75 to 225 ms and decoding time 75 to 330 ms, $p = 0.0012$, and from training time 250 to 360 ms and decoding time 195 to 355 ms, $p = 0.027$).

If the cues induced sensory templates of the expected grating, one would expect these to be revealed in the difference in decoding between valid and invalidly predicted gratings (see *Material and Methods* for details of the subtraction logic). Indeed, this analysis demonstrated that the auditory expectation cues induce orientation-specific neural signals (Fig. 3A, bottom panel). These signals were present already 40 ms before grating presentation, and extended into the post-stimulus period (from decoding time -40 to 230 ms, $p = 0.0092$, and from 300 to 530 ms, $p = 0.016$). Furthermore, these signals were uncovered when the decoder was trained on around 120 to 160 ms post-stimulus during the grating localiser (Fig. 3B), suggesting that these cue-induced signals were similar to those evoked by task-irrelevant gratings. In other words, the auditory expectation cues

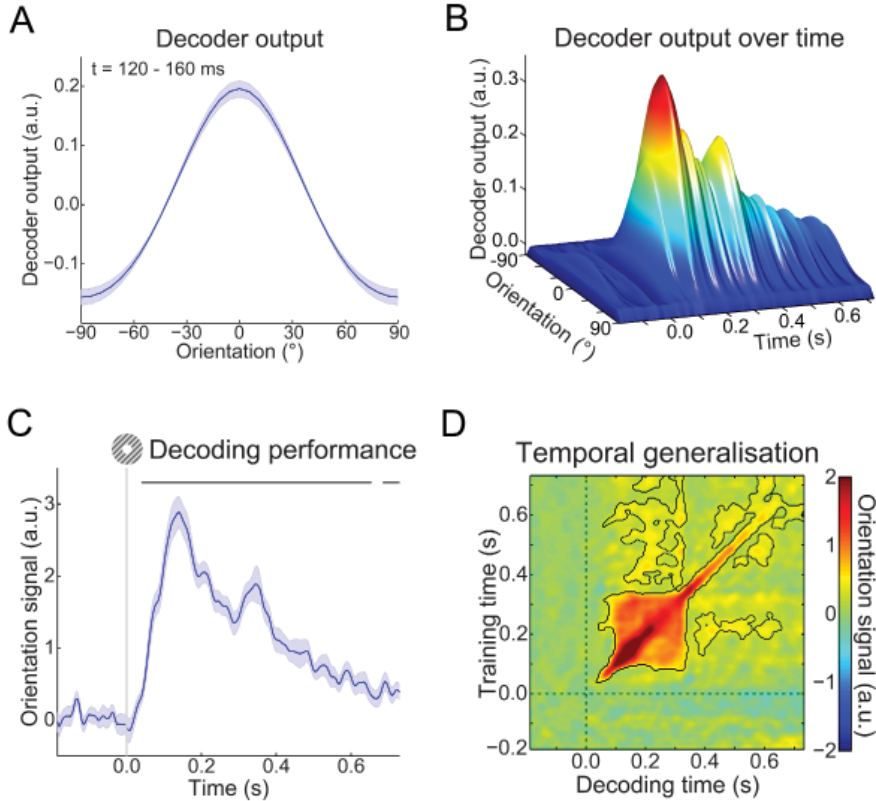


Figure 2. Localiser orientation decoding. **(A)** The output of the decoder consisted of the responses of 32 hypothetical orientation channels, shown here decoders trained and tested on the MEG signal 120-160 ms post-stimulus during the grating localiser (cross-validated). Shaded region represent SEM. **(B)** Decoder output over time, trained and tested in 5 ms steps (sliding window of 29.2 ms), showing the temporal evolution of the orientation signal. **(C)** The response of the 32 orientation channels collapsed into a single metric of decoding performance (see *Supporting Materials and Methods*), over time. Shaded region represent SEM, horizontal lines indicate significant clusters ($p < 0.05$). **(D)** Temporal generalisation matrix of orientation decoding performance, obtained by training decoders on each time point, and testing all decoders on all time points (as above, steps of 5 ms and a sliding window of 29.2 ms). This method provides insight into the sustained versus dynamical nature of orientation representations (King and Dehaene, 2014). Solid black lines indicate significant clusters ($p < 0.05$), dashed lines indicate grating onset ($t = 0s$).

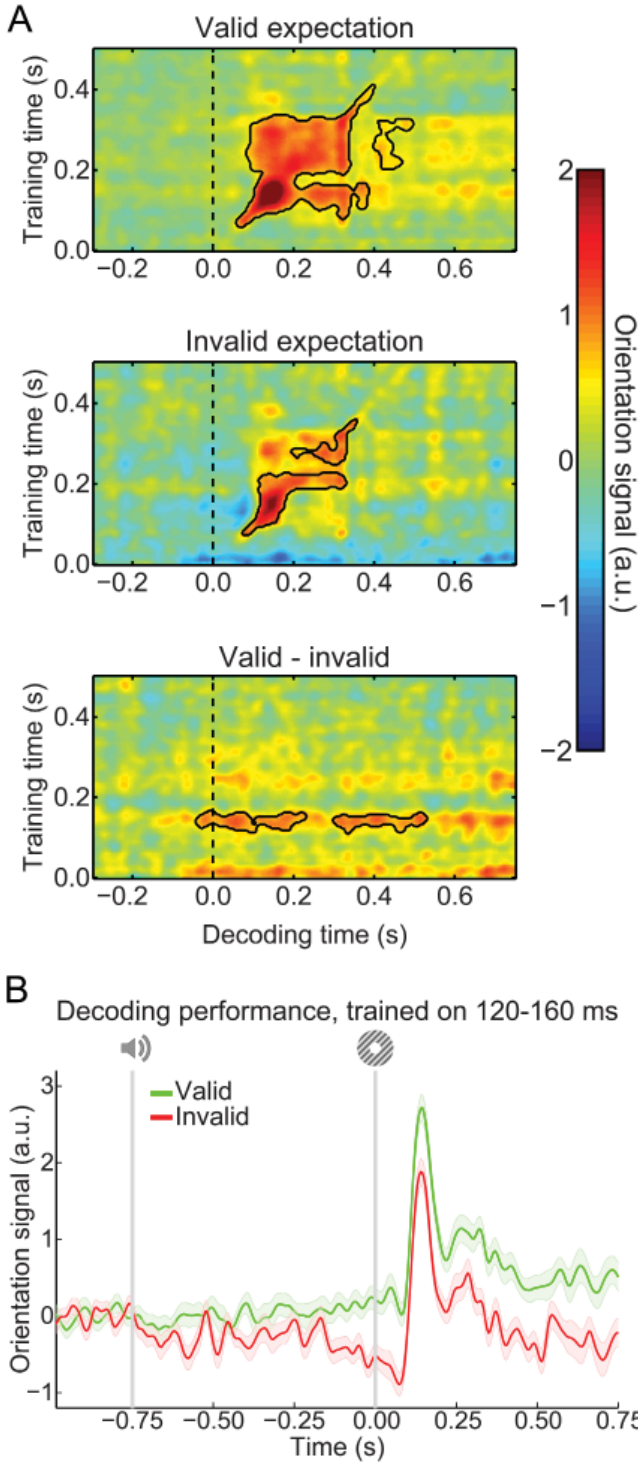


Figure 3. Expectation induces stimulus templates. **(A)** Temporal generalisation matrices of orientation decoding during the main experiment. Decoders were trained on the grating localiser (training time on the y-axis) and tested on the main experiment (time on the x-axis; dashed vertical line indicates $t = 0$ s, onset of the first grating). Decoding shown separately for gratings preceded by a valid expectation (top row), invalid expectation (middle row), and the subtraction of the two conditions (i.e., the expectation cue effect, bottom row). Solid black lines indicate significant clusters ($p < 0.05$). **(B)** Orientation decoding during the main task, averaged over training time 120 – 160 ms post-stimulus during the grating localiser. That is, a horizontal slice through the temporal generalisation matrices above at the training time for which we see a significant cluster of expected orientation decoding, for visualisation. Shaded regions indicate SEM.

evoked orientation-specific signals that were similar to sensory signals evoked by the corresponding actual grating stimuli (Fig. S1A).

In sum, expectations induced pre-stimulus sensory templates that influenced post-stimulus representations as well; invalidly expected gratings had to ‘overcome’ a pre-stimulus activation of the opposite orientation, while validly expected gratings were facilitated by a compatible pre-stimulus activation (Fig. S1B). The post-stimulus carryover of these expectation signals lasted throughout the trial (Fig. S1C).

As in previous studies using a similar paradigm (Kok et al., 2012; Kok, van Lieshout et al., 2016), there was no interaction between the effects of the expectation cue and the task (orientation vs. contrast discrimination) participants performed (no clusters with $p < 0.05$; Fig. S2A). In other words, expectations evoked pre-stimulus orientation signals to a similar degree in both tasks (Fig. S2B). This suggests that influences of expectation on neural representations are relatively independent of the task-relevance of the expected feature, in line with our previous fMRI study (Kok et al., 2012). Note though that, unlike in that study, there was no significant modulation of the orientation signal by task-relevance (no clusters with $p < 0.05$, Fig. S2A). The reason for this lack of difference is unclear, although it should be noted that there was a trend towards participants having higher accuracy and faster reaction times (see above) on the contrast task than on the orientation task. This may suggest the two tasks were not optimally balanced in terms of difficulty, precluding a proper comparison of the effect of task set in the current study.

In our previous fMRI study, we found a relationship between the effects of expectation on neural stimulus representations and performance on the orientation discrimination task. Specifically, participants for whom valid expectations led to the largest improvement in neural stimulus representations, also showed the strongest benefit of valid expectations on behavioural performance during the orientation discrimination task (Kok et al., 2012). This relationship was absent for the contrast discrimination task, when grating orientation was task-irrelevant. The current study allowed us to test for a similar relationship, with an important extension: here, we could test whether neural *pre-stimulus* expectation signals are related to behavioural performance improvements. We quantified the decoding of the expected orientation just before grating presentation (-50 to 0 ms, training window 120 to 160 ms) and correlated this with the difference in task accuracy for valid and invalid expectation trials, across participants. This analysis revealed that participants with a stronger pre-stimulus reflection of the expected orientation in their neural signal also had a greater benefit from valid expectations on performance on the orientation task ($r = 0.44$, $p = 0.035$; Fig. 4, left panel). No such relationship was found for the contrast task, where the orientation of the gratings was not task-relevant ($r = -0.13$, $p = 0.55$; Fig. 4, right panel). This is exactly the pattern of results we found in our previous fMRI study, but with the important extension that it is the *pre-stimulus* expectation effect that is correlated with behavioural performance, whereas the previous study did not have the temporal resolution to distinguish pre- from post-stimulus signals.

In the current study, neural orientation signals were probed by applying a forward model that takes the noise covariance between MEG sensors into account (see *Supporting Materials and Methods* for details). This model was superior to a forward model that did not correct for the noise covariance (Fig. S3), suggesting that feature covariance is an important factor to take into account when applying multivariate methods to MEG data. Corroborating this notion, a two-class decoder that corrected for noise covariance (Mostert et al., 2015) was able to reproduce our effects of interest (Fig. S4), demonstrating that the expectation effects do not depend on a specific analysis technique, as long as the covariance between MEG sensors is taken into account.

Finally, there was no difference in the overall amplitude of the neural response evoked between validly and invalidly expected gratings (no clusters with $p < 0.4$, Fig. S5).

Discussion

Here, we show that expectations can induce sensory templates of the expected stimulus already before the stimulus appears. These results extend previous fMRI studies demonstrating stimulus-specific patterns of activation in sensory cortex induced by expectations, but which could not resolve whether these templates indeed reflected pre-stimulus expectations, or instead stimulus specific error signals induced by the unexpected omission of a stimulus (Kok et al., 2014; Hindy et al., 2016). Furthermore, the strength of these pre-stimulus expectation signals correlated with the behavioural benefit of a valid expectation, when the expected feature (i.e., orientation) was task-relevant (Kok et al., 2012). These results suggest that valid expectations facilitate perception by allowing

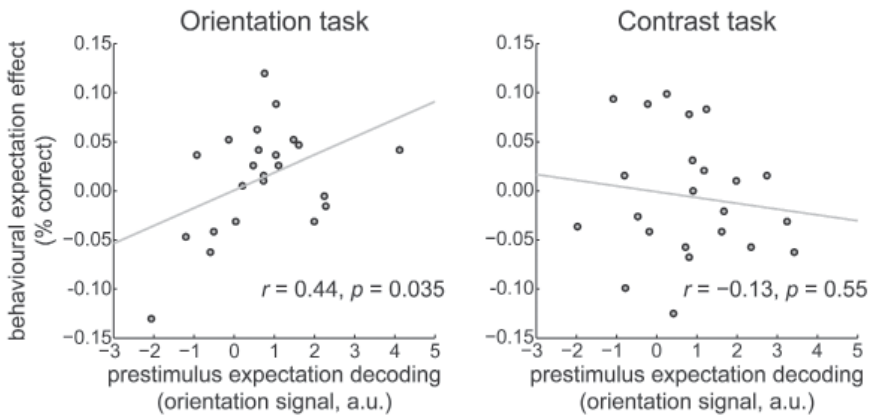


Figure 4. Correlation between neural expectation signals and behavioural improvement by expectation. Neural pre-stimulus expectation decoding (on the x axis) correlated with behavioural improvement induced by valid expectations (on the y axis) during the orientation discrimination task (left panel). This correlation was absent during the contrast discrimination task (right panel).

sensory cortex to prepare for upcoming sensory signals. As in a previous fMRI study using a very similar experimental paradigm (Kok et al., 2012), the neural effects of orientation expectations reported here were independent of the task-relevance of the orientation of the gratings, suggesting that the generation of expectation templates may be an automatic phenomenon.

The fact that expectation signals were revealed by a decoder trained on physically presented (but task-irrelevant) gratings suggests that these expectation signals resemble activity patterns induced by actual stimuli. The expectation signal remained present throughout the trial, extending into the post-stimulus period, suggesting the tonic activation of a stimulus template. These results are in line with a recent monkey electrophysiology study (Bell et al., 2016), which showed that neurons in the face patch of IT cortex encode the prior expectation of a face appearing, both prior to and following actual stimulus presentation. When the subsequently presented stimulus is noisy or ambiguous, such a pre-stimulus template could conceivably bias perception towards the expected stimulus (Chalk et al., 2010; Kok et al., 2013; Pajani et al., 2015; St. John-Saaltink et al., 2016).

What is the source of these cue-induced expectation signals? One candidate region is the hippocampus, which is known to be involved in encoding associations between previously unrelated, discontinuous stimuli (Wallenstein et al., 1998), such as the auditory tones and visual gratings used in the present study. Furthermore, fMRI studies have revealed predictive signals in the hippocampus (Hindy et al., 2016; Schapiro et al., 2012; Davachi and DuBrow, 2015), and Reddy et al. (2015) reported anticipatory firing to expected stimuli in the medial temporal lobe, including the hippocampus. One intriguing possibility is that predictive signals from the hippocampus are fed back to sensory cortex (Hindy et al., 2016; Lavenex and Amaral, 2000; Bosch et al., 2014).

Previous studies have suggested, both on theoretical (Bastos et al., 2012) and empirical (Bastos et al., 2015; Bauer et al., 2014) grounds, that top-down (prediction) and bottom-up (stimulus-driven, or prediction error) signals are subserved by distinct frequency bands. Therefore, one highly interesting direction for future research would be to determine whether the expectation templates revealed here are specifically manifested in certain frequency bands (i.e., the alpha or beta band).

In addition to expectation, several other cognitive phenomena have been shown to induce stimulus templates in sensory cortex, such as preparatory attention (Myers et al., 2015; Stokes et al., 2009a), mental imagery (Stokes et al., 2009b; Lee et al., 2012; Albers et al., 2013), and working memory (Harrison and Tong, 2009; Serences et al., 2009). In fact, explicit task preparation can also induce pre-stimulus sensory templates that last into the post-stimulus period (Myers et al., 2015). Note that in the current study the task did not require explicit use of the expectation cues, the task response was in fact orthogonal to the expectation. Furthermore, there was no difference in the expectation signal between runs in which grating orientation was task-relevant (orientation discrimination task) and

when it was irrelevant (contrast discrimination task), suggestion expectation may be a relatively automatic phenomenon (Kok et al., 2012; Den Ouden et al., 2009). In fact, neural modulations by expectation have even been observed during states of inattention (Näätänen, 1990), sleep (Nakano et al., 2008) and in patients experiencing disorders of consciousness (Bekinschtein et al., 2009). One important question for future research will be to establish whether the same neural mechanism underlies the different cognitive phenomena that are capable of inducing stimulus templates in sensory cortex, or whether different top-down mechanisms are at work. Indeed, it has been suggested that expectation and attention, or task preparation, may have different underlying neural mechanisms (Kok, van Lieshout et al., 2016; Summerfield and Egner, 2009, 2016). For instance, predictive coding theories suggest that attention may modulate sensory signals in the superficial layers of sensory cortex, while predictions modulate the response in deep layers (Friston, 2005; Kok, Bains et al., 2016).

One may wonder why the current study does not report a modulation of the overall neural response by expectation, while previous studies have found an increased neural response to unexpected stimuli (Den Ouden et al., 2009; Summerfield et al., 2008; Alink et al., 2010; Meyer and Olson, 2011; Todorovic et al., 2011; Wacongne et al., 2011), including some using an almost identical paradigm as the current study (Kok et al., 2011; Kok, van Lieshout et al., 2016). Of course, the current study reports a null effect, from which it is hard to draw firm conclusions. However, it is possible that the type of measurement of neural activity plays a role in the absence of the effect. Most previous studies reporting expectation suppression in visual cortex used fMRI, while the current study used MEG. It is possible that the BOLD signal, a mass-action signal that integrates synaptic and neural activity, as well as integrating over time, is sensitive to certain neural effects that MEG, which is predominantly sensitive to synchronised activity in pyramidal neurons oriented perpendicular to the cortical surface, is not. It is even possible that within MEG, different types of sensors (i.e. magnetometers, planar and axial gradiometers) differ in their sensitivity to expectation suppression (Cashdollar et al., 2016).

Recent theories of sensory processing state that perception reflects the integration of bottom-up inputs and top-down expectations, but ideas diverge on whether the brain continuously generates stimulus templates in sensory cortex to pre-empt expected inputs (Bell et al., 2016; Pajani et al., 2015; Berkes et al., 2011; Fiser et al., 2016), or rather engages in perceptual inference only after receiving sensory inputs (Rao and Ballard, 1999; Bar et al., 2006). Our results are in line with the brain being proactive, constantly forming predictions about future sensory inputs. These findings bring us closer to uncovering the neural mechanisms by which we integrate prior knowledge with sensory inputs to optimise perception.

Materials and Methods

Participants

Twenty-three (15 female, age 26 ± 9 , mean \pm SD) healthy individuals participated in the MEG experiment. All participants were right-handed and had normal or corrected-to-normal vision. The study was approved by the local ethics committee (CMO Arnhem-Nijmegen, The Netherlands) under the general ethics approval (“Imaging Human Cognition”, CMO 2014/288), and the experiment was conducted in accordance with these guidelines. All participants gave written informed consent according to the declaration of Helsinki.

Experimental design

Each trial consisted of an auditory cue, followed by two consecutive grating stimuli (750 ms SOA between auditory and first visual stimulus) (Fig. 1A). The two grating stimuli were presented for 250 ms each, separated by a blank screen (500 ms). A central fixation bull’s eye (0.7°) was presented throughout the trial, as well as during the intertrial interval (ITI, 2250 ms). The auditory cue consisted of either a low- (500 Hz) or high-frequency (1000 Hz) tone, which predicted the orientation of the first grating stimulus (45° or 135°) with 75% validity (Fig. 1B). In the other 25% of trials, the first grating had the orthogonal orientation. Thus, the first grating had an orientation of either exactly 45° or 135° , and a luminance contrast of 80%. The second grating differed slightly from the first in terms of both orientation and contrast (see below), as well as being in antiphase to the first grating (which had a random spatial phase). The contingencies between the auditory cues and grating orientations were flipped halfway through the experiment (i.e., after four runs), and the order was counterbalanced over subjects.

In separate runs (64 trials each, ~ 4.5 minutes), subjects performed either an orientation or a contrast discrimination task on the two gratings. When performing the orientation task, subjects had to judge whether the second grating was rotated clockwise or anticlockwise with respect to the first grating. In the contrast task, a judgment had to be made on whether the second grating had lower or higher contrast than the first one. These tasks were explicitly designed to avoid a direct relationship between the perceptual expectation and the task response. Furthermore, as in a previous fMRI study (Kok et al., 2012), these two different tasks were designed to manipulate the task-relevance of the grating orientations, to investigate whether the effects of orientation expectations depend on the task-relevance of the expected feature.

Interleaved with the main task runs, subjects performed eight runs of a grating localiser task (Fig. 1C). Each run (~ 2 min) consisted of 80 grating presentations (ITI uniformly jittered between 1000 and 1200 ms). The grating annuli were identical to those presented during the main task (80% contrast, 250 ms duration, 1.0 cycles/ $^\circ$, random spatial phase).

Each grating had one of eight orientations (spanning the 180° space, starting at 0° , in steps of 22.5°), each of which was presented ten times per run in pseudorandom order. A black fixation bull's eye (4 cd/m², 0.7° diameter, identical to the one presented during the main task runs) was presented throughout the run. On 10% of trials (counterbalanced across orientations), the black fixation point in the centre of the bull's eye (0.2° , 4 cd/m²) briefly turned gray (324 cd/m²) during the first 50 ms of grating presentation. Participants task was to press a button (response deadline: 500 ms) when they perceived this fixation flicker. This simple task was meant to ensure central fixation, while rendering the gratings task-irrelevant. Trials containing fixation flickers were excluded from further analyses.

Orientation decoding analysis

To probe sensory representations in the visual cortex, we used a forward modelling approach to reconstruct the orientation of the grating stimuli from the MEG signal (Meyers et al., 2015; Brouwer et al., 2009, 2011; Garcia et al., 2013). This method has been shown to be highly successful at reconstructing circular stimulus features, such as colour (Brouwer and Heeger, 2009), orientation (Meyers et al., 2015; Brouwer and Heeger, 2011; Garcia et al., 2013), and motion direction (Kok et al., 2013), from neural signals. Neural representations in MEG signals have also been successfully investigated using binomial classifiers (Cichy et al., 2015), however, when it comes to a continuous stimulus feature such as orientation, forward model reconstructions provide a richer decoding signal than binomial classifier accuracy (Ester et al., 2015). We made certain changes to the forward model proposed by Brouwer and Heeger (2009; most notably, taking the noise covariance into account; (see *Supporting Materials and Methods* for details) in order to optimise it for MEG data, given the high correlations between neighbouring sensors, based on Mostert et al. (2015). In sum, this previously published and theoretically motivated decoding model was optimally suited for recovering a continuous feature from MEG data. For our main analyses, the forward model was trained on the data from the localiser runs, in which the gratings were task-irrelevant, and then applied to the main task data, in order to uncover sensory templates induced by pre-stimulus expectations (see *Supporting Materials and Methods* for details). Our effects of interest (see Fig. 3) were reproduced using a two-class decoder (Fig. S4).

Supporting Materials and Methods

Participants

Twenty-three (15 female, age 26 ± 9 , mean \pm SD) healthy individuals participated in the experiment. All participants were right-handed and had normal or corrected-to-normal vision. The study was approved by the local ethics committee (CMO Arnhem-Nijmegen, The Netherlands) under the general ethics approval ("Imaging Human Cognition", CMO 2014/288), and the experiment was conducted in accordance with these guidelines. All participants gave written informed consent according to the declaration of Helsinki.

Stimuli

Grayscale luminance-defined sinusoidal grating stimuli (spatial frequency: 1.0 cycles/°) were generated using MATLAB (MathWorks, Natick, MA) in conjunction with the Psychophysics Toolbox (60). Gratings were displayed in an annulus (outer diameter: 15° of visual angle, inner diameter: 1°), surrounding a black fixation bull's eye (4 cd/m²), on a gray (580 cd/m²) background. The visual stimuli were presented with an LCD projector (1024 × 768 resolution, 60 Hz refresh rate) positioned outside the magnetically shielded room, and projected on a translucent screen via two front-silvered mirrors. The projector lag was measured at 36 ms, which was corrected for by shifting the time axis of the data accordingly. The auditory cue consisted of a pure tone (500 or 1000 Hz, 250 ms duration, including 10 ms on and off-ramp time), presented over MEG-compatible earphones.

Experimental design

Each trial consisted of an auditory cue, followed by two consecutive grating stimuli (750 ms SOA between auditory and first visual stimulus) (Fig. 1A). The two grating stimuli were presented for 250 ms each, separated by a blank screen (500 ms). A central fixation bull's eye (0.7°) was presented throughout the trial, as well as during the intertrial interval (ITI, 2250 ms). The auditory cue consisted of either a low- (500 Hz) or high-frequency (1000 Hz) tone, which predicted the orientation of the first grating stimulus (45° or 135°) with 75% validity (Fig. 1B). In the other 25% of trials, the first grating had the orthogonal orientation. Thus, the first grating had an orientation of either exactly 45° or 135°, and a luminance contrast of 80%. The second grating differed slightly from the first in terms of both orientation and contrast (see below), as well as being in antiphase to the first grating (which had a random spatial phase). The contingencies between the auditory cues and grating orientations were flipped halfway through the experiment (i.e., after four runs), and the order was counterbalanced over subjects.

In separate runs (64 trials each, ~4.5 minutes), subjects performed either an orientation or a contrast discrimination task on the two gratings. When performing the orientation task, subjects had to judge whether the second grating was rotated clockwise or anticlockwise with respect to the first grating. In the contrast task, a judgment had to be made on whether the second grating had lower or higher contrast than the first one. These tasks were explicitly designed to avoid a direct relationship between the perceptual expectation and the task response. Furthermore, as in a previous fMRI study (Kok et al., 2012), these two different tasks were designed to manipulate the task-relevance of the grating orientations, to investigate whether the effects of orientation expectations depend on the task-relevance of the expected feature. Subjects indicated their response (response deadline: 750 ms after offset of the second grating) using an MEG-compatible button box. The orientation and contrast differences between the two gratings were determined by an adaptive staircase procedure (61), being updated after each trial. This was done to yield comparable task difficulty and performance (~ 75% correct) for the different tasks. In the current study,

unlike in our previous fMRI study (Kok et al., 2012), we did not run separate staircasing for the valid and invalid expectation conditions. In that study, we were concerned that once the first grating on a trial violated the expectation, this might lead to increased difficulty in comparing this unexpected grating to the second grating. This hypothetical chain of events would be triggered only after the first grating had been processed, however, in an fMRI study one cannot separate such a relatively late effect from any early expectation effects. In the current study on the other hand we could, since we used MEG. Any pre-stimulus expectation effects, which were the target of the current study, could not possibly be affected by any post-stimulus difficulty effects. Therefore, for simplicity, we ran a single staircase for all trials, one per task. Staircase thresholds obtained during one task were used to set the stimulus differences during the other task, in order to make the stimuli as similar as possible in both contexts.

All subjects completed eight runs (four of each task, alternating every two runs, order was counterbalanced over subjects) of the experiment, yielding a total of 512 trials. The staircases were kept running throughout the experiment. Before the first run, as well as in between runs four and five, when the contingencies between cue and stimuli were flipped, subjects performed a short practice run containing 32 trials of both tasks (~4.5 minutes).

Interleaved with the main task runs, subjects performed eight runs of a grating localiser task (Fig. 1C). Each run (~2 min) consisted of 80 grating presentations (ITI uniformly jittered between 1000 and 1200 ms). The grating annuli were identical to those presented during the main task (80% contrast, 250 ms duration, 1.0 cycles/°, random spatial phase). Each grating had one of eight orientations (spanning the 180° space, starting at 0°, in steps of 22.5°), each of which was presented ten times per run in pseudorandom order. A black fixation bull's eye (4 cd/m², 0.7° diameter, identical to the one presented during the main task runs) was presented throughout the run. On 10% of trials (counterbalanced across orientations), the black fixation point in the centre of the bull's eye (0.2°, 4 cd/m²) briefly turned gray (324 cd/m²) during the first 50 ms of grating presentation. Participants task was to press a button (response deadline: 500 ms) when they perceived this fixation flicker. This simple task was meant to ensure central fixation, while rendering the gratings task-irrelevant. Trials containing fixation flickers were excluded from further analyses.

Finally, participants were exposed to a tone localiser (~1.5 min), presented at the start, end, and halfway through the MEG session. These runs consisted of 81 presentations of the two tones used in the main experiment. Data from these runs were not analysed further.

Prior to the MEG session (1–3 days), all participants completed a behavioural session. The aim of this session was to familiarise participants with the tasks and to initialise the staircase values for both the orientation and the contrast discrimination task (see above). The behavioural session consisted of written instructions and 32 practice trials of each task, followed by four runs (~4.5 min each) of the main experiment (each task twice, alternating between runs, cue contingencies switching between the second and third run).

Finally, participants were exposed to one run each of the grating and tone localiser, to familiarise them with the procedure.

MEG recording and preprocessing

Whole-head neural recordings were obtained using a 275-channel MEG system with axial gradiometers (CTF Systems, Coquitlam, BC, Canada) located in a magnetically shielded room. Throughout the experiment, head position was monitored online, and corrected if necessary, using three fiducial coils that were placed on the nasion and on earplugs in both ears (Stolk et al., 2013). If subjects had moved their head more than 5 mm from the starting position they were repositioned during block breaks. Furthermore, both horizontal and vertical electrooculograms (EOGs), as well as an electrocardiogram (ECG) were recorded to facilitate removal of eye- and heart-related artifacts. The ground electrode was placed at the left mastoid. All signals were sampled at a rate of 1200 Hz.

The data were preprocessed offline using FieldTrip (Oostenveld et al., 2011; www.fieldtriptoolbox.org). In order to identify artifacts, the variance (collapsed over channels and time) was calculated for each trial. Trials with large variances were subsequently selected for manual inspection and removed if they contained excessive and irregular artifacts. Independent component analysis was subsequently used to remove regular artifacts, such as heartbeats and eye blinks. Specifically, for each subject, the independent components were correlated to both EOGs and the ECG to identify potentially contaminating components, and these were subsequently inspected manually before removal. For the main analyses, data were low-pass filtered using a two-pass Butterworth filter with a filter order of 6 and a frequency cutoff of 40 Hz. To rule out that the temporal smoothing caused by low-pass filtering may have artificially decreased the onset latency of neural signals, we repeated the decoding analyses (see below) on data that were not low-pass filtered (Fig. S6). Here, only notch filters were applied at 50, 100 and 150 Hz to remove line noise and its harmonics. The absence of any low-pass filtering or smoothing in this analysis precluded the possibility that any pre-stimulus effects could be due to backwards smoothing of stimulus-driven effects. No detrending was applied for any analysis. Finally, main task data were baseline corrected on the interval of -250 to 0 ms relative to auditory cue onset, and grating localiser data were baseline corrected on the interval of -200 to 0 ms relative to visual grating onset.

Event-related field analysis

Event-related fields (ERFs) were calculated per participant, and subjected to a planar gradient transformation (Bastiaansen and Knösche, 2000) before averaging across participants. The planar transformation simplifies the interpretation of the sensor-level data because it typically places the maximal signal above the source. To avoid differences in the amount of noise when comparing conditions with different numbers of trials, we matched the trial count by randomly selecting a subsample of trials from the conditions with more trials (i.e., valid expectations).

Orientation decoding analysis

To probe sensory representations in the visual cortex, we used a forward modelling approach to reconstruct the orientation of the grating stimuli from the MEG signal (Myers et al., 2015; Brouwer and Heeger, 2009, 2011; Garcia et al., 2013). This method has been shown to be highly successful at reconstructing circular stimulus features, such as colour (Brouwer and Heeger, 2009), orientation (Myers et al., 2015; Brouwer and Heeger, 2011; Garcia et al., 2013), and motion direction (Kok et al., 2013), from neural signals. Neural representations in MEG signals have also been successfully investigated using binomial classifiers (Cichy et al., 2015), however, when it comes to a continuous stimulus feature such as orientation, forward model reconstructions provide a richer decoding signal than binomial classifier accuracy (Ester et al., 2015). We made certain changes to the forward model proposed by Brouwer and Heeger (2009; most notably, taking the noise covariance into account; see below for details) in order to optimise it for MEG data, given the high correlations between neighbouring sensors, based on Mostert et al. (2015). In sum, this previously published and theoretically motivated decoding model was optimally suited for recovering a continuous feature from MEG data.

The forward modelling approach was two-fold. First, a theoretical forward model was postulated that described the measured activity in the MEG sensors, given the orientation of the presented grating. Second, this forward model was used to obtain an inverse model that specified the transformation from MEG sensor space to orientation space. The forward and inverse models were estimated on the basis of the grating localiser data. The inverse model was then applied to the data from the main experiment, in order to generalise from sensory signals evoked by task-irrelevant gratings to the gratings and expectation signals evoked in the main task. To test the performance of the model we also applied it to the localiser data itself, using a cross-validation approach in which in each iteration one trial of each orientation was used at the test set, and the remaining data were used as the training set.

The forward model was based on work by Brouwer and Heeger (2009, 2011) and involved 32 hypothetical channels, each with an idealised orientation tuning curve. Each channel consisted of a half-wave-rectified sinusoid raised to the fifth power, and the 32 channels were spaced evenly within the 180° orientation space, such that a tuning curve with any possible orientation preference could be expressed exactly as a weighted sum of the channels. Arranging the hypothesised channel activities for each trial along the columns of a matrix \mathbf{C} (32 channels \times n trials), the observed data could be described by the following linear model:

where \mathbf{B} are the (m sensors \times n trials) MEG data, \mathbf{W} is a weight matrix (m sensors \times 32 channels) that specifies how channel activity is transformed into sensory activity, and \mathbf{N} are the residuals (i.e., noise).

$$\mathbf{B} = \mathbf{WC} + \mathbf{N}$$

In order to obtain the inverse model, we estimated an array of spatial filters that, when applied to the data, aimed to reconstruct the underlying channel activities as accurately as possible. In doing so, we extended Brouwer and Heeger's (2009) approach in three respects. First, since the MEG signal in (nearby) sensors is correlated, we took into account the correlational structure of the noise. Second, we estimated a spatial filter for each orientation channel independently. As a result, the number of channels used in our model was not constrained, whereas the maximum number of channels would otherwise be dependent on the number of presented orientations. In practice, this resulted in smoothing in orientation space, because the channels were not truly independent. Third, each filter was normalised such that the magnitude of its output matched the magnitude of the underlying channel activity it was designed to recover. Prior to estimating the inverse model, \mathbf{B} and \mathbf{C} were demeaned such that their average over trials equalled zero, for each sensor and channel, respectively.

As stated above, the inverse model was estimated on the basis of the grating localiser data. On each localiser trial, one of eight orientations was presented (see above), and the hypothetical responses of each of the channels could thus be calculated for each trial, resulting in the response row vector $\mathbf{c}_{train,i}$, of length n_{train} trials, for each channel i . The weights on the sensors \mathbf{w}_i could now be obtained through least squares estimation, for each channel:

$$\mathbf{w}_i = \mathbf{B}_{train} \mathbf{c}_{train,i}^T \left(\mathbf{c}_{train,i} \mathbf{c}_{train,i}^T \right)^{-1}$$

where \mathbf{B}_{train} are the (m sensors \times n_{train} trials) localiser MEG data. Subsequently, the optimal spatial filter \mathbf{v}_i to recover the activity of the i -th channel was obtained as follows (Mostert et al., 2015):

$$\mathbf{v}_i = \frac{\tilde{\Sigma}_i^{-1} \mathbf{w}_i}{\mathbf{w}_i^T \tilde{\Sigma}_i^{-1} \mathbf{w}_i}$$

where $\tilde{\Sigma}_i$ is the regularised covariance matrix for channel i . Incorporating the noise covariance in the filter estimation leads to the suppression of noise that arises from correlations between sensors. The noise covariance was estimated as follows:

$$\hat{\Sigma}_i = \frac{1}{n_{train} - 1} \mathbf{e}_i \mathbf{e}_i^T \quad \mathbf{e}_i = \mathbf{B}_{train} - \mathbf{w}_i \mathbf{c}_{train,i}$$

where n_{train} is the number of training trials. For optimal noise suppression, we improved this estimation by means of regularization by shrinkage, using the analytically determined optimal shrinkage parameter (for details, see Blankertz et al., 2011), yielding the regularised covariance matrix $\tilde{\Sigma}_i$.

Such a spatial filter was estimated for each hypothetical channel, yielding an m sensors \times 32 channel filter matrix \mathbf{V} . Given that we performed our decoding analysis in a time-resolved manner, \mathbf{V} was estimated at each time point of the training data, in steps of 5 ms, resulting in array of filter matrices, or decoders. To improve the signal-to-noise ratio, the data were first averaged within a window of 29.2 ms centred on the time point of interest. The window length of 29.2 ms was based on an a priori chosen length of 30 ms, but minus one sample such that the window contained an odd number of samples for symmetric centring (Mostert et al., 2015). These filter matrices could now be applied to estimate the orientation channel responses in independent data – in this case, the trials from the main experiment:

$$\mathbf{C}_{test} = \mathbf{V}^T \mathbf{B}_{test}$$

where \mathbf{B}_{test} are the (m sensors $\times n_{test}$ trials) main experiment data. These channel responses were estimated at each time point of the test data, in steps of 5 ms, with the data being averaged within a window of 29.2 ms at each step. This procedure resulted in a four-dimensional (training time \times testing time \times 32 channel $\times n_{test}$) matrix of estimated channel responses for each trial in the main experiment. Each trials' channel responses were shifted such that the channel with its hypothetical peak response at the orientation presented on that trial (i.e. 45° or 135°) ended up in the position of the 0° channel, before averaging over trials within each condition (i.e., valid vs. invalid expectation). Thus, the presented orientation was defined as 0°, by convention. Note that for 3D surface plots that show the evolution of channel responses over time (e.g., Fig. 2B), the response of the 90° channel (i.e., orthogonal to the presented orientation) was used as a baseline, to avoid negative numbers for visualisation purposes.

To quantify decoding performance, the channel responses for a given condition were converted into polar form and projected onto a vector with angle 0° (the presented orientation, see above).

$$r = |z| \cos(\arg(z)) \quad z = \mathbf{c} e^{2i\phi}$$

where \mathbf{c} is a vector of estimated channel responses, and ϕ is the vector of angles at which the channels peak (multiplied by 2 to project the 180° orientation space onto the full 360° space). The scalar projection r indicates the strength of the decoder signal for the orientation presented on screen. (Note that this approach is practically identical to subtracting the estimated response of the 90° channel from that of the 0° channel.) This quantification yielded (training time \times testing time) temporal generalisation matrices of orientation decoding performance.

In order to establish the importance of including the noise covariance term in the forward model, we compared decoding performance to a standard forward model that did not take noise covariance into account (Fig. S3). This analysis showed that the adapted forward model was far superior to the standard model. To ensure that our results were not dependent on one specific decoding implementation, we also reproduced our effects of

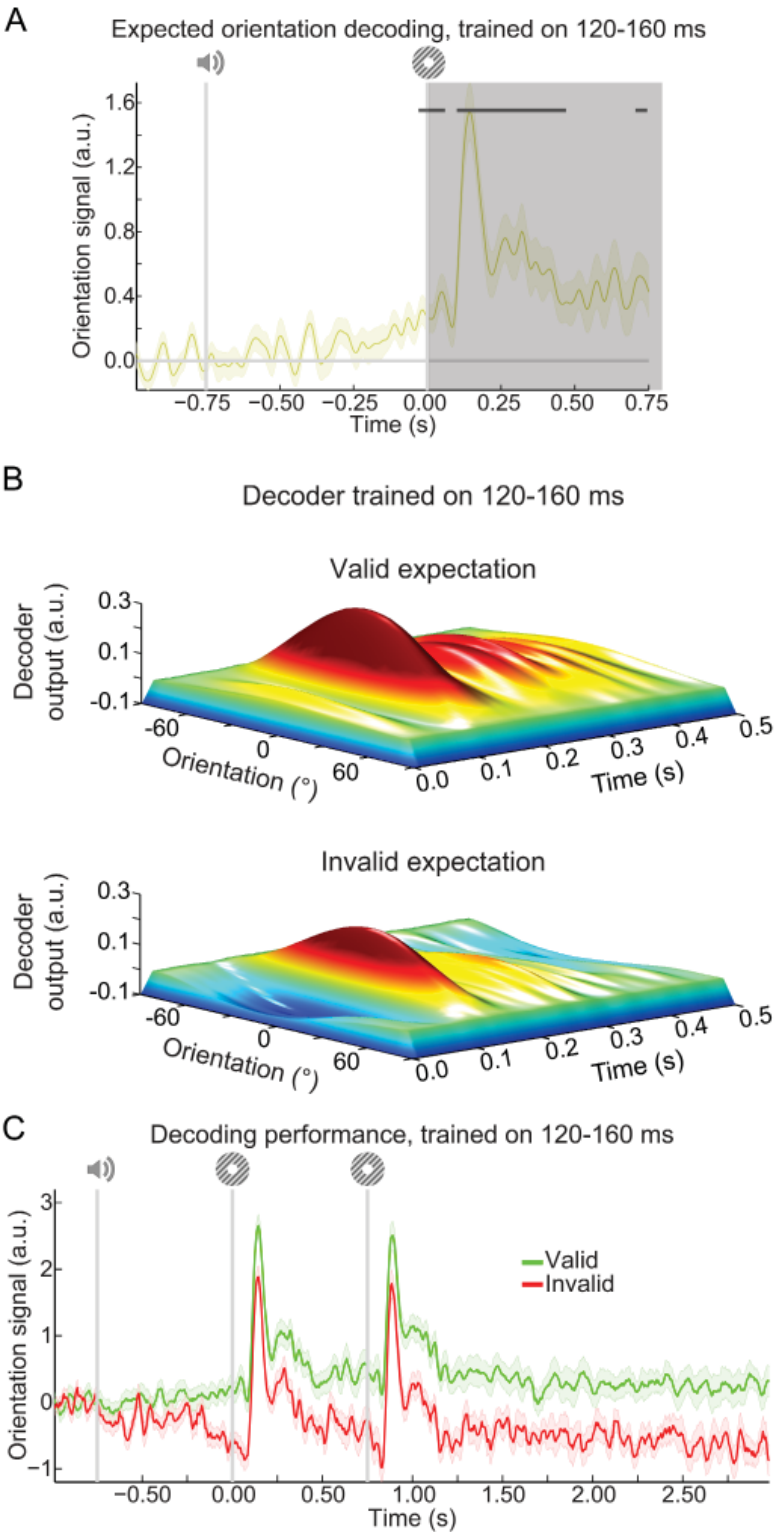
interest using a two-class decoder that takes noise covariance into account, equivalent to a linear discriminant analysis (Mostert et al., 2015). For details and equations underlying this model, see Mostert et al. (2015). Given that training the decoder only on 45° and 135° gratings would only allow us to use 25% of the localiser trials, we trained the decoder to distinguish right-tilted (22.5° , 45° , and 67.5°) from left-tilted (112.5° , 135° and 157.5°), using 75% of the localiser trials. This decoder was then applied to the main task, revealing virtually identical expectation effects as our adapted forward model (Fig. S4).

In order to isolate any orientation-specific neural signals evoked by the expectation cues, we applied the following subtraction logic. On valid expectation trials, the expected and presented orientations are identical, and thus the orientation signal induced by both the cue and stimulus be expected to be positive, by convention. On invalid expectation trials on the other hand, the expected and presented orientations are orthogonal, and thus the orientation signal induced by the stimulus would be positive and the signal induced by cue would be expected to be negative. Thus, subtracting the orientation decoding signal on invalid trials from that on valid trials would subtract out the stimulus-evoked signal while revealing any cue-induced orientation signal. Additionally, we investigated cue-induced orientation signals by simply aligning trials by the expected, rather than the presented orientation (Fig. S1A).

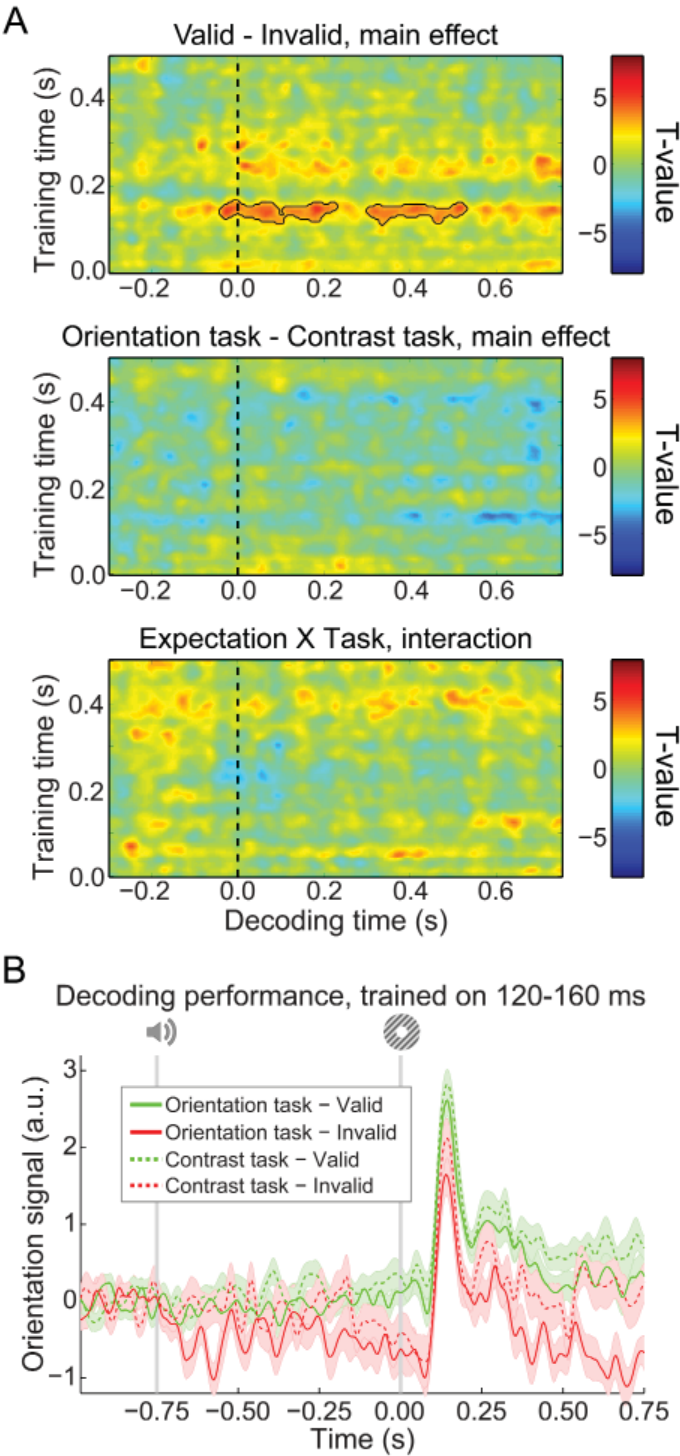
Statistical testing

Neural signals evoked by the different conditions were statistically tested using nonparametric cluster-based permutation tests (Maris and Oostenveld, 2007). For ERF analyses, we averaged over the spatial (sensor) dimension, on the basis of independent localisation of the 10 sensors that showed the strongest visual-evoked activity during the grating localiser between 50 and 150 ms post-stimulus. Therefore, our statistical analysis considered one-dimensional (temporal) clusters. For orientation decoding analyses, the data consisted of two-dimensional (training time \times testing time) decoding performance matrices, and the statistical analysis thus considered two-dimensional clusters. For both one- and two-dimensional data, univariate t -statistics were calculated for the entire matrix and neighbouring elements that passed a threshold value corresponding to a p -value of 0.01 (two-tailed) were collected into separate negative and positive clusters. Elements were considered neighbours if they were directly adjacent, either cardinally or diagonally. Cluster-level test statistics consisted of the sum of t -values within each cluster, and these were compared to a null distribution of test statistics created by drawing 10,000 random permutations of the observed data. A cluster was considered significant when its p -value was below 0.05 (two-tailed).

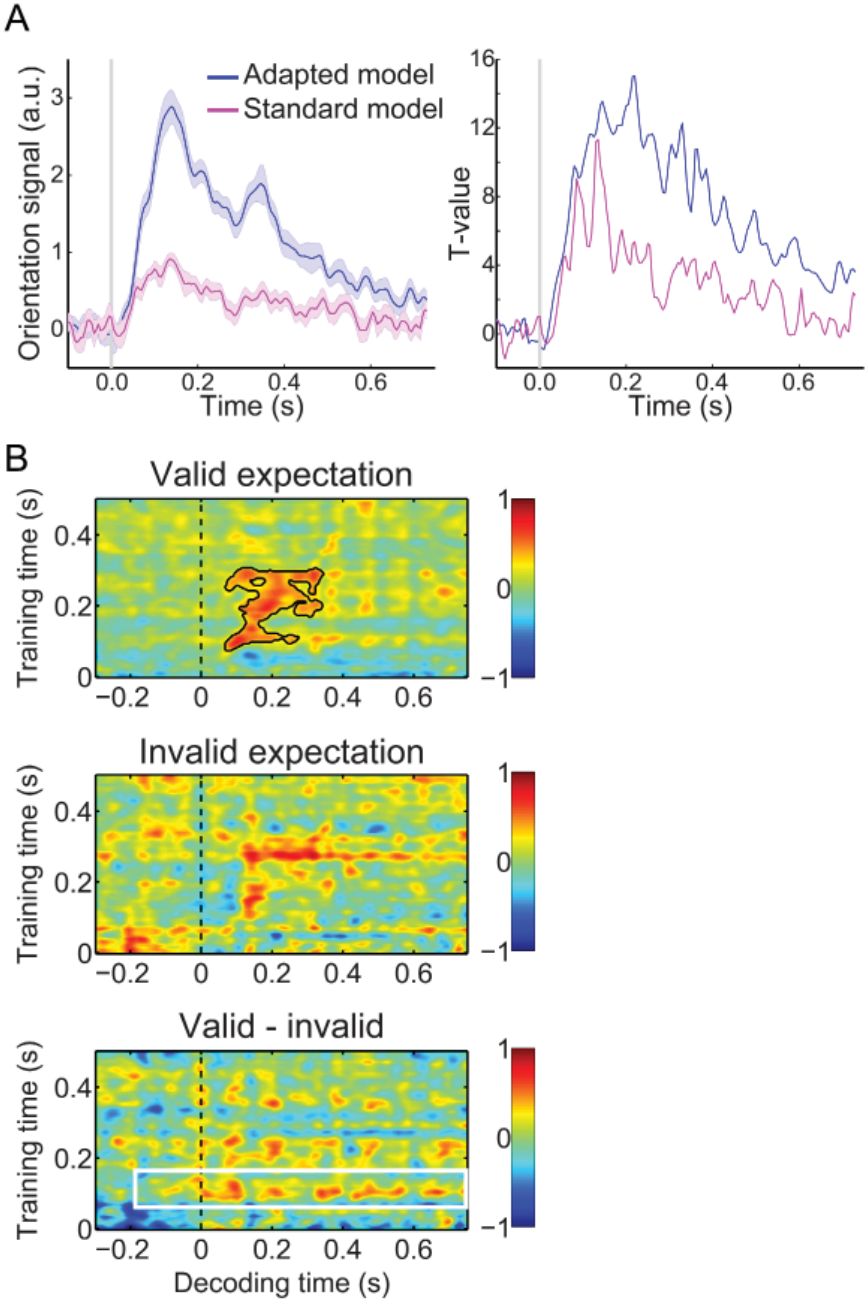
Supporting Figures



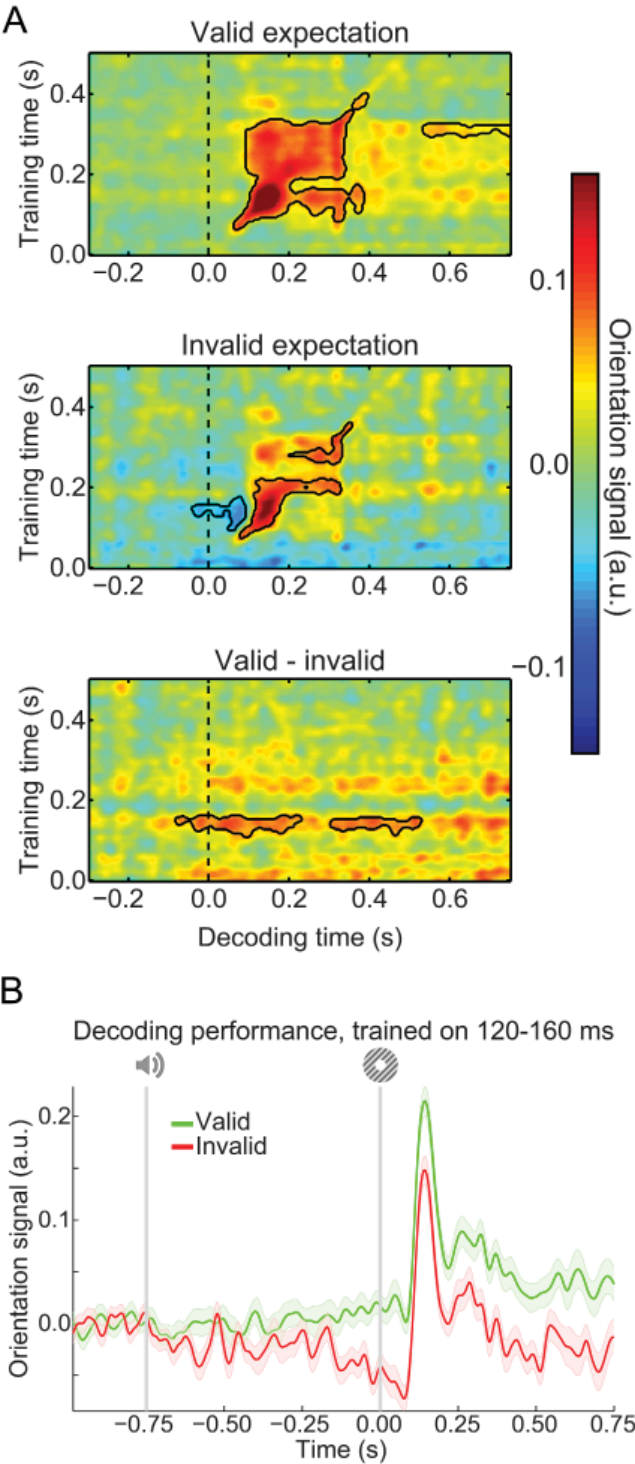
Supplementary Figure S1. Expectation affects orientation decoding throughout the trial. **(A)** Decoded orientation signals resulting from aligning trials to the expected, rather than the presented orientation. Note that post-stimulus ($t > 0$ ms, shaded region) decoding signals cannot be unambiguously assigned to the expectation cue, since on 75% of trials the presented orientation is the same as the expected orientation. It can be seen that there is already significant decoding of the cued orientation before $t = 0$ ms. Note that this pre-stimulus decoding signal is also significant when only time points before $t = 0$ ms are submitted to the cluster-based permutation test. Horizontal stripes indicate significant clusters. **(B)** Output of the 32 orientation channels over time, during the main task, separately for gratings preceded by a valid (top panel) and invalid (bottom panel) expectation cue. Here, decoders were trained on 120 – 160 ms post-stimulus during the grating localiser, similar to the data in Fig. 3B. That is, this figure depicts similar results as in Fig. 3B, but displaying the output of all 32 orientation channels, rather than collapsing the channel responses into a decoding performance score (cf. Fig. 2B). **(C)** Orientation decoding during the main task, averaged over training time 120 – 160 ms post-stimulus during the grating localiser. Same as in Fig. S6B, but with an extended x-axis, in order to shown the sustained nature of the expectation templates. Shaded regions indicate SEM.



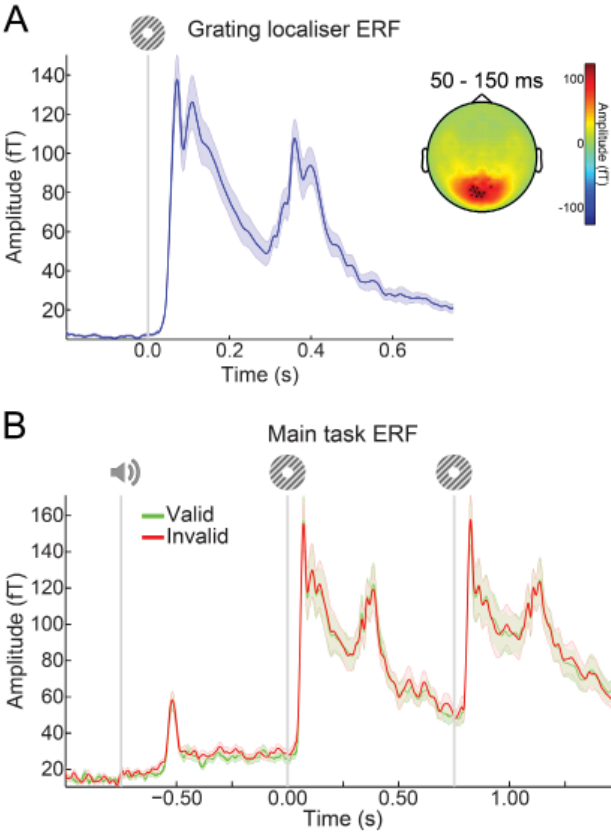
Supplementary Figure S2. Effects of expectation and task-relevance on neural orientation signals. **(A)** Temporal generalisation matrices of the effects of expectation (top), task-relevance (middle) and the interaction between the two factors (bottom panel). Note that the top panel is identical to the bottom panel of Fig. 3A. Solid black lines indicate significant clusters ($p < 0.05$). **(B)** Orientation decoding, separately for validly and invalidly cued gratings, split up for the two tasks, averaged over training time 120 – 160 ms post-stimulus during the grating localiser. Shaded regions indicate SEM.



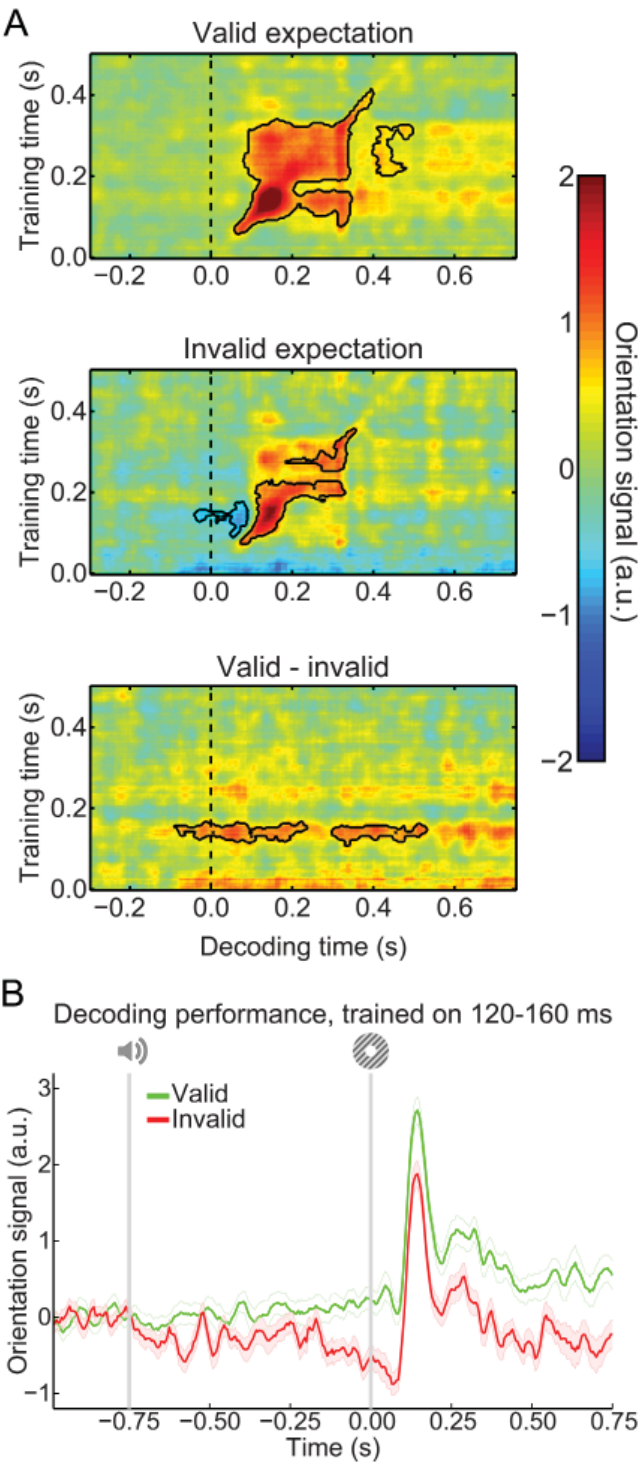
Supplementary Figure S3. Orientation decoding using a standard forward model. Unlike the main analyses, this model does not take the noise covariance between MEG sensors into account (see *Supporting Materials and Methods* for details). **(A)** Orientation decoding within the localiser, separately for our adapted forward model (in blue) and a standard forward model (in pink). Left panel shows mean decoding performance (shaded regions represent SEM), right panel shows T-values over participants. **(B)** Temporal generalisation matrices of orientation decoding during the main experiment. Same format as Fig. 3A, using a sub-optimal standard forward model rather than our adapted forward model. As expected, applying this less sensitive standard model to our main task data results in far less reliable decoding in the main task (note that the unexpected gratings are no longer significantly decoded at all, middle panel). Decoding of the expected orientation (bottom panel) is numerically still reflected by a horizontal stripe around training time 120-160ms (indicated by the white box), but this effect is not statistically significant. This is likely due to the fact that the standard model is far less sensitive to neural orientation signals, as illustrated by panel A. Solid black lines indicate significant clusters ($p < 0.05$).



Supplementary Figure S4. Expectation effects as revealed by a two-class decoder. Same format as Fig. 3, except that orientation signals were decoded using a two-class decoding method, rather than a forward model as in the main analyses (see *Supporting Materials and Methods*). **(A)** Temporal generalisation matrices of orientation decoding during the main experiment. Solid black lines indicate significant clusters ($p < 0.05$). **(B)** Orientation decoding during the main task, averaged over training time 120 – 160 ms post-stimulus during the grating localiser. Shaded regions indicate SEM.



Supplementary Figure S5. Event related fields. **(A)** Event-related fields during the grating localiser, in the 10 channels showing the largest response from 50 – 150 ms post-stimulus. **(B)** Event-related fields during the main task, in the same 10 channels as in panel A, selected on the basis of the grating localiser response. No significant differences between ERF for validly and invalidly predicted gratings. Shaded regions indicate SEM.



Supplementary Figure S6. Expectation effects in the absence of low-pass filtering. **(A)** Temporal generalisation matrices of orientation decoding during the main experiment, without low-pass filtering the data. Otherwise, identical to Fig. 3A. Solid black lines indicate significant clusters ($p < 0.05$). **(B)** Orientation decoding during the main task, averaged over training time 120 – 160 ms post-stimulus during the grating localiser. Shaded regions indicate SEM.



4

**Similar neural activity patterns evoked
by expected and unexpected object images**

Abstract

Numerous studies have reported that expected stimuli evoke a reduced neural response as compared to unexpected stimuli. This phenomenon, known as expectation suppression, has aided in the development of influential theories. The goal of the current study was to address a number of questions regarding expectation suppression that have so far remained open. We subjected human volunteers to a simple statistical learning paradigm in which they were presented with two images of objects in succession, while recording their brain activity using magnetoencephalography (MEG). The leading image could be either of three, and predicted which of two leading images would follow. This resulted in an expected, neutral and unexpected condition. Subjects performed an oddball task in which they detected the occurrence of a rubber ducky, rendering the predictive relations irrelevant. To our surprise, we observed no effect of expectation on evoked sensory activity. We discuss a number of potential factors, in relation to the existing literature, that might have played a role in the absence of the effect. We conclude that our null finding may provide an important boundary condition that can help advance our understanding of how predictive mechanisms are implemented in the brain.

Introduction

The currently most prevalent hypothesis regarding the nature of perception is that it represents a process of inference, whereby the brain attempts to infer the state of the external world on the basis of currently available sensory data and prior knowledge (Lee and Mumford, 2003; Friston, 2005; Fiser et al., 2010). On the neural level, this means that the brain's response to stimulation is codetermined by non-sensory factors, such as expectations regarding the upcoming stimulus (Summerfield and de Lange, 2014). A prime example of this is the phenomenon known as *expectation suppression*: expected stimuli elicit an attenuated neural response as compared to unexpected stimuli. Expectation suppression has been observed across a variety of species, sensory modalities and neuroimaging methods (e.g. Summerfield et al., 2008; Egner et al., 2010; Meyer and Olson, 2011; Todorovic et al., 2011; Kok et al., 2012; Todorovic and de Lange, 2012; Meyer et al., 2014; Summerfield and de Lange, 2014; St. John-Saaltink et al., 2015; Kaposvari et al., 2016; Ramachandran et al., 2016, 2017; Richter et al., 2018; Manahova et al., 2018; Utzerath et al., 2017). Moreover, it is a key ingredient in the influential predictive coding theory (Friston, 2005; Bogacz, 2017). However, despite the large amount of research that has been devoted to expectation suppression, a number of questions have remained open, the answers to which could help further develop contemporary theories.

First, despite its name, expectation suppression may in fact also refer to an enhancement of neural responses to *surprising* events, rather than to an attenuation to expected events (Kaliukhovich and Vogels, 2014; Bell et al., 2016; Ramachandran et al., 2017). Although this issue has been recognized previously (Meyer and Olson, 2011; Kaliukhovich and Vogels, 2014; Amado et al., 2016; Bell et al., 2016; Kaposvari et al., 2016; Ramachandran et al., 2017), it has so far received relatively little attention. Nevertheless, it may have important consequences for our understanding of how perception is accomplished by the brain. According to the predictive coding hypothesis, bottom-up sensory information is matched against top-down predictions by calculating their mismatch, i.e. prediction error, which is believed to lead to the recorded neural signal (Rao and Ballard, 1999; Friston, 2005; Bogacz, 2017). Therefore, when predictions are confirmed, there is no or little prediction error and accordingly a relatively low neural response is evoked as compared to an unexpected stimulus. However, it is currently unclear how exactly these responses compare to a situation where there is no expectation, i.e. a neutral condition.

A second open question concerns the representational specificity of expectation suppression. Is the reduction in activity specific to neural populations that are either sensitive (representational dampening) or non-sensitive (representational sharpening) to the stimulus about which the prediction is formed? This question has been addressed before in the literature, but the results so far have been mixed, and evidence has been found for both the sharpening (Kok et al., 2012) as well as the dampening account (Meyer and Olson, 2011; Kumar et al., 2017).

Finally, we investigated the temporal profile of expectation suppression. If the brain is to perform optimal inference regarding the outside world, then a reliable internal model is required that accurately reflects that world (Fiser et al., 2010). This includes temporal statistics: a given stimulus may be predictive of upcoming stimulation at a particular lag. In speech for instance, after having processed a syllable, the next one is expected to come in approximately 200 ms later (Arnal and Giraud, 2012). Moreover, low-level visual features in natural scenes are known to correlate strongly across short time lags, and this correlation decays rapidly over longer timescales (Dong and Atick, 1995; Kayser et al., 2003). Therefore, if the brain makes optimal use of natural statistics, then sensory predictions may be modulated by the interval between the cue and the predicted stimulus. Expectation suppression, being a proxy of the brain's predictive processing, provides a means of testing this hypothesis. We investigated this question by varying the time delay between the stimulus that induces a prediction and the stimulus about which the prediction is formed.

In the present study, we addressed these questions using a simple statistical learning paradigm, while recording magnetoencephalography (MEG) in humans. The paradigm included three conditions, in which a predictive image evoked a valid, an invalid or no prediction about an upcoming image. In addition, we introduced a block-wise manipulation in which the two images were either presented with a 300 ms temporal gap in-between or with no gap (0 ms gap). Furthermore, subjects also performed separate functional localizer blocks in which they perceived instances of the trailing images, but without modulation by any top-down components such as attention or expectation. This approach allowed us to trace the sensory-specific representational contents of the neural signal using a multivariate decoding approach (Mostert et al., 2015) as well as how these are modulated under the different conditions.

Materials and Methods

This study was pre-registered, and the research questions, design and proposed analyses, as described before the start of data acquisition, may be consulted at the following link: <https://aspredicted.org/7un39.pdf>

Subjects

Twenty-nine human volunteers were recruited from the local institute's subject pool for participation in both a behavioral training session and an experimental MEG session (see *Experimental design and procedure*). One subject was excluded from participation in the MEG session because of the presence of an orthodontic wire. Of the remaining 28 subjects, two more were excluded from further analysis due to poor MEG data quality. The final sample thus consisted of 26 subjects, of which 8 males, with a mean age of 23.4 (range 18-29). The study was approved by the local ethics committee (CMO Arnhem-

Nijmegen) and conducted according to the prescribed guidelines. All participants provided written informed consent and received either monetary compensation or study credits for participation.

Stimuli

Stimulation consisted of visually presented images of objects, obtained from the database provided at <http://cvcl.mit.edu/MM/uniqueObjects.html> (Brady et al., 2008). The images were manually inspected to remove any potentially arousing stimuli, such as pictures of insects. A random sample of five images (three leading, two trailing; see *Experimental design and procedure*) was drawn from this database separately for each subject. The stimuli were presented on a white background (luminance: 901 cd/m² for the first 8 subjects; 294 cd/m² for the remaining 18) and subtended 4° of visual angle in both horizontal and vertical direction. The fixation dot (diameter 0.2°) consisted of an outer black ring with an inner white circle. The stimuli were generated using Matlab (The Mathworks, Inc., Natick, Massachusetts, United States) with the PsychtoolBox extension (Kleiner et al., 2007) and presented by a PROPixx projector (VPixx Technologies, Saint-Bruno, QC Canada).

Experimental design and procedure

The experiment consisted of two sessions: a behavioral training session and the experimental MEG session. In the MEG session, participants performed two different types of blocks: 8 blocks in which they performed the main task and 6 functional localizer blocks.

In the main task, subjects were presented with two stimuli in succession (see Fig. 1A for details). The first stimulus (the leading image) could be either of three images with equal probability, whereas the second stimulus (the trailing image) could be either of two images. Importantly, the leading image was predictive of the trailing image with probabilities as specified in Fig. 1B. Subjects were explicitly made aware of these relations before the start of the session. This led to three conditions regarding the prediction of the trailing image: expected, neutral or unexpected. Subjects performed an oddball task in which they had to press a button with the right index finger whenever an image of a rubber ducky was presented. Importantly, the occurrence of a ducky was unrelated to the conditions, rendering the transitional probabilities task-irrelevant for the subjects. We used 16 different instances of ducky images, with a variety of colors and vantage points. This was done to motivate the subject to pay attention to the object as a whole, rather than to a single low-level feature such as the color yellow. The oddball could be presented in place of either the leading or trailing image and occurred on 10% of the trials, though never in the unexpected condition. This was done in order to avoid having too few trials in that condition, as the oddball trials were excluded from MEG analysis. Subjects received feedback at the end of the block about their performance. We introduced an

additional condition regarding the timing of the two images. In half of the blocks, there was a temporal gap of 300 ms between the leading and the trailing images, whereas the images were presented back-to-back (i.e. 0 ms gap) in the other half. Each block consisted of 100 trials and there were four blocks per temporal gap condition. This led to a total of 192 expected, 120 neutral and 48 unexpected trials after exclusion of oddballs, for each of the 0 ms and 300 ms gap conditions.

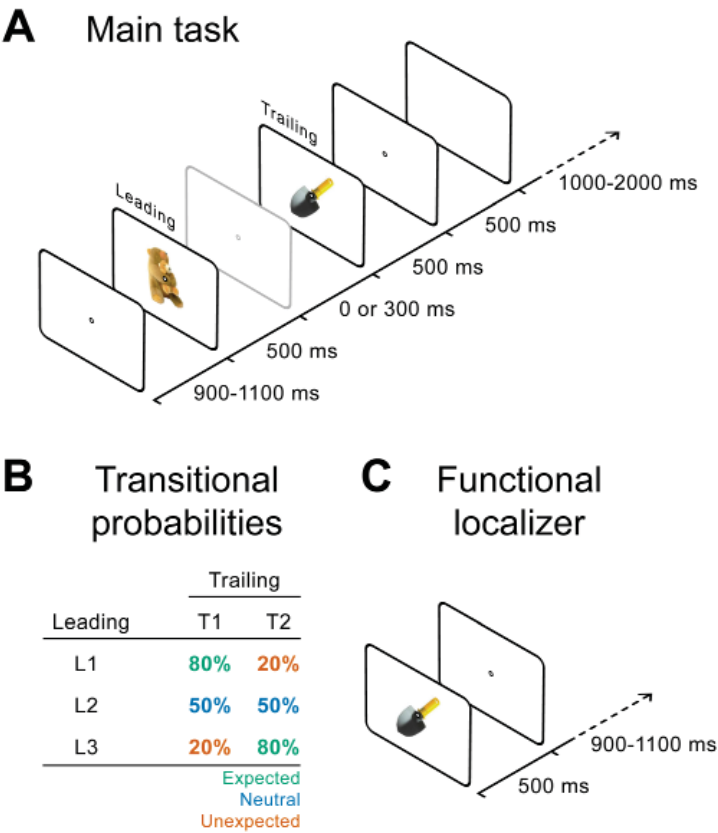


Figure 1. Paradigm and trial structure. **(A)** Each trial consisted of two images that were presented in succession, either back-to-back (0 ms gap) or with a temporal gap of 300 ms in between. Each image was presented for 500 ms and was preceded and followed by a fixation dot. A blank screen was shown in between trials for a random interval between 1 and 2 seconds. **(B)** The leading image was predictive of the trailing image. If leading image L1 was shown, then trailing image T1 followed in 80% cases whereas T2 followed in 20% of the cases, and vice versa for leading image L3. L2 provided a neutral condition, where T1 and T2 were equally likely to follow. **(C)** In separate blocks, subjects performed a functional localizer task in which they had to detect a brief ‘blink’ in the fixation dot while being continuously presented with the two leading images in pseudorandomized order. The data from these blocks were used for the decoding analysis in order to extract a bottom-up, sensory-specific activation pattern corresponding to each of the two images.

In the functional localizer, subjects were continuously presented with the two trailing images (Fig. 1C) in pseudorandomized order. The task was again an oddball, but this time subjects were instructed to detect ‘blinks’ in the fixation dot: on 10% of the trials, the inner white circle of the fixation dot would turn black during the interval of 300-400 ms after stimulus onset. This design ensured that attention was drawn to the fixation dot and not to the objects. Importantly, no predictions were induced about the stimuli. Hence, this task provided us with data in which the neural signals reflected bottom-up, sensory-specific signals, independent from top-down factors such as attention and expectation. These data were used to construct a multivariate decoder (see *Decoding analysis*) that was specifically targeted at sensory representations (Mostert et al., 2015).

The order of the blocks within the MEG session was as follows: two functional localizer blocks; four main blocks, all having the same temporal gap; two functional localizers blocks; four main blocks, again all with the same temporal gap, but different from the first four; two functional localizer blocks. Whether the first four main blocks were of the 0 ms or 300 ms condition was counterbalanced across subjects. At the end of the MEG session, subjects’ awareness of the predictive relations was assessed with five trials. Each trial tested one of the five images and subjects were required to reply which of the other four would most likely follow. The other four images were displayed on the screen, as well as an option that stated “None of these. The above image was never presented first”. For the neutral leading image, either of the two trailing images was counted as correct.

One or two days before the MEG session took place, subjects performed a behavioral training session. This session served to familiarize the subjects with the tasks as well as to entrain the predictive relations. As in the MEG session, the participants completed four full blocks of the main task, for both the 0 ms and 300 ms lags. Unlike in the MEG session however, the predictions were now 100% valid. That is, T1 always followed L1, and T2 always followed L3. Participants were explicitly made aware of these relations before the start of the experiment. After these eight blocks, subjects were briefly introduced to the functional localizer task.

MEG data acquisition and analysis

MEG data were recorded using a whole-head, 275-channel MEG system with axial gradiometers (VSM/CTF Systems, Coquitlam, BC, Canada) located in a magnetically shielded room. Using three fiducial coils, located at the nose and inside the ears, head position was monitored online and corrected if necessary. Both vertical and horizontal electrooculography (vEOG and hEOG) as well as electrocardiography (ECG) were recorded to aid in artifact rejection. Eye movements were recorded using an eye-tracker (EyeLink, SR Research Ltd., Mississauga, Ontario, Canada). All signals were sampled at 1200 Hz. The data were offline pre-processed and analyzed using FieldTrip (Oostenveld et al., 2010) (www.fieldtriptoolbox.org) and custom-made scripts in Matlab. Notch filters were applied to remove line noise (50 Hz) and its harmonics. Segments contaminated by

irregular artifacts were identified in a semi-automatic manner, whereby trials with a high variance over time and channels were marked for subsequent manual inspection. After removal of irregular artifacts, independent component analysis (ICA) was used to remove regular artifacts, such as those caused by eye blinks and heartbeats. These components were identified by correlating all components with the vEOG, hEOG and ECG. For three subjects, further stereotypical artifacts, possibly caused by miniscule traces of metal, were removed by means of principal component analysis (PCA). Finally, the data were baseline-corrected on an interval of -0.2 to 0 ms for the localizer and on -0.5 to 0 ms for the main task.

Anatomical MRI scan acquisition

Anatomical magnetic resonance imaging (MRI) scans were obtained either by inviting the participants for a third session, or by consulting the local institute's data base in case the subject had participated in an experiment before. The images were acquired using a T1-weighted MP-RAGE sequence, with a GRAPPA acceleration factor of 2 (TR = 2300 ms, TE = 3.03 ms, voxel size 1 x 1 x 1 mm, 192 transversal slices, 8° flip angle). These anatomical scans were used for source reconstruction of the MEG signals. No scan was obtained for one subject because she was MRI incompatible, but a template source model provided by FieldTrip was used instead.

MEG event-related fields and source reconstruction

Event-related fields (ERFs) were calculated by performing a synthetic planar gradiometer transformation. Occipital channels were selected as defined by the labels provided by the MEG system. In order to counter a positivity bias due to the unequal number of trials in each of the three expectation-related conditions, we performed a sub-sampling procedure. We randomly selected a subset of trials from each of the three expectation-related conditions such that the number of trials per condition matched that in the condition with the lowest number (i.e. the unexpected condition). 100 sub-selections were obtained, and the resulting ERFs were averaged to improve the signal-to-noise ratio. This procedure was performed separately for the 0 ms and 300 ms temporal gap conditions.

In addition to ERFs, we also obtained time-resolved activity traces at the source level in both the functional localizer and main task using linearly constrained maximum variance (LCMV) beamformers (Van Veen et al., 1997). Individual regular source grids were constructed on the basis of the anatomical MRI scans and subsequently normalized to MNI space using FieldTrip. Volume conduction models were based on a single shell model of the inner surface of the skull. Lead fields were calculated and rank-reduced to a dimensionality of two to accommodate the fact that MEG is blind to tangential sources. After that, spatial filters were constructed using the data covariance over a window of 80-500 ms relative to stimulus onset (leading image in the main task). The covariance was subsequently regularized using shrinkage (as described by Blankertz et al., 2011) with a

parameter of 0.05. Applying the spatial filters to the data resulted in a two-dimensional estimate of the dipole moment, per grid point, over time. A region of interest was selected on the basis of the spatial topography of the activity evoked by both stimuli in the localizer pooled. Since we were interested in the magnitude of source activity and not dipole orientation, as well as to facilitate visualization, we further reduced these estimates to a scalar by taking the norm of the vector. However, as this always results in a positive value, the noise in the data inherently results in a positivity bias. We countered this bias by calculating an estimate of the noise using a permutation procedure (see Manahova et al., 2018 for more details), whereby we randomly flipped the sign on half of the trials. Such a procedure effectively cancels out the stimulus-evoked signal, leaving an estimate of the noise. This estimate was subtracted from the true data. In addition, when inspecting the spatial topography of the source reconstruction, the data were also divided by the noise estimate as a countermeasure to the depth bias. The number of permutations was 1,000.

MEG population decoding analysis

As we were interested in whether the effect of expectation suppression was different for neural populations that prefer the trailing image versus populations that do not prefer the image, we applied a multivariate decoding technique to trace the activity of two populations that specifically code for the two trailing images. We postulated that the activity of such a population can be described by a latent component whose activity reflects the degree to which its associated stimulus is represented in the brain. Each of these components has a corresponding sensor topography that is composed of the aggregate pattern elicited by the underlying neural population. In a first step, these spatial patterns are extracted from the data, effectively formulating a forward model that describes how the sensor data vary as function of the latent component. In a second step, this model is inverted, yielding spatial filters that recover the activity of the latent component, given the sensor data. Specifically, the forward model is simply the average signal evoked by the stimulus. Then, if \mathbf{m} is a column vector that contains this average signal at a given time point, then the corresponding spatial filter \mathbf{w} is calculated as follows:

$$\mathbf{w} = \frac{\mathbf{S}^{-1}\mathbf{m}}{\mathbf{m}^T\mathbf{S}^{-1}\mathbf{m}}$$

where \mathbf{S} is the regularized noise covariance (using shrinkage as described in Blankertz et al., 2011; regularization parameter of 0.01). This formula is equivalent to the equation used to calculate LCMV filters (Van Veen et al., 1997). In other words, decoding of the population activity may be regarded as source reconstruction using beamformers, whereby the “lead fields” are defined functionally on the basis of empirically observed data, rather than on a theoretical volume conduction model (Mostert et al., 2015).

The result is two separate independent decoders, for each of the two trailing images, which yield the degree to which that image is represented in the neural signal. Application of the spatial filter to testing data \mathbf{x} is accomplished by the taking inner product:

$$\mathbf{y} = \mathbf{w}^T \mathbf{x}$$

where \mathbf{y} is the decoded signal. We constructed the decoders on the basis of the functional localizer and subsequently applied them to the data from the main task. Moreover, we trained as well as applied the decoders across all time points in order to inspect the temporal dynamics of the underlying code (King and Dehaene, 2014).

Statistical inference

Statistical significance was tested at the group level by means of permutation tests, using a cluster-based correction for multiple comparisons (Maris and Oostenveld, 2007). Clusters were defined along both the spatial (sensors) and temporal dimension for ERF analysis. For the temporal generalization matrices, clusters were defined along the two-dimensional time axes. Individual data points were labeled for clustering if their univariate p -values, as obtained using two-tailed paired t -tests, were below a threshold of $p = 0.05$. The t -values were summed within each cluster, separately for negative and positive clusters. A cluster in the true data was considered significant if its corresponding p -value was lower than 0.05. The number of permutations was 1,000 for the ERF analysis, and 10,000 for all other analyses.

In addition to frequentist null hypothesis-testing, we also performed a Bayesian statistical analysis by calculating the Bayes factor for the contrast of expected versus unexpected. For brevity, as well as to contain the number of multiple comparisons, this was done only for averaged occipital activity in the ERF analysis. An advantage of the Bayes factor is that it allows for testing of invariances between conditions (Rouder et al., 2009; Jarosz and Wiley, 2014). We calculated the Bayes factor on the basis of equation 1 in Rouder et al. (2009), using the non-informative JZF prior (see Rouder et al., 2009), separately per time point. This was done for the ERFs, for the reconstructed source traces and for both the preferring and non-preferring populations within the decoding analysis. The Bayes factor is interpreted as a ratio between the likelihood that the data is observed under the null hypothesis (i.e. no difference) and the likelihood that the data is observed under the alternative hypothesis (i.e. there is a difference).

Standard error of the means displayed in the figures are calculated using an adjustment for within-subject comparisons (Rouder et al., 2009).

Results

Twenty-six human participants performed an oddball task in a statistical learning paradigm. On each trial, two images (leading and trailing) were shown in rapid succession (Fig. 1A). On half of the blocks, they were presented back-to-back, whereas there was a temporal gap of 300 ms in the other half. The leading image could be one of three, whereas the trailing image had two options. The leading image was predictive of the trailing image

according to the transitional probabilities as summarized in Fig. 1B, which led to three conditions: expected, neutral and unexpected. The task was to detect the rare occurrence of a rubber ducky, thus rendering the predictive relations task-irrelevant.

Behavioral results

In the main task, subjects correctly detected the oddball on 96% of the occurrences (SEM = 0.6%, range = 90% - 100%), while falsely reporting one on only 0.06% of the non-oddball trials (SEM = 0.02%, range = 0% - 0.3%), indicating that they were well able to identify the presented objects. The assessment at the end of the experiment indicated that subjects were generally well aware of the predictive relations. For the three leading images, they correctly identified the trailing image on 92% of the questions.

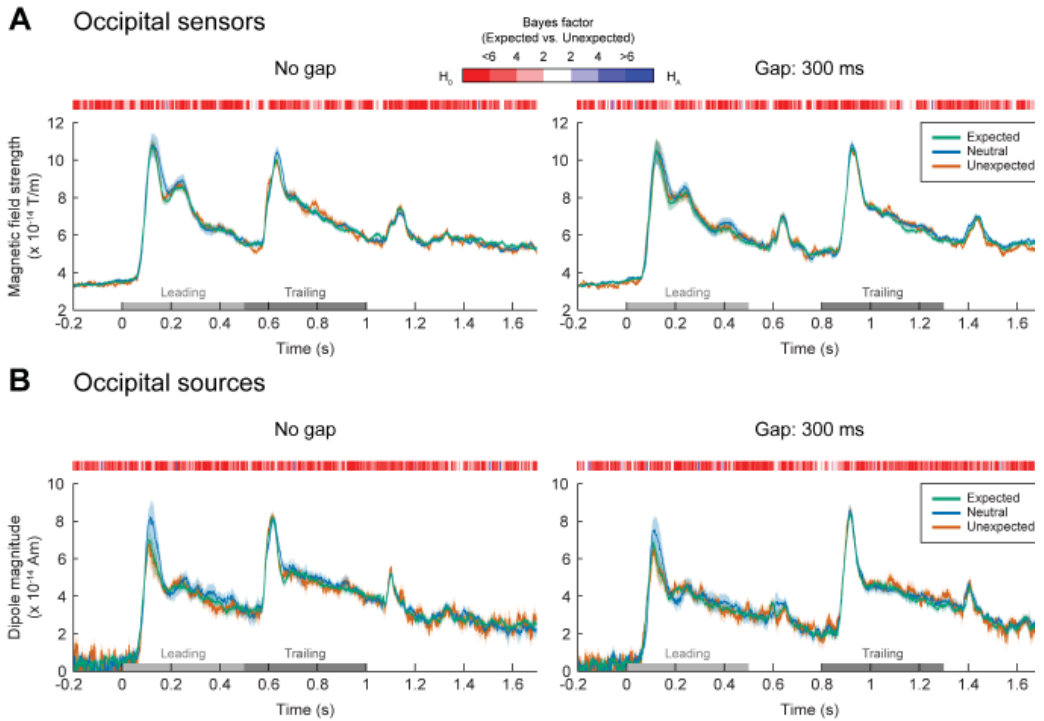


Figure 2. Neural activity over time, averaged across occipital synthetic planar sensors **(A)** and in reconstructed occipital sources **(B)**, separately for the expected, neutral and unexpected conditions, as well as separately for the no temporal gap (left figures) and 300 ms gap (right figures) conditions. The (mostly) red stripes above each figure represent the Bayes factor between the expected and unexpected condition. No significant differences were found between the three prediction-related conditions, and the Bayesian analyses provided “positive” or “substantial” evidence for invariance between the expected and unexpected conditions, for all of the four panels. The shaded areas demarcate the standard error of the mean, adjusted for within-subjects contrasts. The gray bars at the bottom of each figure indicate the stimulus presentations.

For the two trailing images, they correctly recognized that they were never presented first on 85% of the questions. In the localizer task, subjects correctly detected the oddball on 99% of the occurrences (SEM = 0.4%, range = 94% - 100%), while incorrectly reporting an oddball on 0.1% of the non-oddball trials (SEM = 0.01%, range = 0% - 1.2%). These results verify that they were paying attention to the task at hand.

No modulation of event-related fields by expectation

In order to assess if the visual neural response was modulated by whether the trailing image was expected, unexpected or neutral, we looked at the synthetic planar ERF. The results showed a clear evoked response to the onset of both the leading and the trailing image (Fig. 2A). However, no differences were found between the three prediction-related conditions. Contrary to our hypothesis, the activity in the expected, neutral and unexpected all overlapped, and indeed no significant difference clusters were found in any of the three comparisons. This was true both when the images were presented back-to-back and when there was a temporal delay of 300 ms in between them. Note that although only the occipital channels are shown in Fig. 2A, the statistical test was performed across all sensors (and time points). For both temporal gap conditions, the Bayesian analysis provided some evidence for equality of the expected and unexpected condition. That is, throughout most of the segment, the Bayes factor ranged between approximately 4 and 7 in favor of the null hypothesis, which may be regarded as “positive” or “substantial” evidence (Jarosz and Wiley, 2014). Note that the Bayesian analysis was performed only on the averaged occipital data, i.e. did not involve the sensor dimension.

No modulation of occipital sources by expectation

In addition to ERFs at the sensor level, we also reconstructed activity traces at the source level. These may provide a more spatially precise view of the activity in visual areas, as compared to the spatially coarse signals at the sensor level. We first selected a region of interest on the basis of the functional localizer. In these data, we reconstructed the source topography of the activity evoked by both (trailing) images, averaged over the time window of 110 to 140 ms. This analysis identified the occipital poles as most responsive (Supplementary Fig. S1). We subsequently calculated the time-resolved activity traces for these sources in the main task, averaging over the two hemispheres. Similarly to the ERFs, these results showed a clear evoked response to the onset of both the leading and trailing image (Fig. 2B). However, no significant differences were found between the three prediction-related conditions. As in the ERF analysis, this was true for both the 0 ms and the 300 ms temporal gap condition. Furthermore, the Bayes factor again ranged between approximately 4 and 7 throughout most of the trial, providing “positive” or “substantial” (Jarosz and Wiley, 2014) evidence for no difference between the expected and unexpected conditions.

No modulation of population-specific activity by expectation

The analyses above focused on the overall activity evoked by the trailing images. However, it is possible that the effect of expectation suppression may be specific to whether a neural population prefers the stimulus or not (Meyer and Olson, 2011; Kok et al., 2012; Kumar et al., 2017). For verification purposes, we first tested the decoders on the functional localizer data using cross-validation. We found that the uncovered latent components indeed showed a larger response when the preferred stimulus was presented as compared to when the non-preferred stimulus was presented (Supplementary Fig. S2A), verifying that we were able to decode object identity from the MEG signal. Moreover, source localization of the decoder's corresponding spatial patterns revealed a primary contribution from occipital sources, which suggests that the latent components indeed correspond to vision-related neural populations.

Having established the validity of the method, we then decoded the signals in the main task using decoders trained on the basis of the functional localizer. We specifically looked at a training time window of 100-200 ms, because the decoding analysis within the localizer showed a peak during that time period (Supplementary Fig. S2A). As expected, the recovered latent components also show an increased response for preferred trailing images as compared to non-preferred images (Fig. 3, Supplementary Fig. S3).

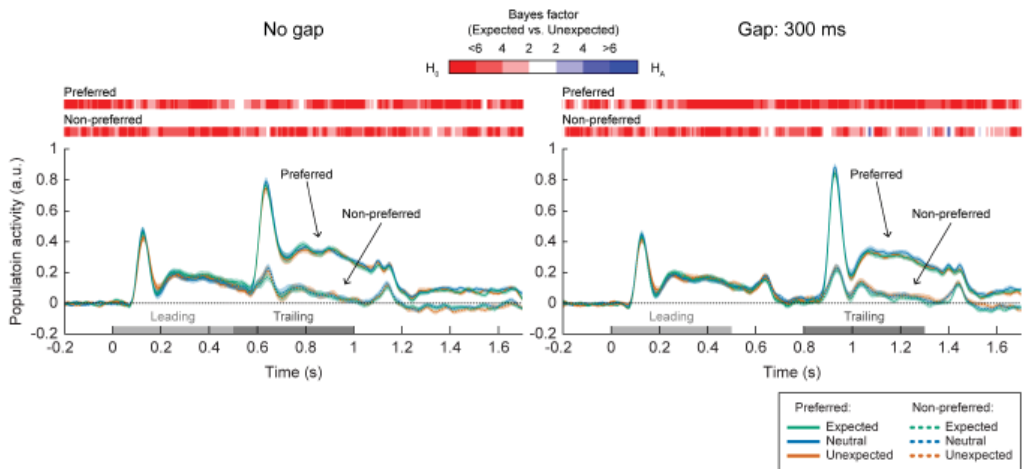


Figure 3. Decoding population activity traces, binned according to their preference to the trailing image, separately for the 0 ms (left) and the 300 ms (right) gap condition. The (mostly) red stripes above each figure represent the Bayes factor between the expected and unexpected condition, separately for the case where the preferred stimulus was shown and the case where the non-preferred stimulus was shown. No significant differences were found between the three prediction-related conditions, and the Bayesian analyses provided “positive” or “substantial” evidence for invariance between the expected and unexpected conditions, for both the 0 ms and 300 ms gap condition and within both the preferred and non-preferred populations. The shaded areas demarcate the standard error of the mean, adjusted for within-subjects contrasts. The gray bars at the bottom of each figure indicate the stimulus presentations.

However, no significant differences were found between the expected, neutral and unexpected condition. As before, the Bayesian analysis provided “substantial” or “positive” evidence for the null hypothesis at the vast majority of time points. This was the case for both preferred and non-preferred stimuli, as well as for the 0 ms and 300 ms gap condition.

The absence of an effect cannot be explained by the specific choice of the training time window of 100-200 ms, because the pattern is highly stable throughout large parts of the trial as revealed by temporal generalization matrices, for both the within-localizer decoding (Supplementary Fig. S2B) and the between-task generalization (Supplementary Fig. S3). In other words, the results would look very similar, regardless of the exact choice of training time.

Discussion

Expectation suppression is a well-studied phenomenon that has delivered a significant contribution to the development of contemporary theories about brain functioning (Friston, 2005; Summerfield and de Lange, 2014). However, a number of important questions have yet remained unanswered. Using a simple statistical learning paradigm, we aimed to address three main questions: 1) Is expectation suppression the result of attenuation by expected events, or instead due to an enhanced response by a surprising event? 2) Is expectation suppression specific to neural populations that either prefer or do not prefer the stimulus in question? 3) Is the magnitude of expectation suppression modulated by a temporal delay between the predictive and predicted event? Surprisingly, we found no evidence for expectation suppression in our paradigm. We are therefore unable to draw any firm conclusions about our main questions. However, the lack of evidence for expectation suppression itself may be regarded as a potentially interesting result. Given the widespread occurrence of expectation suppression, across a variety of species, sensory modalities and neuroimaging techniques (e.g. Summerfield et al., 2008; Egner et al., 2010; Meyer and Olson, 2011; Todorovic et al., 2011; Kok et al., 2012; Todorovic and de Lange, 2012; Meyer et al., 2014; Summerfield and de Lange, 2014; St. John-Saaltink et al., 2015; Kaposvari et al., 2016; Ramachandran et al., 2016, 2017; Richter et al., 2018; Manahova et al., 2018; Utzerath et al., 2017), a counterexample to its ubiquity may help constrain and set boundary conditions for current theories. Although the definitive answer as to why our experiment failed to show expectation suppression remains open, we discuss a number of potential factors below that may be important to consider for future research.

A potentially important factor is the number of stimuli involved and the complexity of their relations. In our study, we defined three leading images and two trailing, resulting in a 2-by-3 contingency table. This is a simple design, whereby the stimuli about which a prediction is formed could be one of only two options. This is in contrast to previous studies that did observe expectation suppression in a statistical learning paradigm very

similar to ours, but where the number of options was six (Meyer and Olson, 2011) or eight (Ramachandran et al., 2016). Indeed, a recent study with a very similar design as ours did find widespread expectation suppression using fMRI, with the main difference that they used eight stimuli (Richter et al., 2018). Thus, in our case one could argue that both trailing images were actually highly expected, leading to the possibility that both conditions already displayed maximum expectation suppression.

Other relevant factors to consider are attentional and/or task demands. In the current experiment, the leading and trailing images were task-relevant in order to detect the oddball, but the predictive relations between them were not. Although a similar arrangement has previously successfully led to the observation of expectation suppression (Todorovic and de Lange, 2012, 2012; Manahova et al., 2018), the fact that we did not observe it contributes to the growing controversy surrounding the phenomenon. Indeed, while there is convincing evidence that shows that expectation suppression is a largely automatic phenomenon that does not require the predicted features to be attended or task relevant (Kok et al., 2012), a more recent study found an abolishment of expectation suppression in a very similar paradigm where the stimuli were attended and task relevant (St. John-Saaltink et al., 2015). A similar abolishment of expectation suppression when attention is drawn away has been reported before (Larsson and Smith, 2012). Moreover, in addition to the effects of attention, St. John-Saaltink et al. (2015) also found an effect of task demands on expectation suppression. Neural activity was only suppressed for expected stimuli when subjects were concurrently engaged in a perceptually demanding task, but not while engaged in a working memory task. The authors explained this by noting that both expectations and working memory rely on a similar underlying mechanism, namely the formation of stimulus templates in sensory cortex (Albers et al., 2013; Kok et al., 2014, 2017; Reddy et al., 2015). Thus, enforcing a load on working memory may hamper the formation of stimulus templates for the expected stimulus and thereby abolish expectation suppression (St. John-Saaltink et al., 2015). A similar explanation could be put forward for our (lack of) expectation suppression. Subjects were instructed to press a button in response to the occurrence of a rubber ducky. Therefore, while the predictive relations were task-irrelevant and subjects presumably did not form any explicit expectation about the upcoming stimulus, it is plausible that subjects continuously maintained an expectation of the oddball stimulus. In other words, the ongoing instantiation of a sensory template of rubber duckies may have impeded the (automatic) formation of expectations about the trailing image.

Furthermore, comparing our experiment previous studies that also used a statistical learning paradigm (Meyer and Olson, 2011; Meyer et al., 2014; Ramachandran et al., 2016, 2017), one big difference (except for the species) is the extent of the training period. In these studies, monkeys were trained for weeks, whereas in our experiment the humans only received a short ~1h training session one or two days before the experimental session. This difference in amount of training may have resulted in differential recruitment of neural mechanisms responsible for effects of prediction. The studies mentioned above

recorded from sensory cortex and, due to the extensive exposure periods, it is possible that the prediction-related effects are instantiated by local, gradually emerged connections. On the other hand, rapid learning after a short training session is thought to be mediated by hippocampal regions (Schapiro et al., 2012).

Another factor that could contribute to the manifestation of expectation suppression is the exact manner by which the visual stimuli are presented. In line with previous studies (Meyer and Olson, 2011; Ramachandran et al., 2016, 2017), we chose to present images in pairs. However, the MMN is generally observed in a stream of stimuli (Garrido et al., 2009). Given that expectation suppression as sought after in the current study is temporal in nature - that is, the leading image induces an expectation about the trailing image at a later point in time - it is conceivable that the temporal context may play a role. Exposing subjects to the rhythmicity of a visual stream induces oscillations not only in neural activity, but also in perception (Spaak et al., 2014). Put differently, the entrainment of ongoing internal oscillations to external stimuli allows for an optimal preparation in order to process an upcoming stimulus. In the context of expectation suppression, this suggests that a visual stream may aid in instantiating top-down predictions. Indeed, the importance of neural oscillations in guiding the temporal aspect of sensory predictions has been pointed out before (Arnal and Giraud, 2012). Accordingly, a recent study did find expectation suppression using the same stimuli and oddball task as ours, but presented the images in rapid streams instead (Manahova et al., 2018).

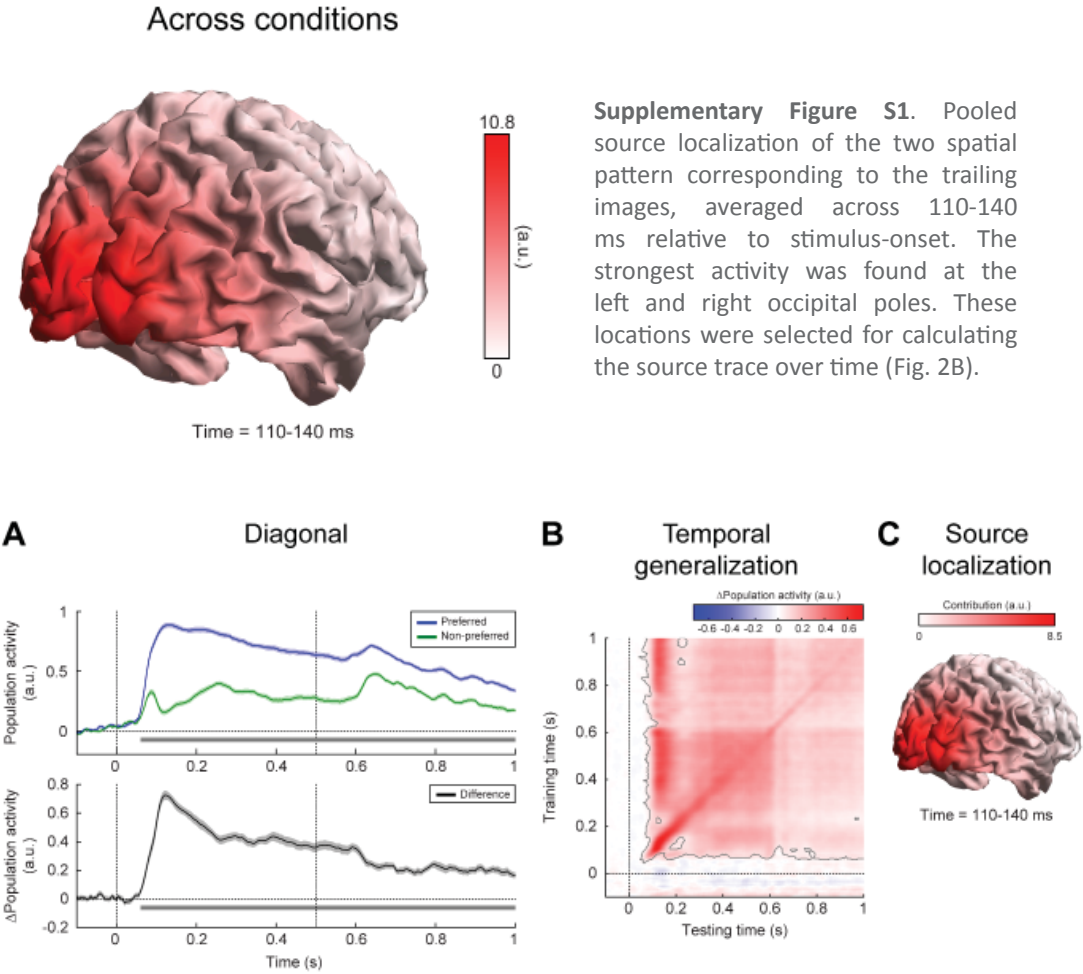
Furthermore, one could consider methodological reasons for our lack of expectation suppression. An obvious culprit would be the lack of sufficient statistical power or a poor signal-to-noise ratio (SNR). However, we consider this to be unlikely for the following two reasons. First, we observed clear stimulus-evoked signals, with small errors bars. Moreover, we were well able to decode the stimulus identity with high statistical significance. Second, our Bayesian statistical analysis provided positive/substantial evidence for the null hypothesis. If our data suffered from a low SNR, then these analyses would have yielded inconclusive results. Thus, either there was no difference between the expected and unexpected condition, or the difference was exceedingly small - both of which are surprising in the light of existing literature.

Finally, it is important to consider the specific neuroimaging technique that was used. Expectation suppression has been observed using a wide array of methods, including single-cell recordings (Meyer and Olson, 2011; Kaposvari et al., 2016), functional magnetic resonance imaging (Alink et al., 2010; Kok et al., 2012), electroencephalography (Stefanics et al., 2014; Kimura and Takeda, 2015), functional near-infrared spectroscopy (Emberson et al., 2015) and magnetoencephalography (Todorovic et al., 2011; Wacongne et al., 2011; Todorovic and de Lange, 2012; Cashdollar et al., 2016). Nevertheless, using varying neuroimaging techniques has led to mixed results. A clear example is a study by Kok et al. (2017) in which they used a paradigm very similar to the one used in Kok et al. (2012), except neural activity was now measured using MEG instead of fMRI. Although they replicated the effect that expected gratings could be decoded more accurately as

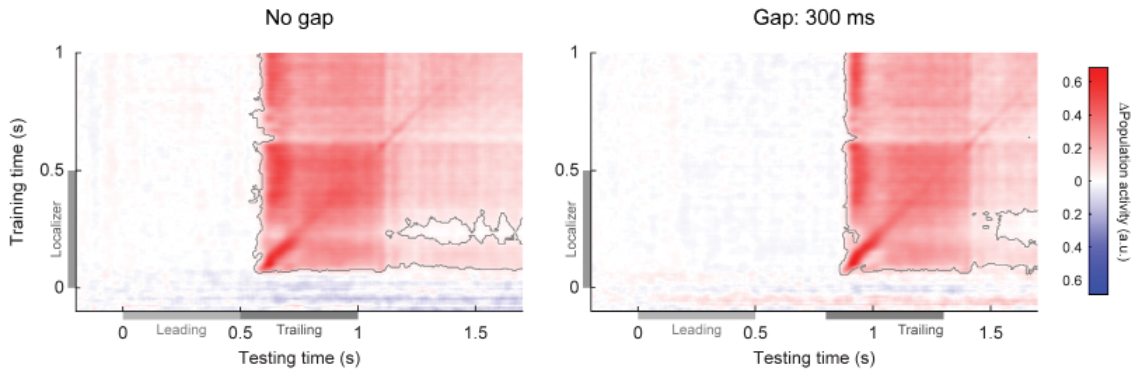
compared to unexpected gratings, the MEG data did not display an attenuated evoked response for expected stimuli (Kok et al., 2017), in stark contrast to the fMRI data (Kok et al., 2012). Another example can be found in the MEG study by Cashdollar et al. (2016), where expected stimuli showed a decrease response as compared to unexpected stimuli, but only in magnetometers and not in combined gradiometers.

In summary, we aimed to address three outstanding questions regarding the well-established phenomenon of expectation suppression using a simple statistical learning paradigm. However, to our surprise, we did not observe any expectation suppression at all. Given the literature, this is a remarkable result that may help in further understanding the phenomenon. We reviewed a number of factors that may potentially have contributed to our finding, though the definitive answer remains elusive. We propose that future research be conducted in order to further refine the constraints and boundary conditions under which expectation suppression manifests itself. This will help development and refinement of contemporary theories about brain functioning.

Supplementary Figures



Supplementary Figure S2. Populating decoding results within the functional localizer using cross-validation, showing that the two stimuli could readily be decoded. **(A)** Decoded population activity for matched training and testing time points [i.e. diagonal in temporal generalization matrix in **(B)**], separately for preferred and non-preferred stimuli (top) and their difference (bottom). The horizontal gray bar (identical in top and bottom figure) denotes time points belonging to the significant cluster shown in **(B)**. The shaded areas demarcate the standard error of the mean, adjusted for within-subjects contrasts, and the vertical dotted lines indicate the onset and offset of the stimulus. **(B)** Temporal generalization matrix, showing the difference between population activity for preferred and non-preferred stimuli (cf. **(A)**, bottom). The gray outline indicates a significant cluster ($p \approx 0$). **(C)** Source localization of the difference of the spatial patterns corresponding to each of the two stimuli, averaged over 110-140 ms after stimulus onset. This topography displays the degree to which a given neural source contributes to the decoding performance (i.e. the difference between preferred and non-preferred).



Supplementary Figure S3. Full temporal generalization matrices, displaying decoding performance in the main task, pooled across all leading images (i.e. irrespective of expectation condition), using decoders training on the functional localizer. Decoding performance refers to the difference population activity in response to preferred and non-preferred stimuli (cf. Supplementary Fig. S2A, bottom and Supplementary Fig. S2B). The results show that the identity of the trailing image can be decoded reliably on the basis of decoders trained on the functional localizer data. The gray outlines demarcate significant clusters ($p \approx 0$, for both temporal gap conditions). The gray bars along the figure's axes indicate presentation of the leading and trailing images.



5

**Eye movement-related confounds
in neural decoding of
visual working memory representations**

Abstract

A relatively new analysis technique, known as neural decoding or multivariate pattern analysis, has become increasingly popular for cognitive neuroimaging studies over recent years. These techniques promise to uncover the representational contents of neural signals, as well as the underlying code and the dynamic profile thereof. A field in which these techniques have led to novel insights in particular is that of visual working memory (VWM). In the present study we subjected human volunteers to a combined VWM/imagery task while recording their neural signals using MEG. We applied multivariate decoding analyses to uncover the temporal profile underlying the neural representations of the memorized item. Analysis of gaze position however revealed that our results were contaminated by systematic eye movements, suggesting that the MEG decoding results from our originally planned analyses were confounded. In addition to the eye movement analyses, we also present the original analyses to highlight how these might have readily led to invalid conclusions. Finally, we demonstrate a potential remedy, whereby we train the decoders on a functional localizer that was specifically designed to target bottom-up sensory signals and as such avoids eye movements. We conclude by arguing for more awareness of the potentially pervasive and ubiquitous effects of eye movement-related confounds.

This chapter has been published as:

Mostert, P., Albers, A. M., Brinkman, L., Todorova, L., Kok, P., & de Lange, F. P. (2018) Eye movement-related confounds in neural decoding of visual working memory representations. *eNeuro* 0401-17.2018. doi: 10.1523/ENEURO.0401-17.2018

Introduction

Neural decoding, or multivariate pattern analysis (MVPA), is a popular analysis technique that has obtained considerable momentum in the field of cognitive neuroimaging (Haxby et al., 2014; Grootswagers et al., 2016). It refers to uncovering a factor of interest, for instance stimulus identity, from multivariate patterns in neural signals such as those measured by magnetoencephalography (MEG) or functional magnetic resonance imaging (fMRI). Decoding allows one to probe the representational contents of a neural signal, rather than overall activity levels, with superior sensitivity. However, this sensitivity may require extra vigilance at the end of the user, because these analyses may also be particularly sensitive to potentially confounding factors. Here we demonstrate such an example, specifically in the context of visual working memory (VWM), where a decoding analysis is contaminated by stimulus-specific eye movements. Given the widespread use of these techniques and its pivotal contributions to contemporary VWM theories, we argue that appreciation of these potential caveats is important.

Visual working memory is the ability to retain and utilize visual information about the world for a short period of time, even when the original external source of that information is no longer available. Neural decoding has been frequently applied in the study of VWM in order to elucidate where, when and how a memorandum is encoded in the brain. This was first demonstrated by Harrison and Tong (2009) and Serences et al., (2009), who were able to decode the orientation of a memorized grating from visual cortex. Further VWM decoding studies extended Harrison and Tong's (2009) paradigm in varying ways to study, among others, mental imagery, mental transformations and spatial working memory (Albers et al., 2013; Christophel et al., 2015; Foster et al., 2016; Christophel et al., 2017; Gayet et al., 2017). The paradigm has also been ported to electrophysiological studies using MEG or electroencephalography to capitalize on the high temporal resolution offered by those methods (Wolff et al., 2015; Foster et al., 2016; King et al., 2016; Wolff et al., 2017). These results have led to important new theories, among others the idea that high-fidelity VWM representations are stored in early sensory cortex (Albers et al., 2013; Sreenivasan et al., 2014), the activity-silent coding hypothesis (Stokes, 2015; Wolff et al., 2015; Rose et al., 2016; Rademaker and Serences, 2017; Wolff et al., 2017) and the dynamic coding framework (Stokes et al., 2013; Stokes, 2015; King et al., 2016; Spaak et al., 2017).

In the current study, human volunteers performed a combined VWM/imagery task, while we traced the representational contents of their neural activity as measured by MEG. The experiment was designed to elucidate the temporal profile of the memorized item's neural representation. However, control analyses revealed that our data were severely contaminated by small eye movements. In this paper, we first describe the eye movement analysis to show how the identity of the memorized item could be decoded from gaze position. Next, we present the naive results as they would have been, had we not been aware of the confound. This highlights how these could easily have been mistaken to

provide genuine insight into the neural mechanisms underlying VWM. Finally, we present a potential solution by training the decoders on separate functional localizer blocks, which allowed us to extract the sensory-specific neural patterns, thereby effectively bypassing the eye-movements confounds.

Materials and Methods

Subjects

Thirty-six human volunteers were recruited from the local institute's subject pool to participate in a behavioral screening session. Of these, 24 (thirteen male; mean age: 26.8 year, range: 18-60) were selected to participate in the MEG experiment (see *Experimental design and procedure*). Of these 24 selected subjects, three were excluded from MEG analysis due to poor data quality and another four were excluded from the analyses regarding eye movements, because the eye-tracker failed to track the eye reliably in those subjects. The experiment was approved by the local ethics committee and conducted according to the guidelines set out by the committee. All participants provided written informed consent and received either monetary compensation or course credits.

Stimuli

Stimulation was visual and consisted of sinusoidal gratings with a spatial frequency of 1 cycle/°, 80% contrast and one random phase per experimental block. The gratings were masked at an outer radius of 7.5° and an inner aperture radius of 0.7°,

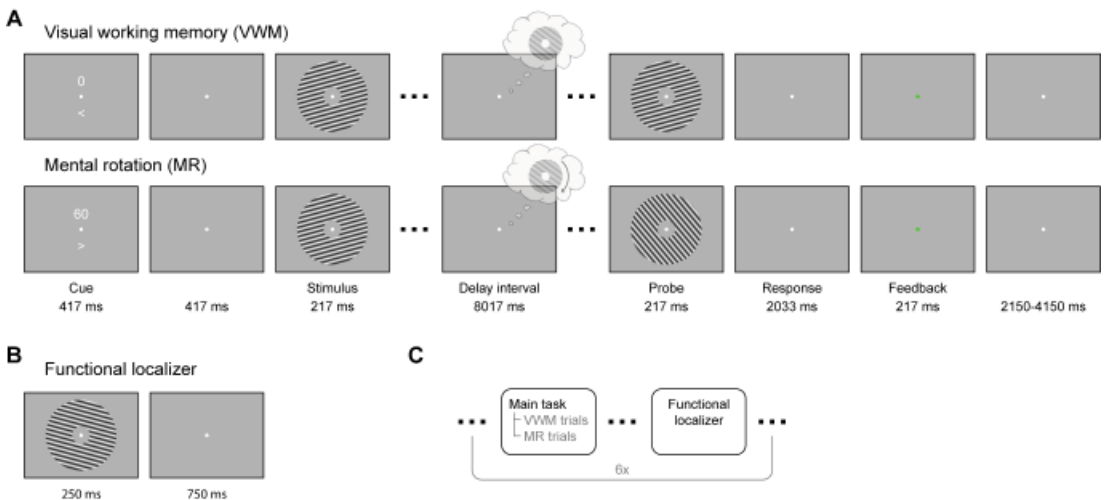


Figure 1. Experimental paradigm. **(A)** In the combined VWM/imagery blocks, subjects were instructed to vividly imagine a grating and to either keep that in mind (VWM condition) or rotate it mentally over a cued number of degrees (MR condition). **(B)** In the functional localizer block, oriented gratings were continuously presented while the subject's attention was drawn to a task at fixation. **(C)** The VWM/imagery and localizer blocks were performed in alternating order.

and presented on a gray background (luminance: 186 cd/m²). Stimuli were generated and presented using MATLAB with the Psychtoolbox extension (Kleiner et al., 2007).

Experimental design and procedure

The main task was to vividly imagine and remember an oriented grating and, in some conditions, mentally rotate this grating over a certain angle. Each trial began with a dual cue that indicated both the amount (presented above fixation) and the direction ('>' for clockwise and '<' for counter-clockwise, presented below fixation) of mental rotation that was to be performed in that trial (Fig. 1). The amount could be either 0°, 60°, 120° or 180°, in either clockwise or counter-clockwise direction, where 0° corresponded to a VWM task. This condition will henceforth be referred to as the VWM condition, and the other three conditions, which corresponded to imagery, as the mental rotation (MR) conditions. This cue lasted for 417 ms, after which a blank screen was shown for another 417 ms. A fixation dot (4 pixels diameter) was present throughout the entire trial, and throughout the entire block. After the blank, a grating was presented for 217 ms that could have either of three orientations: 15°, 75° or 135° (clockwise with respect to vertical). Next, a blank delay period of 8017 ms followed, during which subjects were required to keep the starting grating in mind and, in a subset of trials, mentally rotate it. The delay period was terminated by the presentation of a probe grating for 217 ms, whose orientation was slightly jittered (see *Staircase procedure*) with respect to the orientation that subjects were supposed to have in mind at that moment. Subjects then indicated with a button press whether the probe was oriented clockwise or counterclockwise relative to their internal image. The response period lasted until 2033 ms post probe, after which feedback was given. There were 3 trials per design cell (4 arcs of rotation and 2 directions) per block, resulting in 24 trials per block. In addition, there were two catch trials per block, in which the probe grating was presented at an earlier moment in the delay interval in order to gauge ongoing rotation. All trials were presented in pseudorandomized order. The catch trials were excluded from further analysis, because subjects indicated to find them difficult and confusing. In general, each experiment consisted of six experimental blocks (though some subjects performed 5, 7 or 8 blocks), preceded by one or more practice blocks, resulting in a total of 144 experimental trials for most of the participants.

Interleaved with the VWM/imagery blocks, there were six functional localizer blocks (Fig. 1B, C). In these blocks, gratings of six different orientations (15° to 165°, in steps of 30°) were presented for 250 ms with an inter-trial interval of 750 ms. Each block consisted of 120 trials, resulting in a total of 120 trials per orientation. The task was to press a button when a brief flicker of the fixation dot occurred. Such a flicker occurred between 8 to 12 times (randomly selected number) per block, at random times. Using such a task we ensured that spatial attention was drawn away from the gratings while stimulating subjects to maintain fixation, allowing us to record activity that predominantly reflected bottom-up, sensory-specific signals (Mostert et al., 2015).

Prior to the MEG session, the volunteers participated in a behavioral screening session that

served to both train the subjects on the task as well as to assess their ability to perform it. Subjects were instructed to mentally rotate the stimulus at an angular velocity of 30 °/s by demonstrating examples of rotations on the screen. Moreover, during this session subjects were required to press a button as soon as they achieved a vivid imagination of the grating upon completion of the cued rotation. This provided a proxy of the speed at which they actually performed the rotation and was used as selection criterion for participation in the MEG session.

Staircase procedure

The amount of jitter of the probe grating was determined online using an adaptive staircasing procedure to equalize subjective task difficulty across conditions and subjects. The starting difference was set to 15°, and was increased by 1° following an incorrect response and decreased by 0.5488 after two consecutive correct responses. Such a procedure has a theoretical target performance of approximately 80% correct (García-Pérez, 1998). Four separate staircases were utilized, one for each of the VWM and MR conditions.

MEG recordings, eye-tracker recordings and pre-processing

Neural activity was measured using a whole-head MEG system with 275 axial gradiometers (VSM/CTF Systems, Coquitlam, BC, Canada) situated in a magnetically shielded room. A projector outside the room projected via a mirror system onto the screen located in front of the subject. Fiducial coils positioned on the nasion and in the ears allowed for online monitoring of head position and for correction in between blocks if necessary. Both vertical and horizontal EOG as well as electrocardiogram were obtained to aid in the recognition of artifacts. All signals were sampled at 1200 Hz and analyzed offline using the FieldTrip toolbox (Oostenveld et al., 2010). The data were notch-filtered at 50 Hz and corresponding harmonics to remove line noise, and subsequently inspected in a semi-automatic manner to identify irregular artifacts. After rejection of bad segments, independent component analysis was used to remove components that corresponded to regular artifacts such as heartbeat, blinks and eye movements (although our results suggest that the removal of eye movement-related artifacts was imperfect, see *Results and Discussion*). The cleaned data were baseline-corrected on the interval of -200 to 0 ms, relative to stimulus onset.

Gaze position and pupil dilation were continuously tracked throughout the experiment using an Eyelink 1000 (SR Researcher) eye-tracker. The eye-tracker was calibrated before each session and signals were sampled at 1200 Hz. Because we were interested in eye-movements induced by the experimental stimulation, we removed any slow drifts in the signal by baseline-correcting the signal on an interval of -200 to 0 ms relative to cue onset.

Data sharing

All data, as well as analysis scripts required to obtain the presented figures, are available from the Donders Institute for Brain, Cognition and Behavior repository at http://hdl.handle.net/11633/di.dccn.DSC_3018016.04_526.

Classification and decoding analyses

Originally, we first focused on the neural data. Broadly, we conducted two lines of decoding analyses. In the first, we focused only on the blocks in which participants performed the combined VWM/imagery task, using 8-fold cross-validation. We trained a three-class probabilistic classifier that returns the probability that a given trial belongs to either of the three presented grating orientations. In order to improve signal-to-noise ratio, yet retain the ability to draw firm conclusions regarding the timing of any decoded signal, we smoothed the data using a moving average with a window of 100 ms. The classifier was trained across the spatial dimension (i.e. using sensors as features), on trials from all conditions (i.e., all amounts and directions of rotation). This may seem counterproductive, because the mental contents diverge over the delay interval and there should therefore be no systematic relationship between the MEG data and the stimulus label. However, our rationale was that regardless of condition, subjects need to first perceive, encode and maintain the presented stimulus before they can even commence the task, be it VWM or MR. Thus, we expected to be able to extract the neural pattern of the presented stimulus during at least the physical presentation and a brief moment after that. This classifier was then trained and applied across all time points, resulting in a temporal generalization matrix (King and Dehaene, 2014). It is important to note that we trained the classifiers only using the labels of the presented stimulus, but sorted the data in varying ways when testing the performance. For example, by looking at an early training time point, but a late testing time point, we tested whether we could decode the orientation of the grating kept in mind near the end of the delay period, on the basis of the pattern evoked by the presented stimulus early in the trial.

In the second line of analysis, we trained a continuous orientation decoder on the functional localizer. The larger number of orientations sampled in the functional localizer allowed us to decode a continuous estimate of represented orientation, rather than a discrete one from a fixed number of classes. We applied this decoder to the VWM/imagery task and subsequently related the decoded orientation to the true presented orientation by calculating a quantity intuitively similar to a correlation coefficient (see *Continuous orientation decoder*). Here too, we extended the procedure to include all pairwise training and testing time points, resulting in temporal generalization matrices (King and Dehaene, 2014).

In the control analysis, where we tested for a systematic relationship between gaze position and VWM contents, we repeated the first line of analysis described above, but

instead used the gaze position (x- and y-coordinates) as features rather than the MEG data.

Multi-class probabilistic classifier

The three-class classifier was based on Bishop (2006, p. 196-199). Briefly, the class-conditional densities were modeled as Gaussian distributions with assumed equal covariance. By means of Bayes' theorem, and assuming a flat prior, this model was inverted to yield the posterior probabilities, given the data. Specifically, let \mathbf{x} be a column vector with length equal to the number of features [number of sensors for MEG data, two for gaze position (horizontal and vertical location)] containing the data to be classified, then the posterior probability that the data belongs to class k is given by the following equations:

$$P(\text{class} = k \mid \mathbf{x}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

$$a_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + w_{k0}$$

$$\mathbf{w}_k = \mathbf{S}^{-1} \mathbf{m}_k$$

$$w_{k0} = -\frac{1}{2} \mathbf{m}_k^T \mathbf{S}^{-1} \mathbf{m}_k$$

where \mathbf{m}_k is the mean of class k and \mathbf{S} is the common covariance, both obtained from the training set. The latter was calculated as the unweighted mean of the three covariance matrices for each individual class, and subsequently regularized using shrinkage (Blankertz et al., 2011) with a regularization parameter of 0.05 for the MEG data and 0.01 for the eye-tracker analysis.

Continuous orientation decoder

The continuous orientation decoder was based on the forward-modeling approach as described in (Brouwer and Heeger, 2009, 2011) but adapted for improved performance (Kok et al., 2017). The forward model postulates that a grating with a particular orientation activates a number of hypothetical orientation channels, according to a characteristic tuning curve, that subsequently lead to the measured MEG data. We formulated a model with 24 channels spaced equally around the circle, whose tuning curves were governed by a Von Mises curve with a concentration parameter of 5. Note that all circular quantities in the analyses were multiplied by two, because the formulas we used operate on input that is periodic over a range of 360° , but grating orientation only ranges from 0° to 180° . Next, we inverted the forward model to obtain an inverse model. This model reconstructs activity of the orientation channels, given some test data. In this step we departed from Brouwer and Heeger's (2009, 2011) original formulation in two aspects. First, we estimated each channel independently from each other, allowing us to include more channels than there

are stimulus classes. Second, we explicitly took into account the correlational structure of the noise, which is a prominent characteristic of MEG data, in order to improve decoding performance (Blankertz et al., 2011; Mostert et al., 2015). For full implementational details, see (Kok et al., 2017). The decoding analysis yields a vector \mathbf{c} of length equal to number of channels (24 in our case) with the estimated channel activity in a test trial, for each pairwise training and testing time point. These channels activities were then transformed into a single orientation estimate θ by calculating the circular mean (Berens, 2009) across all the orientations the channels are tuned for, weighted by each individual activation:

$$\theta = \arg \left[\sum_j c_j \exp(i\mu_j) \right]$$

where the summation is over channels, μ_j is the orientation around which the j th channel tuning curve is centered, and i is the imaginary unit. These decoded orientations can then be related to the true orientation, across trials, as follows:

$$z = \frac{1}{N} \sum_k^N \exp[i(\theta_k - \varphi_k)]$$

$$\rho = |z| \cos(\arg(z))$$

where N is the number of trials and φ_k is the true orientation on trial k . The quantity ρ is also known as the test statistic in the V-test for circular uniformity, where the orientation under the alternative hypothesis is pre-specified (Berens, 2009). This quantity has properties that make it intuitively similar to a correlation coefficient: it is +1 when decoded and true orientations are exactly equal, -1 when they are in perfect counter-phase and 0 when there is no systematic relationship or when they are perfectly orthogonal.

Statistical testing

All inferential statistics were performed by means of a permutation test with cluster-based multiple comparisons correction (Maris and Oostenveld, 2007). These were applied to either whole temporal generalization matrices, or horizontal cross-sections thereof (i.e. a fixed training window). These matrices/cross-sections were tested against chance-level (33%) in the classification analysis, or against zero in the continuous decoding analysis. In the first step of each permutation, clusters were defined by adjacent points that crossed a threshold of $p < 0.05$ according to a two-tailed one-sample t-test. The t-values were summed within each cluster, but separately for positive and negative clusters, and the largest of these were included in the permutation distributions. A cluster in the true data was considered significant if its p-value was less than 0.05. For each test, 10,000 permutations were conducted.

Spatial patterns and source analysis

To interpret the signals that the classifier and decoder pick up, we looked at the corresponding spatial patterns (Haufe et al., 2014). The spatial pattern is the signal that would be measured if the latent variable that is being decoded is varied by one unit. For both the probabilistic classifier and the continuous orientation decoder, this comes down to the difference ERF between each category and the average across all categories. This yields one spatial pattern for each class, and these were subsequently averaged across classes, as well as across time of interest, and fed into source analysis and synthetic planar gradient transformation. This transformation refers to a procedure whereby MEG data recorded with axial gradiometers is transformed as if it were measured by planar gradiometers (Bastiaansen and Knösche, 2000). The main advantage is that the spatial distribution of the resulting data is more readily interpretable.

For source analysis, we used a template anatomical scan provided by FieldTrip to create a volume conduction model based on a single shell model of the inner surface of the skull. The source model consisted of a regular grid spaced 0.5 cm apart that encompassed the entire brain. Leadfields were calculated and rank-reduced to two dimensions, to accommodate the fact that MEG is blind to tangential sources. The covariance of the data was calculated over the window of 1 to 8 s post-stimulus and regularized using shrinkage (Blankertz et al., 2011) with a regularization parameter of 0.05. The leadfields and data covariance were then used to calculate linearly constrained minimum variance spatial filters (LCMV, also known as beamformers; Van Veen et al., 1997). Applying these filters to sensor-level data yields activity estimates of a two-dimensional dipole at each grid point. We further reduced these estimates to a scalar value by means of the Pythagorean theorem. This leads to a positivity bias however, that we corrected for using a permutation procedure (see Manahova et al., 2018, for details). The number of permutations was 10,000. The final result was interpolated to be projected on a cortical surface, and quantifies the degree to which a particular area contributed to the performance of the classifier/decoder.

Results

Behavioral results

The average accuracies in the MEG session for the four conditions ranged from 68-72%, confirming that subjects were able to do the task, as well as that the staircase procedure was successful. The average final jitter estimate from the staircase procedure for the 0°, 60°, 120° and 180° conditions were as follows (95%-CI in parentheses): 3.1° (0.98-5.2), 11.0° (8.87-13.0), 13.4° (11.35-15.52) and 6.5° (4.42-8.60), respectively. With the exception of the 180° condition, the rising trend in these values suggests that subjects found the task more difficult when the amount of rotation was larger. The relatively low value for the 180° condition however indicates that this condition was relatively easy. One explanation may be the fact that subjects did not require the final product of the mental rotation in order to perform well on the task. It is possible that they simultaneously

memorized the starting orientation. After having finished the mental rotation - regardless of how well they were able to do so - they could simply reactivate the initial image and use that in their judgment.

Gaze position tracks VWM contents

In the eye movement analysis, we investigated whether there is a relation between gaze position and the item held in VWM. We adopted the same analysis in our original main analysis (see *Methods*), but instead entered the horizontal and vertical gaze position, measured by the eye-tracker, as features in the decoding analysis. Specifically, we constructed a three-class probabilistic classifier that yields the posterior probabilities that any given data belong to either of three presented orientations. That is, the classifier was trained according to the labels of the presented stimulus. The classifier was trained on trials from all conditions (i.e., all amounts and directions of rotation) pooled together to obtain maximum sensitivity (see *Methods* for rationale). To verify whether we could decode stimulus identity from the gaze position, we first applied the classifier to the same (pooled) data using cross-validation. We found above-chance decoding in a time period of approximately 0.5-3.5 s post-stimulus that was marginally significant (Supplementary Figure S1). This indicates that subjects moved their eyes in a way consistent with the present stimulus, and kept it there for approximately 2-3 seconds.

Then, when looking at the decoded signal within the VWM condition only, we found a sustained pattern (Fig. 2A), though again only marginally significant. This suggests that upon perceiving and encoding the stimulus, subjects move their eyes in a way systematically related to the identity of the stimulus, and keep that gaze position stable throughout the entire delay period.

Contrary to previously used paradigms, where two stimuli were displayed at the beginning of a trial and a retro-cue signaled the item that was to be remembered (e.g. Harrison and Tong, 2009; Albers et al., 2013; Christophel et al., 2015, 2017), in the present experiment we only showed one stimulus. It is therefore possible that the sustained decoding performance does not necessarily reflect VWM contents, but simply that the subjects moved their gaze according to the presented stimulus rather than to their mental contents. However, if this were true, then we should find a similar effect in the three MR conditions. If, on the other hand, the classifier picked up the item kept in mind, then the probability that a trial is assigned to the same class as the presented stimulus should drop over time, as the subject rotates the mentally imagined grating away from the starting orientation. Our results were consistent with the latter scenario (Fig. 2B). Whereas the probability that the data belong to the same class as the presented stimulus stays steadily above chance in the VWM condition, it drops to lower levels in the three MR conditions. Moreover, we found evidence that the gaze moves towards a position consistent with the orientation of the presented grating plus or minus 60° (depending on the cued direction of rotation) in the MR conditions, but not any further (Fig. 3A,B).

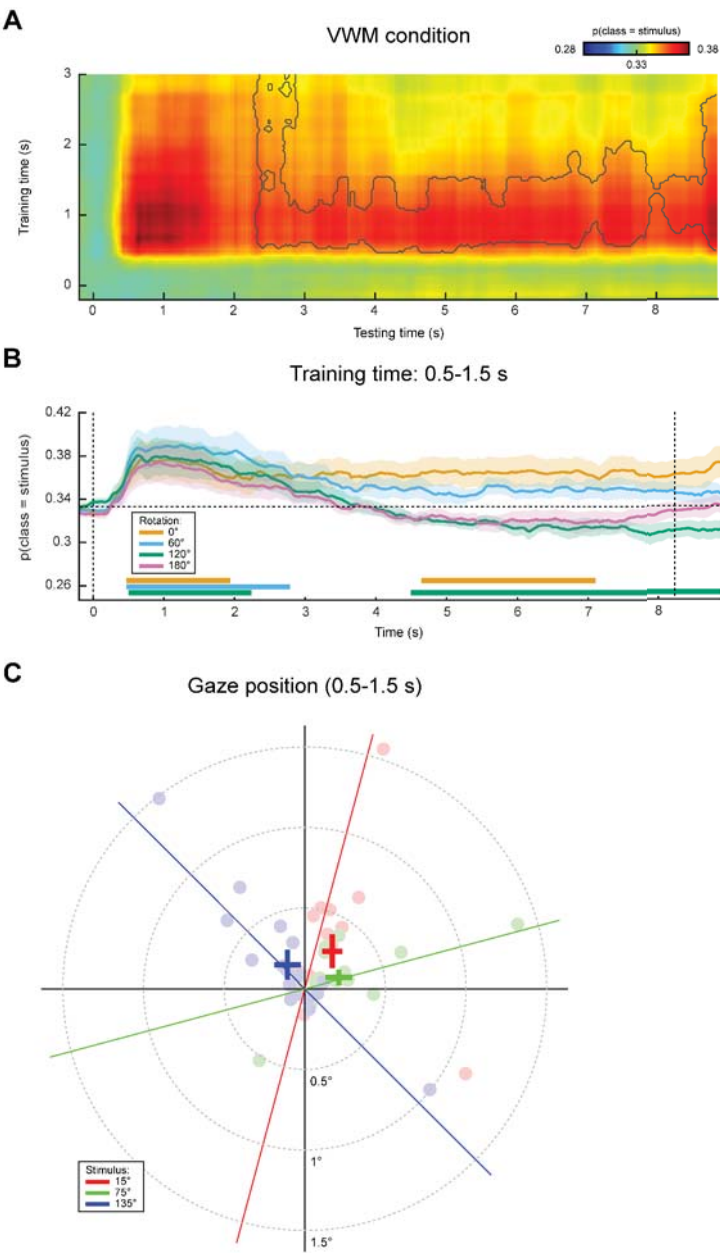


Figure 2. Gaze position classification results, cross-validation within VWM/imagery task. **(A)**, Temporal generalization matrix of classification performance in the VWM condition. The color scale denotes the average posterior probability that the data belong to the same class as the presented stimulus. The gray outline demarcates a near-significant cluster ($p = 0.069$). Note that this matrix is asymmetric because only the VWM condition is shown, while the classifier was trained on the data from all VWM/MR conditions. For this reason, the data after approximately 3 s are not expected to contain systematic patterns and therefore the training time axis has been truncated (see *Supplementary Figure S1*). **(B)**, Classification performance averaged over the training time window of 0.5-1.5 s, separately for the VWM and the three MR conditions. Note that the 0° condition corresponds directly to the matrix in **(A)**. The two vertical dashed lines indicate stimulus and probe onset. Shaded areas indicate the standard error of the mean (SEM). Significant clusters are indicated by the horizontal bars in the lower part of the figure. **(C)**, Average gaze position during 0.5-1.5 s after stimulus onset, separately per stimulus orientation. Each transparent dot corresponds to an individual subject. The crosses are the grand averages, where the vertical and horizontal arms denote the SEM. The three colored lines depict the orientation of the three stimuli.

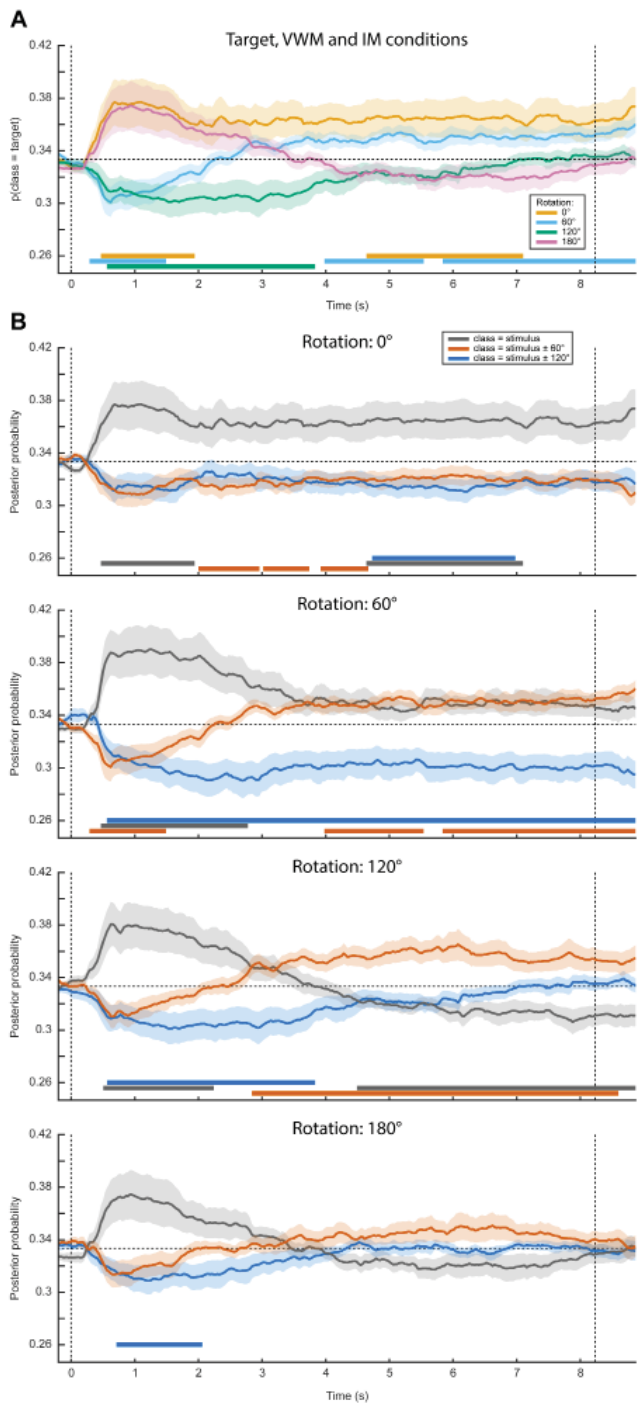


Figure 3. Complete gaze position classification results.

(Figure 3, cont.) Analogously to Fig. 2, the classifier was trained on the time window of 0.5-1.5 s post stimulus onset, and according to the labels of the presented stimulus. **A**, Average posterior probability that the data belong to the class of the target orientation, that is, the orientation that the subjects were supposed to have in mind at the end of the delay period. For both the 0° and the 180° conditions, the target orientation was the same as the presented stimulus. For the 60° and 120° conditions however, the target and presented stimulus were different, hence the below-chance probabilities at the beginning of the delay period. **B**, The average posterior probabilities that the data belong to either of three classes: the same orientation as the presented stimulus, the orientation of the presented stimulus $\pm 60^\circ$ or the orientation of the presented stimulus $\pm 120^\circ$, plotted separately for the four VWM/MR conditions. The plus/minus-sign is due to the mental rotation being performed either clockwise or counter-clockwise. This figure gives insight into whether the feature patterns corresponding to any intermediate orientations become active during mental rotation, which is particularly relevant for the 120° and 180° conditions. For instance, in the 180° counter-clockwise condition, the subject would start with a mental image with the same orientation as the presented stimulus, then pass through respectively -60° and -120°, ultimately to reach the target of -180° (i.e. 0°). If the gaze positions corresponding to all these orientations become active in sequence, one would first expect a peak in posterior probability that the data belong to the same class as the stimulus (gray line, bottom figure), then a peak in the probability of belonging to the presented stimulus -60° (orange line), then for -120° (blue line) and finally again for 0° (gray line). It is important to realize that below-chance probabilities in these analyses are meaningful. For instance consider **(A)**. Here, the probability that the data belong to the same class as the target is plotted. Hence, in the 60° and 120° conditions, the classifier correctly identifies that the data do not belong to the same class as the target in the beginning of the interval, because the starting orientation was different. As another example, consider the red line in the second panel in **(B)**. This line plots the probability that the data belong to the starting orientation $\pm 60^\circ$. Upon presentation of the starting orientation, the classifier therefore yields a significant below-chance probability. However, as the subject performs the $\pm 60^\circ$ rotation over the course of the trial, the classifier increasingly picks up this rotated image, hence giving above-chance probabilities. Note that **(A)** and **(B)**, as well as Fig. 2 all depict the same data, but visualized in different manners. Shaded areas denote the SEM and significant clusters are depicted by the thick horizontal lines at the bottom of the panels.

Fig. 2C displays the grand average, as well as individual average gaze positions during 0.5-1.5 s after stimulus onset, separately for each of the three stimulus conditions, collapsed across VWM and MR conditions. Although there is large variability among subjects in the magnitude of the eye movements, a general trend can be discerned where subjects position their gaze along the orientation axis of the grating. The mean disparity in visual angle with respect to pre-trial fixation was only 0.23° , which is in the same order of magnitude as reported previously (Foster et al., 2016), though for some subjects it was larger, up to 1.5° .

In short, there was a systematic relationship between gaze position and stimulus orientation, after which the gaze position tracked the orientation kept in mind during the delay period, but only for a maximum of approximately $\pm 60^\circ$ relative to starting orientation. These findings raise the concern that any potential decoding of VWM items from MEG signals, as was the aim of our original analysis, could be the result of stimulus-related eye confounds (see *Discussion* for possible underlying mechanisms).

Sustained decoding of VWM items from MEG signals

The original aim of this study was to assess the representational contents of the neural signals while the subjects were engaged in VWM/imagery. We present these results here, to demonstrate how they could easily have been mistaken for genuine results, had we been oblivious to the systematic eye movements. We constructed a three-class probabilistic classifier in which the MEG sensors were entered as features. As before, the classifier was trained on trials from all conditions (i.e., all amounts and directions of rotation) pooled together for maximum sensitivity (see *Methods* for rationale). To test whether we could decode stimulus identity from the MEG signal, we applied the classifier to the same (pooled) data using cross-validation, and found successful decoding during a period of up to approximately 2.5 seconds after stimulus onset (Supplementary Figure S2). The stimulus itself was presented for only 250 ms. Therefore the later part of this period could have been interpreted as an endogenous representation, for instance stemming from active mental instantiation by the subject, although in reality it is more likely to be the result of eye movements.

Next we looked at the decoding performance in the VWM condition alone, using the classifier trained on all conditions as described above. We found that the identity of the memory item could be decoded during the entire delay interval, using classifiers obtained from a training window of approximately 0.5-1.5 s (Fig. 4A). The performance stayed above chance-level (33.3%) at a stable level of $\sim 37\%$ throughout the entire interval (Fig. 4B).

Again, we found this sustained pattern to be specific to the VWM condition, because the probability that the data belong to the same class as the presented stimulus drops over time in the three MR conditions (Fig. 4B). As explained in the previous paragraph, this indicates that the sustained above-chance classification in the VWM condition cannot

be explained as a long-lasting stimulus-driven effect (e.g. stimulus aftereffect), but must also reflect the memorized item to at least some degree. In the 180° condition, the posterior probability that the data belong to the same classes as the presented stimulus later reemerges as a rising, though non-significant trend. This can be explained by the fact that the final orientation that the subjects should have in mind in the 180° condition is identical to the orientation of the presented stimulus at the start of a trial.

In order to facilitate interpretation of these results, we inspected the classifier's corresponding sensor topography (Fig. 4C) and source localization (Fig. 4D), averaged over the training time period of 0.5-1.5 s. These indicate that both occipital (Harrison and Tong, 2009; Albers et al., 2013) and prefrontal sources (Spaak et al., 2017; Sreenivasan et al. 2014) contributed to the classifier's performance. Indeed, the prefrontal sources could in reality point to ocular sources. Moreover, although the contribution from occipital regions may seem to provide evidence that the decoder genuinely picks up visual representations,

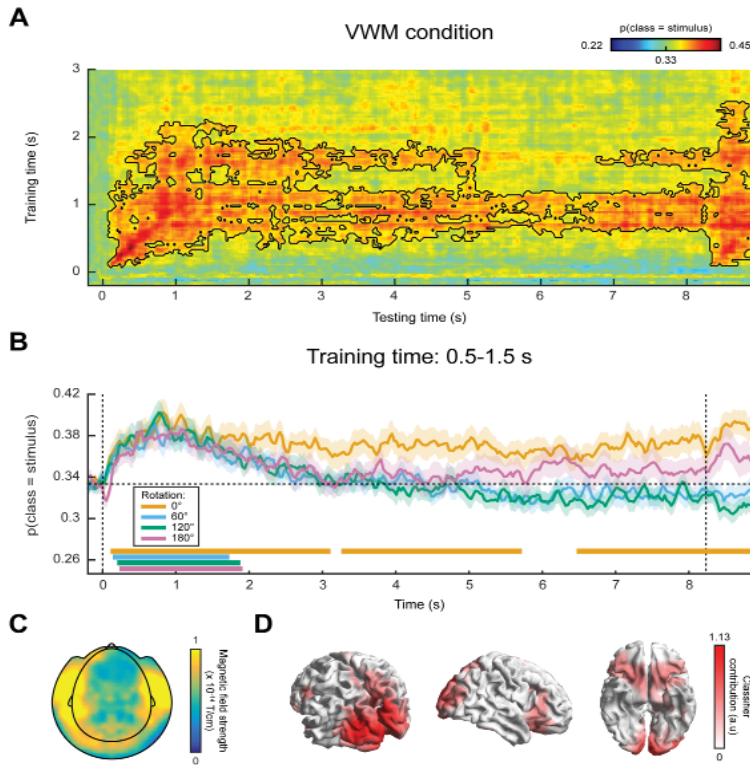


Figure 4. MEG classification results. **(A)** Same as in Fig. 2A, except the classification was performed on MEG data, rather than on gaze position. The black outline demarcates a significant cluster ($p = 0.006$). **(B)** Same as in Fig. 2B, except the analysis was performed on MEG data. **(C)** Synthetic planar gradiometer topography and **(D)** source topography of areas that contributed to the classifier. See also Supplementary Figure S2.

these sources could in fact also be driven by the eye movements (see *Discussion*). Finally, we investigated whether we could decode the intermediate (for the 120° and 180° rotations) and the final orientations in the MR conditions. We found some indication that the final orientation - but not the intermediate ones (Fig. 5B) - indeed emerges halfway through the delay period, but this effect was not statistically significant (Fig. 5A).

Decoding visual representations from sensory areas

In a third analysis, we trained a continuous orientation decoder (see *Methods*) on the functional localizer data (Fig. 1B, C) and applied these decoders to the data from the VWM/imagery task (King and Dehaene, 2014). The main advantage of this method is that it ensures that the decoder is primarily sensitive to sensory signals, and not to higher-level top-down processes involved in mental manipulation of an image. It thus allows us to track sensory-specific activation throughout the delay period (Mostert et al., 2015). Cross-validation within the functional localizer confirmed that we were indeed able to reliably decode orientation-specific information from activity evoked by passively perceived gratings (Supplementary Figure S3). Moreover, we were not able to decode grating orientation on the basis of gaze position, verifying that the data from the functional localizer were not contaminated by stimulus-specific eye movements (Supplementary Figure S4).

In the VWM condition, the decoders trained on the functional localizer data could reliably decode the orientation of the presented stimulus (Fig. 6A). Moreover, for a training time of approximately 90-120 ms, we could decode the stimulus for a prolonged time, lasting over 1 s after stimulus onset. Interestingly, this training time point coincides with the time at which peak performance is obtained within the localizer itself (Supplementary Figure S3). Comparing the decoding trace within this training time window with the three MR conditions, it can be seen that grating orientation can be decoded in all four conditions for a sustained period of approximately 500 ms (Fig. 6B). This is in sharp contrast to the extended decoding of the presented stimulus throughout the entire delay period, found within the VWM/imagery task using cross-validation (Fig. 4).

We inspected the spatial pattern (Fig. 6C) and corresponding source topography (Fig. 6D) for the decoders, averaged across training time 90-120 ms. These highlight primarily occipital regions as contributing to the decoder's performance, consistent with our premise that the functional localizer primarily induced bottom-up sensory signals, especially during this early time interval (Mostert et al., 2015).

In summary, our findings suggest that the stable, persistent representation found in our within-task MEG decoding result may well be attributed to stimulus-specific eye movements, even though the magnitude of the eye movements were only small. In contrast, no clear evidence was found for such a long-lasting representation when training the decoder on the functional localizer. Given that the localizer was not contaminated by stimulus-specific eye movements, these results thus provide a more reliable picture of the sensory representations during the delay interval.

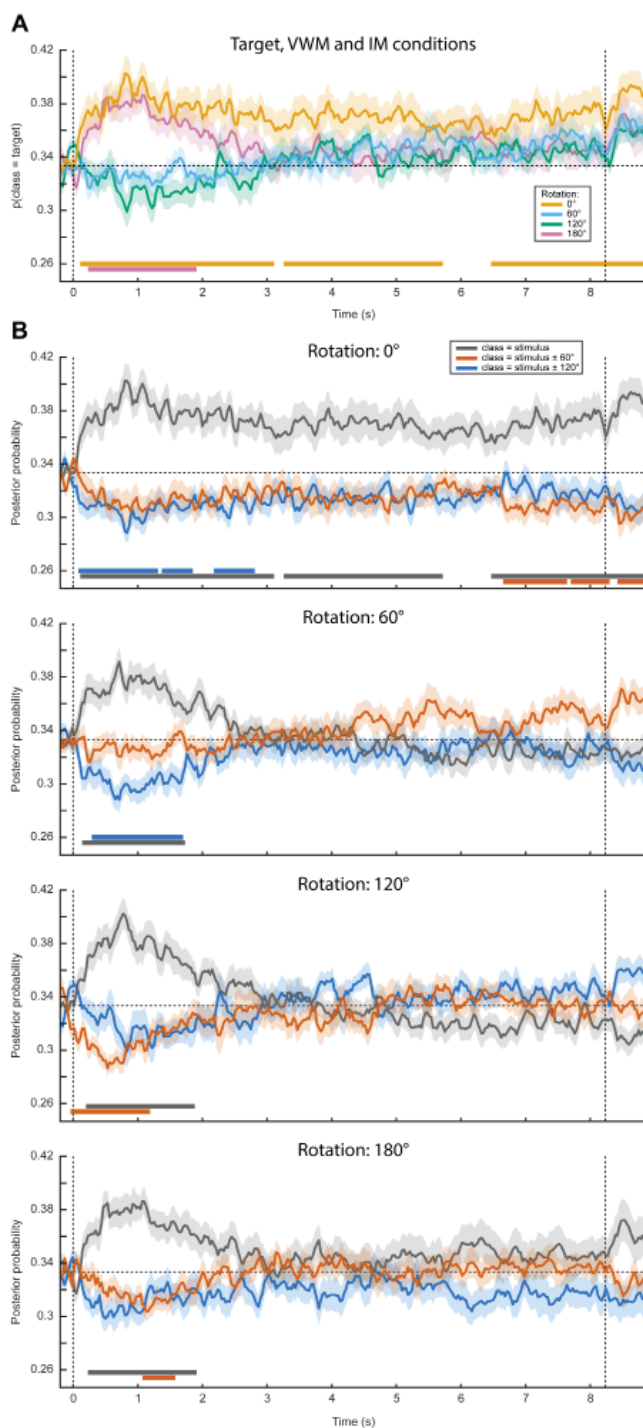


Figure 5. Complete MEG classification results, visualized in a variety of ways. This figure is analogous to Fig. 3, except the classifier is trained and tested on MEG data rather than on gaze position.

Discussion

Neural decoding is a powerful and promising technique for neuroimaging studies (Haxby et al., 2014; Grootswagers et al., 2016) that has led to substantial advancement in the field of VWM over recent years (Harrison and Tong, 2009; Serences et al., 2009; Albers et al., 2013; Wolff et al., 2015, 2017). This study also employed neural decoding techniques, with the original aim to investigate the temporal dynamics of sensory representations during VWM. However, we found that our data were contaminated by small but systematic eye movements, whereby gaze position was related to the stimulus held in mind. This jeopardized our ability to interpret the neural decoding results - results that otherwise would have seemed sensible - because very similar results could be obtained by considering gaze position only.

There are at least three possible mechanisms via which stimulus-specific eye movements may confound our results. First, eye movements are known to cause stereotypical artifacts in MEG recordings. Due to the positively charged cornea and negatively charged retina, the eyeball acts as an electromagnetic dipole whose rotation is picked up by the MEG sensors. The spatial pattern that the dipole evokes is directly related to its angle, or in other words, to the position of the subject's gaze (Plöchl et al., 2012). Thus, if the subject moves his/her eyes in response to the grating in a manner related to the orientation of that grating, then this will induce a specific pattern in the MEG signals, which in turn is directly related to the grating orientation. A decoding analysis applied to these signals is then likely to pick up the patterns evoked by the eyeball dipoles, confounding potential orientation-related information stemming from genuine neural sources. In fact, our source analysis hints at this scenario (Fig. 4D), as the contributions from presumed prefrontal sources closely resemble an ocular source.

Second, if the eyes move, then the projection falling on the retina will also change, even when external visual stimulation remains identical. Thus, if gaze position is systematically modulated by the image that is perceived or kept in mind, then so is the visual information transmitted to the visual cortex. For example, if a vertical grating is presented and kept in VWM, then the subject may subtly move her or his gaze upward. Correspondingly, the fixation dot is now slightly below fixation, thus leading to visual cortex activity that is directly related to the retinotopic position of the fixation dot. Our decoding analysis may thus actually decode the position of the fixation dot, rather than grating orientation, potentially leading to an incorrect conclusion. Source analysis would in this scenario also point to occipital sources, similarly to what we found (Fig. 4D). Note that this mechanism is not specifically dependent on the presence of a fixation dot. A systematic difference in eye position will also lead to changes in the retinotopic position of, for instance, the presentation display or the optically visible part of the MEG helmet.

Third, if gaze position covaries with the mental image, then decoding of the mental image will also reveal areas that encode eye gaze position, such as oculomotor regions in parietal and prefrontal cortex.

Our findings raise the question of why there were task-induced eye movements that were directly related to the grating kept in VWM. In fact, there is a considerable mass of literature that describes the role of eye movements in mental imagery. It has been found that subjects tend to make similar eye movements during imagery as during perception of the same stimulus (Brandt and Stark, 1997; Laeng and Teodorescu, 2002; Laeng et al., 2014). Already proposed by Donald Hebb (Hebb, 1968), it is now thought that eye movements serve to guide the mental reconstruction of an imagined stimulus, possibly by dwelling on salient parts of the image (Spivey and Geng, 2001; Laeng et al., 2014). Moreover, the specificity of the eye movements is also related to neural reactivation (Bone et al., 2017) and recall accuracy (Laeng and Teodorescu, 2002; Laeng et al., 2014; Bone et al., 2017). Our findings are in accordance with these studies. Subjects' gaze was positioned along the orientation axis of the grating - that is, the visual location within the stimulus that provided the highest information regarding its orientation, and is thus exactly what one would expect given that the task was to make a fine-grained orientation comparison with a probe grating. Importantly however, subjects were explicitly instructed to maintain fixation throughout the entire trial. We nevertheless observed that not all subjects adhered to this requirement, albeit involuntarily.

Despite these problems associated with the systematic eye movements in our experiment, it is still possible that our decoding results do in reality stem from genuine orientation information encoded in true neural sources. In fact, we used independent component analysis in our pre-processing pipeline to (presumably) remove eye-movement artifacts. However, it would be very difficult, if not impossible, to convincingly establish that no artifacts remain and, considering the similarities between the decoding results from the MEG data (Fig. 4A,B) and the gaze position (Fig. 2A,B), we feel any attempts at this would be unwarranted.

Given the potential pervasiveness of systematic eye movements in VWM/imagery tasks, and the demonstrated susceptibility of our analysis methods to these confounds, one wonders whether other studies may have been similarly affected. Clearly, the first mechanism described above involving the eyeball dipole would only affect electrophysiological measurements like electroencephalography and MEG, and has indeed been a concern in practice (Foster et al., 2016). The second mechanism however, whereby stimulus identity is confounded with the retinal position of visual input, would also affect other neuroimaging techniques such as fMRI. This confound could be particularly difficult to recognize, because it would also affect activity in visual areas. Moreover, because eye movements during imagery have been found to be positively related to performance (Laeng and Teodorescu, 2002), this could potentially explain correlations between VWM decoding and behavioral performance. The third mechanism, whereby one directly decodes gaze position from motor areas, could be a problem especially for fMRI which, thanks to its high spatial resolution, might be well able to decode such subtle neural signals. This concern may be especially relevant for studies that investigate the role of areas involved in eye movements or planning thereof, such as frontal eye fields or superior

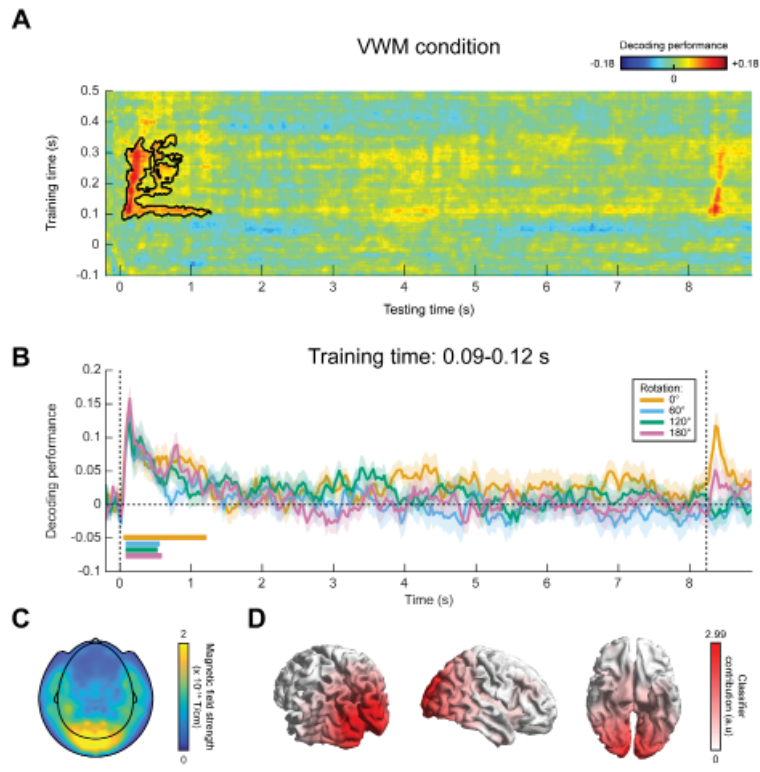


Figure 6. MEG decoding results, generalized from localizer to VWM/imagery task. **(A)** Temporal generalization matrix of orientation decoding performance, for which the decoder was trained on all time points in the functional localizer and tested across all time points in the VWM/imagery task. The color scale reflects the correspondence between true and decoded orientation. The black outline shows a significant cluster ($p = 0.04$). Note that the x- and y-axis in the figure are differently scaled for optimal visualization. **(B)** Decoding performance over time in the VWM/imagery task, averaged over decoders trained in the window of 0.09-0.12 s in the localizer, separately for the VWM and three MR conditions. Shaded areas denote the SEM and significant clusters are indicated by the horizontal bars. **(C)** Synthetic planar gradiometer topography and **(D)** source topography of areas that contribute to the decoder. See also Supplementary Figure S3 and S4.

precentral sulcus, in the maintenance of working memory items (Jerde et al., 2012; Ester et al., 2015; Christophel et al., 2017).

This leaves the question of how to deal with eye movements in VWM/imagery tasks. Naturally, it is important to record eye movements during the experiment, for instance using an eye-tracker or electrooculogram (EOG). One can then test for any systematic relationship and, if found, investigate whether it could confound the main results. In

our case, for example, decoding of gaze position leads to strikingly similar results as those obtained from the MEG data. Foster et al. (2016) on the other hand found that decoding performance of working memory items decreased throughout the trial, whereas the deviation in gaze position increased, suggesting that eye confounds cannot explain the main findings. Another approach might be to design the experimental task in such a way that eye movements are less likely. For example, by presenting gratings laterally (e.g. Pratte and Tong, 2014; Ester et al., 2015; Wolff et al., 2017), and assuming that VWM items are stored in a retinotopically specific manner (Pratte and Tong, 2014), the involuntary tendency to move one's eyes subtly along the remembered grating's orientation axis may become less strong, because those gratings are located distantly from the gaze's initial location (i.e. central fixation). Finally, a powerful approach could be to adopt a separate functional localizer, which allows specific decoding of functionally defined representations such as bottom-up, sensory-specific signals (Harrison and Tong, 2009; Serences et al., 2009; Albers et al., 2013; Mostert et al., 2015). If the localizer is well designed and not systematically contaminated by eye movements, then eye movements in the main task cannot have a systematic effect on the decoded signal, thus effectively filtering them out.

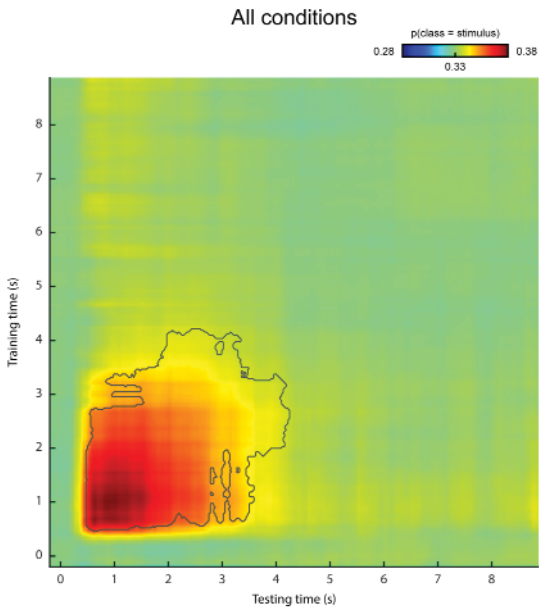
We designed a localizer that is specifically sensitive to the neural representations encoded in bottom-up signals evoked by passively perceived gratings. This allowed us to address the question of whether the imagined stimulus was encoded with a similar neural code as the perceived gratings (Harrison and Tong, 2009; Albers et al., 2013). Using this localizer, we indeed obtained MEG decoding results that were very dissimilar from those obtained using cross-validation within the combined VWM/imagery task. We no longer found persistent activation of an orientation-specific representation throughout the entire delay period. Nevertheless, the sensory pattern did remain above baseline for a period of approximately 1 second, which is relatively long considering that the stimulus was presented for only 250 ms. One explanation is that the stimulus was relevant for the task. Previous work has shown that task relevance may keep the sensory representation online for a prolonged period even after the stimulus is no longer on the screen (Mostert et al., 2015).

It should be pointed out that using a functional localizer also has its intrinsic limitations. The most important being that, while such an approach is primarily sensitive to a functionally defined signal, it may at the same time be blind to other relevant signals that were not a priori included in the functional definition. The VWM literature itself provides an instructive example: while the functional localizer approach has clearly demonstrated sensory representations of the memorandum in associated sensory cortex (Harrison and Tong, 2009; Albers et al., 2013), it would have missed relevant encoding in other regions in the brain such as parietal and prefrontal cortex (Christophel et al., 2015, 2017). Furthermore, the fact that the decoders were trained on a functionally defined signal does not mean that they are necessarily *insensitive* to other signals, such as eye movement-related signals. However, it is important to realize that the exact effect of these

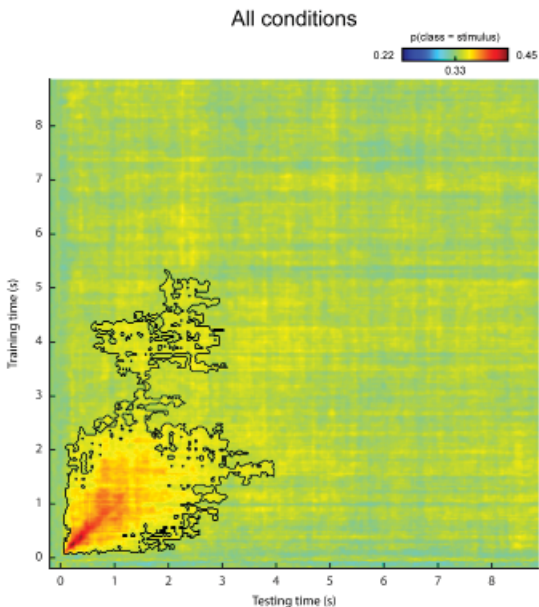
other signals on the decoder's output is not explicitly defined. These potential effects would therefore be idiosyncratic to an individual's data and are expected to cancel out at the group level.

In summary, we demonstrate a case where decoding analyses in a VWM/imagery task are heavily confounded by systematic eye movements. Given the high potential benefit of decoding analyses and its widespread use in the study of working memory and mental imagery, we argue that this problem may be more pervasive than is commonly appreciated. Future studies could target this question specifically, and investigate how strong the confounds are exactly. One approach could be to systematically vary salient input and assess how this impacts decoding performance (cf. the second mechanism described above). Furthermore, it is important to realize that this does not necessarily invalidate all previous studies. While some previous results may have been afflicted, our current understanding of the neural underpinnings of VWM is still firmly grounded in converging evidence from a wide variety of techniques, paradigms and modalities. Nevertheless, we conclude that eye movement confounds should be taken seriously in both the design as well as the analysis phase of future studies.

Supplementary Figures

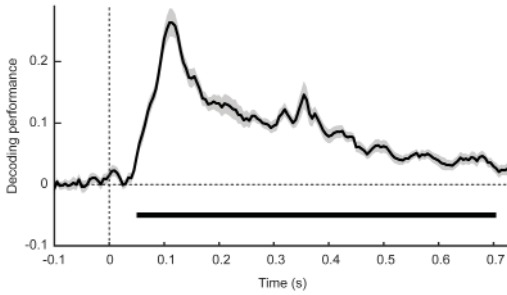


Supplementary Figure S1. Gaze position classification performance within the VWM/imagery task, pooled across VWM and MR conditions. Temporal generalization matrix of the average posterior probability that the data belong to the same class as the presented stimulus, pooled across all rotation conditions (see *Methods*). Note that the relatively short-term classification of approximately 3 seconds is expected, because the subjects rotate their mental image in clockwise direction on some of the trials and in counter-clockwise on others. Hence any reliable relation between the presented stimulus (the factor that the classifier was tested and trained on) and the mental image cancels out over the course of the delay interval. The gray outline corresponds to a near-significant cluster ($p = 0.068$).

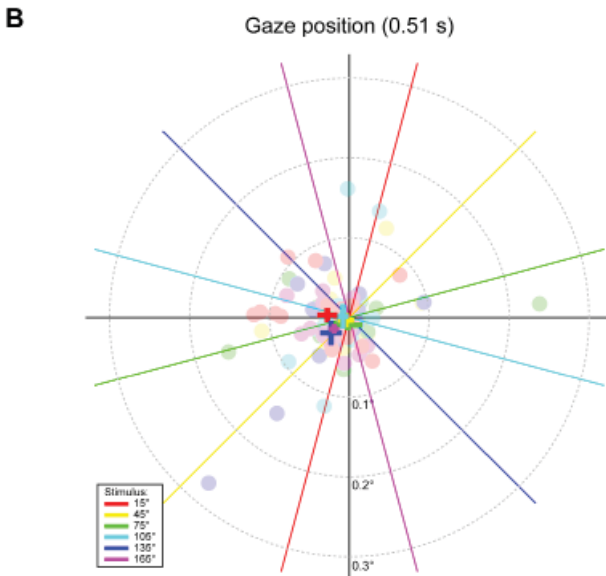
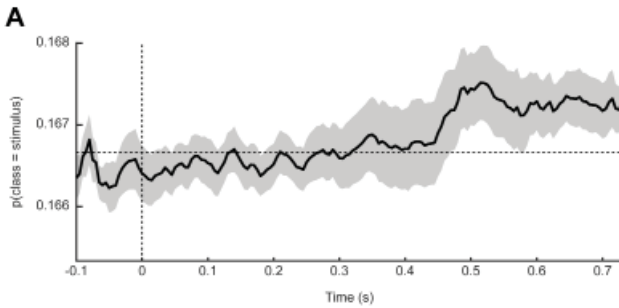


Supplementary Figure S2. MEG classification performance within the VWM/imagery task, pooled across VWM and MR conditions. Similar to Supplementary Figure S1, except the classifier is trained and tested on MEG data rather than on gaze position. Black outline indicates a significant cluster ($p = 0.007$).

Functional localizer



Supplementary Figure S3. MEG decoding performance within the functional localizer using cross-validation. The time axis represents matched training and testing time-points. Shaded areas denote the SEM, and the horizontal line demarcates a significant cluster ($p = \pm 0$).



Supplementary Figure S4. Gaze position classification performance within the functional localizer, using cross-validation. **(A)**, The time axis represents matched training and testing time-points. No significant above-chance classification was found. Note that although there appears to be a rise in performance after approximately 500 ms, this only reached a value of 0.1675 at its peak ($t = 0.51$ s), which is very little above chance (0.1667 for six classes). Shaded areas denote SEM. **(B)**, Average gaze position at 0.51 s after stimulus onset, separately per stimulus orientation. Each translucent dot corresponds to an individual participant. The crosses are the grand averages, where the vertical and horizontal arms denote the SEM. The six colored lines depict the orientation of the six stimuli.



6

General discussion

In my thesis I focused on how visual information is encoded in the brain at a relatively early stage - the sensory representation - and how this is modulated under top-down influences such as decision making, perceptual expectations and working memory. I made use of magnetoencephalography (MEG), together with multivariate decoding analyses, to track the representational content of the neural signal with high temporal resolution.

Summary of the findings

In **Chapter 2** I investigated how a visual stimulus is encoded when it is used for a perceptual judgment. In particular, I was interested in the errors: if a subject failed to see a stimulus, is this because visual areas failed to encode the stimulus or due to faulty decision making at a later stage? I trained multivariate decoders on a separate functional localizer to extract the neural patterns corresponding to passively perceived gratings, and applied these to the perceptual decision making data. The results clearly demonstrated that the stimulus's sensory representation - or the absence thereof - was veridically encoded in the brain, even when the ultimate perceptual decision was incorrect. Moreover, I found that the early part of the sensory representation was actively stabilized and maintained by the brain when it was required for a subsequent decision, in contrast to when the stimulus was only passively viewed.

Chapter 3 describes a perceptual expectation experiment in which we investigated how a predictive auditory cue influences the processing of a subsequent visually presented grating. Capitalizing on the temporal fidelity of MEG to dissociate pre-stimulus signals from post-stimulus signals, we found that the brain instantiates a sensory template of the expected grating already before the actual grating is presented. We interpreted this effect as a preemptive baseline activity shift in neurons encoding for the grating, in order to facilitate subsequent processing. This shift extended into the post-stimulus period, remaining present even well after stimulus offset. The presence of these pre-stimulus sensory templates was not dependent on whether the grating orientation was task-relevant or not, suggesting a relatively automatic prediction mechanism. Moreover, this neural modulation was found to covary with behavior: subjects whose neural data displayed a stronger pre-stimulus template experienced a larger effect of cue validity on a fine orientation-discrimination task.

In **Chapter 4** I focused on a number of open questions regarding the way perceptual expectations modulate neural sensory processing. I adopted a statistical learning paradigm, with visually presented object stimuli, that included a neutral condition in addition to an expected and an unexpected condition. Surprisingly, I did not find any effect of expectation on the magnitude of neural activity, nor on its representational contents. This is striking given the vast body of literature that has demonstrated expectation suppression before, and begs the question under which conditions expectation suppression occurs exactly. It is widely believed that identifying and capitalizing on statistical regularities in the external world is an important aspect of cortical functioning. Thus, exploring the

boundary conditions under which these effects do not occur provides constraints for the development of contemporary theories.

Lastly, in **Chapter 5** we aimed to gauge the evolution of sensory representations while subjects are keeping a grating vividly active in visual working memory for a brief period of time. Disconcertingly however, control analyses revealed that subjects systematically moved their eyes in a way related to the grating they were keeping in mind. This led to systematic eye-movement related differences in the neural signal which were likely picked up by the sensitive neural decoding analyses. As a consequence, I could not reliably interpret the results yielded by the originally planned analysis. The study does however provide a useful cautionary tale, namely to seriously consider the possibility of eye movements during all of the design, acquisition and analysis phases. As a potential countermeasure, I demonstrated that generalization from a functional localizer was not obviously contaminated by the eye movements. These results showed that upon presentation of the to-be-remembered grating, its corresponding sensory representation remained active for an extended period of approximately one second.

Top-down activation of sensory representations

A common finding reported in this thesis is the active top-down instantiation and/or stabilization of stimulus-specific sensory representations, even when the corresponding stimulus was not physically present. A stable sensory representation of a 50 ms stimulus was maintained for approximately 400 ms when that stimulus was used in a perceptual decision, but not when it was passively viewed (chapter 2); a sensory template of an expected grating was evoked even before the grating was actually presented (chapter 3) and a memorized item could be decoded from neural signals for up to 1 second (chapter 5).

This could point at a potentially generic functional mechanism in the brain: the ability to internally activate sensory representations when required for perception and/or task demands. In perceptual decision making, this mechanism may allow the integration of the decision variable to continue, even when the original source of information is no longer available (Ratcliff and Rouder, 2000; Ratcliff and McKoon, 2007). For perceptual expectations, forming a sensory template of the predicted stimulus may act as a baseline shift in the activity of neurons tuned for the upcoming stimulus, such that subsequent processing of that stimulus is facilitated (Kok et al., 2014). Other examples where the brain is able to actively instantiate and maintain internal representations are visual working memory (Harrison and Tong, 2009; Albers et al., 2013) and mental imagery (Albers et al., 2013). Indeed, I found that memorizing a presented item led to its representation being sustained for up to a second (chapter 5).

The results show that the brain is able to intrinsically generate sensory representations when required for optimal performance, across a range of tasks. Note that performance

may be defined in a broad sense, not necessarily *behaviorally*. Specifically, in chapter 3 we found that the preemptive activation of a sensory template was evoked regardless of whether the predicted stimulus feature was relevant for the task at hand. This evocation would not have been required if the brain only generated such templates when relevant for making the correct behavioral decision. This can be explained if we expand the notion of optimal performance to include optimal *perceptual* performance. It has been theorized that one of the primary mechanisms by which sensory cortex operates is to predict its input (Rao and Ballard, 1999; Lee and Mumford, 2003; Bar, 2004; Kersten et al., 2004; Friston, 2005; Summerfield and de Lange, 2014). According to the predictive coding theory, sensory areas attempt to explain away their input by inferring the external cause that caused this input (Rao and Ballard, 1999; Friston, 2005; Bogacz, 2017). This inference is optimized by taking into account statistical regularities, or perceptual priors (Fiser et al., 2010; Girshick et al., 2011). Thus, instantiating a sensory template while the predicted feature is not task-relevant may facilitate optimal performance in terms of sensory processing, rather than in explicit behavioral reporting.

How is this top-down instantiating of a sensory template implemented at the neural level? The ability to internally generate a representation fits naturally in a hierarchical generative model (Clark, 2013). Such a hierarchical model consists of several layers of functionally identical modules that collectively encode for a hierarchy of external causes (Rao and Ballard, 1999; Friston, 2005; Bogacz, 2017). Each layer sends its current prediction of the state of the external world via feedback connections to the layer below, while at the same time sending prediction errors to the layer above. Crucially, the generative part refers to the ability of layers in the model to generate the lower-level features that are expected to be caused by a given external event. Thus, internally evoked sensory representations may be activated via feedback connections from superior, predicting layers.

This begs the question: how is it determined which representation is activated, and at what moment? This involves learning and/or high-level cognitive control. In the case of perceptual expectations, where representations are activated relatively automatically without deliberate control, a potentially relevant system is the hippocampal complex in the medial temporal lobe. These areas rapidly encode statistical relations between presented stimuli, even when these relations are unknown to the observer (Schapiro et al., 2012). Moreover, the hippocampus has been implicated in reinstatement of representations in early visual cortex (Bosch et al., 2014; Hindy et al., 2016). Different neural systems may underlie top-down activation in the case where deliberate control is involved, as in visual working memory, imagery and overt perceptual decision making. Possible candidates are high-level associative areas such as prefrontal cortex and parietal cortex. In visual working memory, for instance, it is believed that sustained firing of prefrontal neurons encode goal-directed variables (Sreenivasan et al., 2014) that may drive the high-fidelity representations in sensory cortex (Harrison and Tong, 2009; Albers et al., 2013) in a top-down manner. Another intriguing neural system that may underlie the top-down instantiation of low-level representations are the basal ganglia, in cooperation with

prefrontal cortex and the thalamus. This system has been proposed as the neural system underlying the “central executive” (Hazy et al., 2007), by implementing a dynamic switchboard between cortical areas (Cohen and Frank, 2009; Stocco et al., 2010). The switching is achieved by conditional routing of information via the thalamus. Indeed, empirical evidence for a general information regulatory function of the thalamus, in particular the pulvinar, has been obtained recently (Saalmann et al., 2012).

In summary, the brain is able to internally activate sensory representations to support behavioral and perceptual performance. While the exact neural systems recruited may vary across tasks, the common denominator is that the system is able to intrinsically generate sensory features *de novo*. This fits well in a hierarchical generative modeling framework such as the predictive coding theory. Finally, it is worth mentioning that the top-down generative mechanism may also be relevant for domains other than perception, for instance motor execution. It has been proposed that the predictive coding theory, or the free-energy framework, may provide a unified theory for cortical functioning, in which motor commands are executed by generating predictions of the state in which the agent’s effectors should be (Friston, 2010; Clark, 2013). It is conceivable that the top-down activation mechanism described in this thesis may be generalized to the motor domain, where it is a way to issue motor commands down the hierarchy.

The sensory representation

Although 19th century’s phrenology has long been abandoned, localization of mental functions to specific brain areas has remained of keen interest in cognitive neuroscience. In visual cortex, a large number of low-level elementary features as well as high-level complex features have been associated with distinct visual regions. A famous example is the work by Hubel and Wiesel (1959, 1962), who showed that neurons in cat primary visual cortex respond primarily to lines of particular orientation. Other well-known examples include the fusiform face area’s (FFA) sensitivity to faces (Kanwisher and Yovel, 2006), the lateral occipital complex for objects (Grill-Spector et al., 2000), the middle temporal area for motion (Gold and Shadlen, 2007) and color in V4 (Brouwer and Heeger, 2009). The specific sensitivity of these neurons pertains to their tuning: the part of stimulus space which causes a neural response. I referred to the sensory representation as the characteristic activity pattern evoked by a given stimulus, which depends on the collective tuning of the brain. This definition may not be so trivial and unambiguous, however.

Firstly, a neuron’s tuning is in principle defined as its response to an external stimulus. However, the brain is a large network of interconnected neurons and a neuron’s response is rarely determined by the stimulus alone. Thus, mapping out a neuron’s receptive field by recording its response while testing a variety of stimulation will not yield necessarily the neuron’s idealized tuning properties, but will yield its tuning properties *in the context of the network* that it lives in. An example can be drawn from the literature on end-stopping

(Rao and Ballard, 1999). When presenting a preferred stimulus - i.e. an oriented line to which the recorded neuron is most sensitive - a V1 neuron's response also depends on stimulation that falls outside of the neuron's spatial receptive field. This poses a conundrum: when mapping out this neuron's receptive field, should we take into account such contextual effects? While these effects are relatively localized to the visual cortex, more global network-contextual effects play a role too. As described in Clark (2013), if the brain indeed is a prediction machine that continuously updates its internal model of the outside world, then any neuron's response must be considered in the light of that internal model. Indeed, it is a well-documented phenomenon that an identical stimulus may lead to different activity, depending on the temporal and/or spatial context (Bar, 2004; Summerfield and de Lange, 2014). Another factor that determines the network context is the behavioral task. Consider a perceptual decision making task, in which the subject presses a button according to which stimulus was shown. If we follow our definition of the sensory representation as the ensemble of neural activity evoked by the stimulus, then we would include decision areas and even motor areas as well. Clearly, this is undesirable, because these areas also respond to a wider array of sensory input than just visual stimuli. The crux is that a neuron's response to a given stimulus is codetermined by the network configuration - in this case a configuration that allows for performance on the decision making task. So, in practice, the sensory representation (as well as tuning properties) may not be so simple to identify. Ideally, we would want to study a neuron's response to bottom-up stimulation only, in the absence of lateral and top-down connections. In practice however, it's more feasible to study a neuron's behavior while attempting to keep the network's context as constant and neutral as possible. One future approach to study neuron's tuning properties more systematically is on the basis of simulations, such as with large-scale deep neural networks (Güçlü and van Gerven, 2015), because this would yield a fine-grained view of, or even control over, the context of the simulated network.

Secondly, an important issue to take into account when defining the sensory representation of a given stimulus is the *format* in which the stimulus is encoded. More specifically, if we follow our definition above and define the sensory representation as the neural pattern evoked by a stimulus, say a face, then we will find that the sensory representation not only includes the FFA, but also upstream areas such as V1 and the lateral geniculate nucleus, or even the retina. Technically this makes sense - early areas necessarily need to encode the stimulus, for else the more downstream, specialized areas wouldn't be able to encode the stimulus either. However, one could argue that a more desirable definition of the sensory representation would only include the area that specifically codes for the stimulus under question, which in the case of faces would arguably be the FFA. While the retina and V1 and may contain all the information that is required to identify the stimulus as a face, it also contains much more information. The central point in this dilemma is the format in which information is encoded. While an early area such as V1 encodes virtually all visual information, it does so in a different format than higher-level areas such as FFA. An empirical example of this scenario is described by Brouwer and Heeger (2009), who found that while color can be decoded from (among others) V1 and V4, it was in V4 that

the encoding format aligned with perceptual experience. Thus, whether or not a given area should be included in an appropriate definition of a stimulus' sensory representation depends on the format of the representation one is interested in.

In this thesis I chose a pragmatic solution, whereby I operationalized the sensory representation as the neural pattern evoked by passively perceived stimuli, while subject's attention was narrowly drawn to the fixation dot. This way, I aimed to keep the network, or overall brain state, as constant as possible. I did not however take into account the specific format of the encoding, and hence our sensory representations could include areas from all levels across the visual hierarchy.

Functional localizer and between-task generalization

Throughout this entire thesis, I made use of functional localizer blocks in which participants passively viewed the stimuli of interest, while having their attention drawn to a task at fixation. This approach allowed me to identify the neural patterns specific to those stimuli - i.e. the sensory representation - and subsequently trace this pattern during different blocks in which participants performed a task specific to the research question (King and Dehaene, 2014).

The use of functional localizers in conjunction with decoding analyses is not always common practice in the literature. Researchers often make use of cross-validation to perform multivariate decoding within one data set. A major downside of this approach is that it's more difficult to understand what exactly is being decoded. It should be realized that decoding results obtained within one and the same data set using cross-validation is equivalent to a multivariate omnibus test. That is, above-chance decoding performance simply tells us that there is a difference between two conditions - it does not necessarily tell us anything about *what* this difference is. A clear example is provided in chapter 5, where we found that within-task decoding of the combined visual working memory and imagery task led to significant decoding of the memorized item. Naively, this could have been interpreted as sustained activation of the stimulus's representation throughout the delay interval. However, we showed that this effect was likely caused by systematic eye movements. The problem is that the within-task approach is sensitive to any difference between the conditions, including trivial differences and differences one would not have thought of *a priori* (e.g., idiosyncratic, observer-specific differences in salience between the stimuli).

Using a functional localizer however, the systematic eye movements no longer appeared to pose a problem. The reason is that the functional localizer was *designed* such that decoders trained on those data were primarily sensitive to the stimulus representation. Thus, by applying between-task generalization, the eye-movement related confounds were effectively filtered out. The net result is that I was able to claim that our decoding results reflected sensory patterns. It should be noted that an important condition for this

inference to be valid is that the localizer was *not* contaminated by eye movements. I found this condition to be met, indeed.

Another illustration of the benefit of using a functional localizer in conjunction with between-task generalization, over within-task decoding with cross-validation, is presented in chapter 2, where I specifically traced the sensory representation while it was being used for a perceptual decision. In the within-task decoding analysis, I discerned two processes on the basis of their temporal profile: an early sensory process and a later decision process. The results suggested that, in the case where a stimulus was presented, the sensory process followed the subject's decision: hits were significantly different from misses during early time windows (Fig. 4H, chapter 2). Accordingly, one would expect to also find an early difference between correct rejects and false alarms. This was not the case, however (Fig. 4E, chapter 2). Moreover, I also found differences during late intervals - which presumably reflected the decision - even when the eventual decision was identical (Fig. 4F,G, chapter 2). Summarizing, the within-task decoding results yielded results that were difficult to interpret. On the other hand, the between-task decoding results using the functional localizer yielded consistent and clear results: the sensory representation was veridically encoded in the neural signal, regardless of the subject's decision. Similarly, the functional localizer approach enabled us to claim in chapter 3 that the brain prepares a sensory template, rather than e.g. a decision-related signal.

An important issue to consider however is that designing a decoder to be sensitive to one factor (e.g. stimulus identity) does *not* make it *insensitive* to another factor (e.g. cue). This consideration especially played a role in the design of the expectation suppression experiment described in chapter 4. Here, I paired a leading image with a trailing image and I decoded the identity of the trailing image. Contrary to chapter 3, I did not flip the contingencies between the leading and trailing stimulus, meaning that the predictive relations remained fixed across the experiment. This also means, experimentally, that these factors (identity of the leading image and identity of the trailing image) were correlated. Therefore, if one finds a difference in the neural signal between the two trailing images, this may well be attributed to long-lasting activity from differences in the leading images, because these factors are correlated. Indeed, a within-task cross-validation analysis would not have been able to dissociate between these factors. In my case however, training the decoders on the functional localizer - which only included the trailing images - made the analysis specifically sensitive to the trailing images. However, as mentioned before, the fact that the decoders are sensitive to the trailing image does not necessarily make them insensitive to the leading image. The pivotal insight however is that exactly how they are sensitive to the leading image is *undefined*. Whereas their sensitivity to the trailing image is carefully defined (e.g. stimulus B leads to a positive response and stimulus A to a negative one), the effect of the leading image on the decoder output could be anything, and is presumably random across subjects - especially since the images were randomly selected per subject. Therefore, at the group level, the effect of the leading image is expected to cancel out and one is left with reliable decoding of the trailing image only.

Boundary conditions on (measuring) top-down modulation

Research into top-down modulation of sensory signals has had a substantial impact on the development of general theories of cortical functioning. For instance, the finding that activity in early visual neurons aligns over time with the observer's decision (Nienborg and Cumming, 2009) has led to the idea of recurrent integration between neural populations (Wimmer et al., 2015); expectation suppression (Summerfield and de Lange, 2014) is one of the key empirical predictions of the predictive coding theory (Friston, 2005) and decoding of sensory representations from visual cortex during visual working memory and imagery (Harrison and Tong, 2009; Albers et al., 2013) has contributed to the idea of viewing early visual cortex as a dynamic blackboard (Roelfsema and Lange, 2016). In this thesis I added to these developments by also demonstrating top-down modulation, and its temporal dynamics, across a variety of tasks.

However, I also demonstrated situations where I did *not* find top-down modulation, despite expecting it on the basis of previous literature. Bearing in mind that null effects should be interpreted with caution and that the absence of the hypothesized effects could simply be attributed to trivial reasons such as insufficient statistical power or limitations of the specific neuroimaging technique used, my null results might add to the field by placing boundary conditions on existing theories. For instance, according to the predictive coding theory, the brain learns statistical relations in its sensory input, resulting in attenuated activity in response to expected stimuli. However, I failed to find this expectation suppression in the experiment described in chapter 4. This might suggest that certain boundary conditions are to be met for this effect to occur - conditions that were not met in my experiment. The exact nature of these conditions remains matter for future research, and may include task-relevance of the stimuli, the amount of exposure to the statistical relations and the specific manner of stimulation. Another example is given in chapter 5, by the sustained activity of the memorized item that only lasted for approximately one second. Despite there being a long interval, and given that the stimulus was presented for only 250 ms, the result does not align with Harrison and Tong (2009) and Albers et al. (2013) who found a clear sensory representation throughout the entire delay period. Instead of the theory that visual working memory items are encoded in activity patterns in sensory cortex (Sreenivasan et al., 2014), my results are potentially more in line with the activity-silent, dynamic coding theory of working memory (Stokes, 2015).

Implications and future directions

Three central themes permeate this thesis: the sensory representation, top-down modulation thereof and the underlying temporal dynamics. These themes, and combinations thereof, have been focused on in the literature before. For instance, Kok et al. (2012) studied the top-down modulation of sensory representations by expectation, but was limited by the low temporal resolution of fMRI. On the other hand, the vast body of research on the visual mismatch negativity (Stefanics et al., 2014) commonly takes advantage of

the superior temporal resolution of electrophysiological methods, but looks at aggregate neural activity rather than sensory representations. Here I combined all three of these themes, and this has led to new insights. Indeed, this approach has recently been applied to increase our understanding of mental imagery (Dijkstra et al., 2018) and I believe it provides a fruitful avenue for future research.

I focused primarily on the visual domain. Although this limits the scope of this thesis's contribution, the hope is that eventually insights from the visual domain will be obtained that generalize to other modalities (Koch, 2004). Indeed, the predictive coding theory (Friston, 2005) and its hypothesized neural microcircuitry (Bastos et al., 2012) are thought to describe generic principles that are shared across cortical areas. Moreover, the theory has been extended to describe the neural implementation of motor functions, and has even been put forward as a general theory of the working mechanisms of the entire cerebral cortex (Friston, 2010; Clark, 2013). Given the promise of these theories to provide us with an increasing understanding of the brain, future research might aim at generalizing the results found in this thesis to other modalities and contexts.

So, how do we see? Naturally, this thesis hasn't provided a clear-cut answer. But it has yielded modest, though significant steps in our scientific endeavor to unraveling the inner workings of the brain and indeed, ultimately to understanding how the projection of light on the retina leads to a vivid, subjective percept.



Appendix

References

- Albers AM, Kok P, Toni I, Dijkerman HC, de Lange FP (2013) Shared Representations for Working Memory and Mental Imagery in Early Visual Cortex. *Curr Biol* 23:1427–1431.
- Alink A, Schwiedrzik CM, Kohler A, Singer W, Muckli L (2010) Stimulus Predictability Reduces Responses in Primary Visual Cortex. *J Neurosci* 30(8):2960–2966.
- Amado C, Hermann P, Kovács P, Grotheer M, Vidnyánszky Z, Kovács G (2016) The contribution of surprise to the prediction based modulation of fMRI responses. *Neuropsychologia* 84:105–112.
- Arnal LH, Giraud A-L (2012) Cortical oscillations and sensory predictions. *Trends Cogn Sci* 16:390–398.
- Bandt C, Weymar M, Samaga D, Hamm AO (2009) A simple classification tool for single-trial analysis of ERP components. *Psychophysiology* 46, 747–757.
- Bar M, (2004) Visual objects in context. *Nat Rev Neurosci* 5:617–629.
- Bar M, Kassam KS, Ghuman, AS, Boshyan J, Schmid AM, Dale AM, Hämäläinen MS, Marinkovic K, Schacter DL, Rosen BR, Halgren E (2006) Top-down facilitation of visual recognition. *Proc Natl Acad Sci U S A* 103(2):449–454.
- Bar M (2009) The proactive brain: memory for predictions. *Philos Trans R Soc Lond B Biol Sci* 364:1235–1243.
- Bastiaansen MCM, Knösche TR (2000) Tangential derivative mapping of axial MEG applied to event-related desynchronization research. *Clin Neurophysiol* 111(7):1300–1305.
- Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ (2012) Canonical Microcircuits for Predictive Coding. *Neuron* 76(4):695–711.
- Bastos AM, Vezoli J, Bosman CA, Schoffelen JM, Oostenveld R, Dowdall JR, de Weerd P, Kennedy H, Fries P (2015) Visual Areas Exert Feedforward and Feedback Influences through Distinct Frequency Channels. *Neuron* 85(2):390–401.
- Bauer M, Stenner M-P, Friston KJ, Dolan RJ (2014) Attentional Modulation of Alpha/Beta and Gamma Oscillations Reflect Functionally Distinct Processes. *J Neurosci* 34(48):16117–16125.
- Bekinschtein TA, et al. (2009) Neural signature of the conscious processing of auditory regularities. *Proc Natl Acad Sci* 106(5):1672–1677.
- Bell AH, Summerfield C, Morin EL, Malecek NJ, Ungerleider LG (2016) Encoding of Stimulus Probability in Macaque Inferior Temporal Cortex. *Curr Biol* 26(17):2280–2290.
- Berens P (2009) CircStat: A MATLAB Toolbox for Circular Statistics. *J Stat Softw* Vol 1 Issue 10 2009 Available at: <https://www.jstatsoft.org/v031/i10>.
- Berkes P, Orban G, Lengyel M, Fiser J (2011) Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment. *Science* 331(6013):83–87.

- Blankertz B, Lemm S, Treder M, Haufe S, Müller K-R (2011) Single-trial analysis and classification of ERP components — A tutorial. *NeuroImage* 56, 814–825.
- Bishop CM (2006) Pattern recognition and machine learning. *springer*.
- Boldt A, Yeung N (2015) Shared Neural Markers of Decision Confidence and Error Detection. *J. Neurosci.* 35, 3478–3484.
- Bone MB, St-Laurent M, Dang C, McQuiggan DA, Ryan JD, Buchsbaum BR (2017) Eye-movement reinstatement and neural reactivation during mental imagery. *bioRxiv*:107953.
- Bogacz R (2017) A tutorial on the free-energy framework for modelling perception and learning. *J Math Psychol* 76:198–211.
- Bosch SE, Jehee JFM, Fernandez G, Doeller CF (2014) Reinstatement of Associative Memories in Early Visual Cortex Is Signaled by the Hippocampus. *J Neurosci* 34(22):7493–7500.
- Brady TF, Konkle T, Alvarez GA, Oliva A (2008) Visual long-term memory has a massive storage capacity for object details. *Proc Natl Acad Sci* 105:14325–14329.
- Brainard DH (1997) The Psychophysics Toolbox. *Spat. Vis.* 10, 433–436.
- Brandt SA, Stark LW (1997) Spontaneous Eye Movements During Visual Imagery Reflect the Content of the Visual Scene. *J Cogn Neurosci* 9:27–38..
- Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J. Neurosci.* 12, 4745–4765.
- Brouwer GJ, Heeger DJ (2009) Decoding and Reconstructing Color from Responses in Human Visual Cortex. *J Neurosci* 29(44):13992–14003.
- Brouwer GJ, Heeger DJ (2011) Cross-orientation suppression in human visual cortex. *J Neurophysiol* 106(5):2108–2119.
- Cashdollar N, Ruhnau P, Weisz N, Hasson U (2016) The Role of Working Memory in the Probabilistic Inference of Future Sensory Events. *Cereb Cortex*:bhw138.
- Curtis CE, D’Esposito M (2003) Persistent activity in the prefrontal cortex during working memory. *Trends Cogn Sci* 7:415–423.
- Chalk M, Seitz AR, Series P (2010) Rapidly learned stimulus expectations alter perception of motion. *J Vis* 10(8):2–2.
- Choe KW, Blake R, Lee S-H (2014) Dissociation between Neural Signatures of Stimulus and Choice in Population Activity of Human V1 during Perceptual Decision-Making. *J Neurosci* 34:2725–2743.
- Christophel TB, Allefeld C, Endisch C, Haynes J-D (2017) View-Independent Working Memory Representations of Artificial Shapes in Prefrontal and Posterior Regions of the Human Brain. *Cereb Cortex*:1–16.

- Christophel TB, Cichy RM, Hebart MN, Haynes J-D (2015) Parietal and early visual cortices encode working memory content across mental transformations. *NeuroImage* 106:198–206.
- Cichy RM, Pantazis D, Oliva A (2014) Resolving human object recognition in space and time. *Nat Neurosci* 17(3):455–462.
- Cichy RM, Ramirez FM, Pantazis D (2015) Can visual information encoded in cortical columns be decoded from magnetoencephalography data in humans? *NeuroImage* 121:193–204.
- Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci* 36:181–204.
- Cohen MX, Frank MJ (2009) Neurocomputational models of basal ganglia function in learning, memory and choice. *Behav Brain Res* 199:141–156.
- Crick F, Koch C (1998) Consciousness and neuroscience. *Cereb. Cortex* 8, 97–107.
- Dale AM, Liu AK, Fischl BR, Buckner RL, Belliveau JW, Lewine JD, Halgren E (2000) Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron* 26:55–67.
- Davachi L, DuBrow S (2015) How the hippocampus preserves order: the role of prediction and context. *Trends Cogn Sci* 19(2):92–99.
- Deco G, Romo R (2008) The role of fluctuations in perception. *Trends Neurosci.* 31, 591–598.
- Den Ouden HEM, Friston KJ, Daw ND, McIntosh AR, Stephan KE (2009) A Dual Role for Prediction Error in Associative Learning. *Cereb Cortex* 19(5):1175–1185.
- van Dijk H, Schoffelen J-M, Oostenveld R, Jensen O (2008) Prestimulus Oscillatory Activity in the Alpha Band Predicts Visual Discrimination Ability. *J. Neurosci.* 28, 1816–1823.
- Dijkstra N, Mostert P, de Lange FP, Bosch S, van Gerven MAJ (2018) Differential temporal dynamics during visual imagery and perception. *eLife* 7:e33904.
- Dong DW, Atick JJ (1995) Statistics of natural time-varying images. *Netw Comput Neural Syst* 6:345–358.
- Egner T, Monti JM, Summerfield C (2010) Expectation and Surprise Determine Neural Population Responses in the Ventral Visual Stream. *J Neurosci* 30:16601–16608.
- Emberson LL, Richards JE, Aslin RN (2015) Top-down modulation in the infant brain: Learning-induced expectations rapidly affect the sensory cortex at 6 months. *Proc Natl Acad Sci* 112:9585–9590.
- Ester EF, Sprague TC, Serences JT (2015) Parietal and Frontal Cortex Encode Stimulus-Specific Mnemonic Representations during Visual Working Memory. *Neuron* 87(4):893–905.
- Fiser J, Berkes P, Orbán G, Lengyel M (2010) Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn Sci* 14:119–130.

- Fiser A, Mahringer D, Oyibo HK, Peterson AV, Leinweber M, Keller GB (2016) Experience-dependent spatial expectations in mouse visual cortex. *Nat Neurosci* 19(12):1658–1664.
- Foster JJ, Sutterer DW, Serences JT, Vogel EK, Awh E (2016) The topography of alpha-band activity tracks the content of spatial working memory. *J Neurophysiol* 115:168–177.
- Friston K (2005) A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci* 360:815836.
- Friston K (2010) The free-energy principle: a unified brain theory? *Nat Rev Neurosci* 11:127–138.
- Garcia JO, Srinivasan R, Serences JT (2013) Near-Real-Time Feature-Selective Modulations in Human Cortex. *Curr Biol* 23(6):515–522.
- García-Pérez MA (1998) Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Res* 38:1861–1881.
- Garrido MI, Kilner JM, Stephan KE, Friston KJ (2009) The mismatch negativity: A review of underlying mechanisms. *Clin Neurophysiol* 120:453–463.
- Gayet S, Guggenmos M, Christophel TB, Haynes J-D, Paffen CLE, Stigchel SV der, Sterzer P (2017) Visual working memory enhances the neural response to matching visual input. *J Neurosci*:3418–16.
- Gayet S, Paffen CLE, van der Stigchel S (2018) Visual Working Memory Storage Recruits Sensory Processing Areas. *Trends Cogn Sci* 22:189:190.
- Girshick AR, Landy MS, Simoncelli EP (2011) Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nat Neurosci* 14:926–932.
- Gold JI, Shadlen MN (2007) The Neural Basis of Decision Making. *Annu Rev Neurosci* 30:535–574.
- Gosselin F, Schyns PG (2003) Superstitious Perceptions Reveal Properties of Internal Representations. *Psychol. Sci.* 14, 505–509.
- Gregory RL (1997) Knowledge in perception and illusion. *Philos Trans R Soc Lond B Biol Sci* 352(1358):1121–1127.
- Grill-Spector K, Kushnir T, Hendler T, Malach R (2000) The dynamics of object-selective activation correlate with recognition performance in humans. *Nat Neurosci* 3:837–843.
- Grootswagers T, Wardle SG, Carlson TA (2016) Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *J Cogn Neurosci* 29:677–697.
- Güçlü U, van Gerven MAJ (2015) Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *J Neurosci* 35:10005–10014.
- Haefner RM, Berkes P, Fiser J (2016) Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. *Neuron* 90:649-660.

- Harrison SA, Tong F (2009) Decoding reveals the contents of visual working memory in early visual areas. *Nature* 458:632–635.
- Haxby JV, Connolly AC, Guntupalli JS (2014) Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annu Rev Neurosci* 37:435–456.
- Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J-D, Blankertz B, Bießmann F (2014) On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage* 87:96–110.
- Haxby JV, Connolly AC, Guntupalli JS (2014) Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annu Rev Neurosci* 37:435–456.
- Hazy TE, Frank MJ, O'Reilly RC (2007) Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Philos Trans R Soc B Biol Sci* 362:1601–1613.
- Hebb DO (1968) Concerning Imagery. In: *Images, Perception, and Knowledge*, pp 139–153 The University of Western Ontario Series in Philosophy of Science. Springer, Dordrecht.
- Heekeren HR, Marrett S, Ungerleider LG (2008) The neural systems that mediate human perceptual decision making. *Nat Rev Neurosci* 9:467–479.
- von Helmholtz H (1866) *Treatise on physiological optics* (The Optical Society of America, Menasha, WI). Translated from the third German edition, 1925.
- Hesselmann G, Kell CA, Kleinschmidt A (2008a) Ongoing Activity Fluctuations in hMT+ Bias the Perception of Coherent Visual Motion. *J Neurosci* 28:14481–14485.
- Hesselmann G, Kell CA, Eger E, Kleinschmidt A (2008b) Spontaneous local variations in ongoing neural activity bias perceptual decisions. *Proc Natl Acad Sci* 105:10984–10989.
- Hesselmann G, Sadaghiani S, Friston KJ, Kleinschmidt A (2010) Predictive Coding or Evidence Accumulation? False Inference and Neuronal Fluctuations. *PLoS ONE* 5, e9926.
- Hindy NC, Ng FY, Turk-Browne NB (2016) Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nat Neurosci* 19(5):665–667.
- Hubel DH, Wiesel TN (1959) Receptive fields of single neurones in the cat's striate cortex. *J Physiol* 148:574–591.
- Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* 160:106–154.2.
- Hulme OJ, Friston KF, Zeki S (2009) Neural correlates of stimulus reportability. *J. Cogn. Neurosci.* 21, 1602–1610.
- Jarosz A, Wiley J (2014) What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *J Probl Solving* 7.

- Jensen O, Hesse C (2010) Estimating distributed representations of evoked responses and oscillatory brain activity. *MEG Introd. Methods* Hansen PC Kringelbach ML Salmelin R Eds 156–185.
- Jerde TA, Merriam EP, Riggall AC, Hedges JH, Curtis CE (2012) Prioritized Maps of Space in Human Frontoparietal Cortex. *J Neurosci* 32:17382–17390.
- Jolij J, Meurs M, Haitel E (2011) Why do we see what's not there? *Commun. Integr. Biol.* 4, 764–767 (2011).
- Kaliukhovich DA, Vogels R (2014) Neurons in Macaque Inferior Temporal Cortex Show No Surprise Response to Deviants in Visual Oddball Sequences. *J Neurosci* 34:12801–12815.
- Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. *Nat Neurosci* 8:679–685.
- Kanwisher N, Yovel G (2006) The fusiform face area: a cortical region specialized for the perception of faces. *Philos Trans R Soc B Biol Sci* 361:2109–2128.
- Kaposvari P, Kumar S, Vogels R (2016) Statistical Learning Signals in Macaque Inferior Temporal Cortex. *Cereb Cortex* 28(1): 250–266.
- Kayser C, Einhäuser W, König P (2003) Temporal correlations of orientations in natural scenes. *Neurocomputing* 52–54:117–123.
- Kelly SP, O'Connell RG (2013) Internal and External Influences on the Rate of Sensory Evidence Accumulation in the Human Brain. *J. Neurosci.* 33, 19434–19441.
- Kersten D, Mamassian P, Yuille A (2004) Object Perception as Bayesian Inference. *Annu Rev Psychol* 55(1):271–304.
- Kimura M, Takeda Y (2015) Automatic prediction regarding the next state of a visual object: Electrophysiological indicators of prediction match and mismatch. *Brain Res* 1626:31–44.
- King J-R, Dehaene S (2014) Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn Sci* 18:203–210.
- King J-R, Pescetelli N, Dehaene S (2016) Brain Mechanisms Underlying the Brief Maintenance of Seen and Unseen Sensory Information. *Neuron* 92:1122–1134.
- Kleiner M, Brainard D, Pelli D, Ingling A, Murray R, Broussard C, others (2007) What's new in Psychtoolbox-3. *Perception* 36:1.
- Koch C (2004) *The Quest for Consciousness: A Neurobiological Approach*, 1 edition. Denver, Colo.: Roberts and Company Publishers.
- Kok P, Bains LJ, van Mourik T, Norris DG, De Lange FP (2016) Selective Activation of the Deep Layers of the Human Primary Visual Cortex by Top-Down Feedback. *Curr Biol* 26(3):371–376.
- Kok P, Brouwer GJ, Van Gerven MAJ, De Lange FP (2013) Prior Expectations Bias Sensory Representations in Visual Cortex. *J Neurosci* 33(41):16275–16284.

- Kok P, Failing MF, de Lange FP (2014) Prior Expectations Evoke Stimulus Templates in the Primary Visual Cortex. *J Cogn Neurosci*:1–9.
- Kok P, Jehee JFM, de Lange FP (2012) Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron* 75:265–270.
- Kok P, van Lieshout LLF, De Lange FP (2016) Local expectation violations result in global activity gain in primary visual cortex. *Sci Rep* 6:37706.
- Kok P, Mostert P, de Lange FP (2017) Prior expectations induce prestimulus sensory templates. *Proc Natl Acad Sci*:201705652.
- Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* 12, 535–540.
- Kumar S, Kaposvari P, Vogels R (2017) Encoding of Predictable and Unpredictable Stimuli by Inferior Temporal Cortical Neurons. *J Cogn Neurosci* 29:1445–1454.
- Laeng B, Bloem IM, D’Ascenzo S, Tommasi L (2014) Scrutinizing visual images: The role of gaze in mental imagery and memory. *Cognition* 131:263–283.
- Laeng B, Teodorescu D-S (2002) Eye scanpaths during visual imagery reenact those of perception of the same visual scene. *Cogn Sci* 26:207–231.
- de Lange FP, Rahnev DA, Donner TH, Lau H (2013) Prestimulus Oscillatory Activity over Motor Cortex Reflects Perceptual Expectations. *J. Neurosci.* 33, 1400–1410.
- Lamme VAF, Zipser K, Spekreijse H (2002) Masking Interrupts Figure-Ground Signals in V1. *J. Cogn. Neurosci.* 14, 1044–1053.
- Larsson J, Smith AT (2012) fMRI Repetition Suppression: Neuronal Adaptation or Stimulus Expectation? *Cereb Cortex* 22:567–576.
- Lavenex P, Amaral DG (2000) Hippocampal-neocortical interaction: a hierarchy of associativity. *Hippocampus* 10(4):420–430.
- Lee S-H, Kravitz DJ, Baker CI (2012) Disentangling visual imagery and perception of real-world objects. *NeuroImage* 59(4):4064–4073.
- Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A* 20(7):1434.
- Manahova ME, Mostert P, Kok P, Schoffelen J-M, de Lange FP (2017) Stimulus familiarity and expectation jointly modulate neural activity in the visual ventral stream. *J Cogn Neurosci* 30(9): 1366-1377.
- Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* 164, 177–190.
- Monto S, Palva S, Voipio J, Palva JM (2008) Very Slow EEG Fluctuations Predict the Dynamics of Stimulus Detection and Oscillation Amplitudes in Humans. *J. Neurosci.* 28, 8268–8272.

- Mostert P, Kok P, De Lange FP (2015) Dissociating sensory from decision processes in human perceptual decision making. *Sci Rep* 5:18253.
- Meyer T, Olson CR (2011) Statistical learning of visual transitions in monkey inferotemporal cortex. *Proc Natl Acad Sci* 108(48):19401–19406.
- Meyer T, Ramachandran S, Olson CR (2014) Statistical Learning of Serial Visual Transitions by Neurons in Monkey Inferotemporal Cortex. *J Neurosci* 34:9332–9337.
- Myers NE, Rohenkohl G, Wyart V, Woolrich MW, Nobre AC, Stokes BG (2015) Testing sensory evidence against mnemonic templates. *eLife* 4:e09000.
- Näätänen R (1990) The role of attention in auditory information processing as revealed by event-related potentials and other brain measures of cognitive function. *Behav Brain Sci* 13(02):201–233.
- Nakano T, Homae F, Watanabe H, Taga G (2008) Anticipatory Cortical Activation Precedes Auditory Events in Sleeping Infants. *PLoS ONE* 3(12):e3912.
- Nienborg HR, Cohen M, Cumming BG (2012) Decision-Related Activity in Sensory Neurons: Correlations Among Neurons and with Behavior. *Annu. Rev. Neurosci.* 35, 463–483.
- Nienborg H, Cumming BG (2009) Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *Nature* 459:89–92.
- Nienborg H, Cumming BG (2014) Decision-Related Activity in Sensory Neurons May Depend on the Columnar Architecture of Cerebral Cortex. *J. Neurosci.* 34, 3579–3585.
- Nienborg H, Roelfsema PR (2015) Belief states as a framework to explain extra-retinal influences in visual cortex. *Curr Opin Neurobiol* 32:45–52.
- O'Connell RG, Dockree PM, Kelly SP (2012) A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nat. Neurosci.* 15, 1729–1735.
- Oostenveld R, Fries P, Maris E, Schoffelen J-M (2010) FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Comput. Intell. Neurosci.* 2011, e156869.
- Pajani A, Kok P, Kouider S, De Lange FP (2015) Spontaneous Activity Patterns in Primary Visual Cortex Predispose to Visual Hallucinations. *J Neurosci* 35(37):12947–12953.
- Parra L, Alvino C, Tang A, Pearlmutter B, Young N, Osman A, Sajda P (2002) Linear spatial integration for single-trial detection in encephalography. *NeuroImage* 17:223–230.
- Philiastides MG, Sajda P (2006) Temporal Characterization of the Neural Correlates of Perceptual Decision Making in the Human Brain. *Cereb. Cortex* 16, 509–518.
- Philiastides MG, Ratcliff R, Sajda P (2006) Neural Representation of Task Difficulty and Decision Making during Perceptual Categorization: A Timing Diagram. *J. Neurosci.* 26, 8965–8975.

- Plöchl M, Ossandón JP, König P (2012) Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Front Hum Neurosci* .
- Pooresmaeili A, Poort J, Roelfsema PR (2014) Simultaneous selection by object-based attention in visual and frontal cortex. *Proc. Natl. Acad. Sci.* 111, 6467–6472.
- Pratte MS, Tong F (2014) Spatial specificity of working memory representations in the early visual cortex. *J Vis* 14:22–22.
- Rademaker RL, Serences JT (2017) Pinging the brain to reveal hidden memories. *Nat Neurosci* 20:767–769.
- Ramachandran S, Meyer T, Olson CR (2016) Prediction suppression in monkey inferotemporal cortex depends on the conditional probability between images. *J Neurophysiol* 115:355–362.
- Ramachandran S, Meyer T, Olson CR (2017) Prediction suppression and surprise enhancement in monkey inferotemporal cortex. *J Neurophysiol* 118:374–382.
- Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci* 2(1):79–87.
- Rao V, DeAngelis GC, Snyder LH (2012) Neural Correlates of Prior Expectations of Motion in the Lateral Intraparietal and Middle Temporal Areas. *J Neurosci* 32(29):10063–10074.
- Ratcliff R, McKoon G (2007) The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Comput.* 20, 873–922.
- Ratcliff R, Philiastides MG, Sajda P (2009) Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proc. Natl. Acad. Sci.* 106, 6539–6544.
- Ratcliff R, Rouder JN (2000) A diffusion model account of masking in two-choice letter identification. *J. Exp. Psychol. Hum. Percept. Perform.* 26, 127–140.
- Reddy L, Poncet M, Self MW, Peters JC, Douw L, van Dellen E, Claus S, Reijneveld JC, Baayen JC, Roelfsema PR (2015) Learning of anticipatory responses in single neurons of the human medial temporal lobe. *Nat Commun* 6:8556.
- Ress D, Backus BT, Heeger DJ (2000) Activity in primary visual cortex predicts performance in a visual detection task. *Nat. Neurosci.* 3, 940–945.
- Ress D, Heeger DJ (2003) Neuronal correlates of perception in early visual cortex. *Nat. Neurosci.* 6, 414–420.
- Richter D, Ekman M, de Lange FP (2017) Suppressed sensory response to predictable object stimuli throughout the ventral visual stream. *J Neurosci* 34:21–17.
- Roelfsema PR, Lange FP de (2016) Early Visual Cortex as a Multiscale Cognitive Blackboard. *Annu Rev Vis Sci* 2:131–151.

- Roitman JD, Shadlen MN (2002) Response of Neurons in the Lateral Intraparietal Area during a Combined Visual Discrimination Reaction Time Task. *J. Neurosci.* 22, 9475–9489.
- Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009) Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev* 16:225–237.
- Rose NS, LaRocque JJ, Riggall AC, Gosseries O, Starrett MJ, Meyering EE, Postle BR (2016) Reactivation of latent working memories with transcranial magnetic stimulation. *Science* 354:1136–1139.
- Saalmann YB, Pinsk MA, Wang L, Li X, Kastner S (2012) The Pulvinar Regulates Information Transmission Between Cortical Areas Based on Attention Demands. *Science* 337:753–756.
- Salti M, Monto S, Charles L, King J-R, Parkkonen L, Dehaene S (2015) Distinct cortical codes and temporal dynamics for conscious and unconscious percepts. *eLife* 4, e05652.
- SanMiguel I, Widmann A, Bendixen A, Trujillo-Barreto N, Schroger E (2013) Hearing Silences: Human Auditory Processing Relies on Preactivation of Sound-Specific Brain Activity Patterns. *J Neurosci* 33(20):8633–8639.
- Scimeca JM, Kiyonaga A, D’Esposito M (2018) Reaffirming the Sensory Recruitment Account of Working Memory. *Trends Cogn Sci* 22:190-192.
- Schapiro AC, Kustner LV, Turk-Browne NB (2012) Shaping of Object Representations in the Human Medial Temporal Lobe Based on Temporal Regularities. *Curr Biol* 22(17):1622–1627.
- Schapiro AC, Kustner LV, Turk-Browne NB (2012) Shaping of Object Representations in the Human Medial Temporal Lobe Based on Temporal Regularities. *Curr Biol* 22:1622–1627.
- Schurger A, Sarigiannidis I, Naccache L, Sitt JD, Dehaene S (2015) Cortical activity is more stable when sensory stimuli are consciously perceived. *Proc. Natl. Acad. Sci.* 112, E2083–E2092.
- Serences JT, Ester EF, Vogel EK, Awh E (2009) Stimulus-Specific Delay Activity in Human Primary Visual Cortex. *Psychol Sci* 20:207–214.
- Siegel M, Donner TH, Oostenveld R, Fries P, Engel AK (2007) High-Frequency Activity in Human Visual Cortex Is Modulated by Visual Motion Strength. *Cereb. Cortex* 17, 732–741.
- Smith ML, Gosselin F, Schyns PG (2012) Measuring Internal Representations from Behavioral and Brain Data. *Curr. Biol.* 22, 191–196.
- Spaak E, de Lange FP, Jensen O (2014) Local Entrainment of Alpha Oscillations by Visual Stimuli Causes Cyclic Modulation of Perception. *J Neurosci* 34:3536–3544.
- Spaak E, Watanabe K, Funahashi S, Stokes MG (2017) Stable and Dynamic Coding for Working Memory in Primate Prefrontal Cortex. *J Neurosci* 37:6503–6516.

- Spivey MJ, Geng JJ (2001) Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychol Res* 65:235–241.
- Sreenivasan KK, Curtis CE, D’Esposito M (2014) Revisiting the role of persistent neural activity during working memory. *Trends Cogn Sci* 18:82–89.
- Stefanics G, Kremláček J, Czigler I (2014) Visual mismatch negativity: a predictive coding view. *Front Hum Neurosci* 8.
- St. John-Saaltink E, Kok P, Lau HC, De Lange FP (2016) Serial Dependence in Perceptual Decisions Is Reflected in Activity Patterns in Primary Visual Cortex. *J Neurosci* 36(23):6186–6192.
- St. John-Saaltink E, Utzerath C, Kok P, Lau HC, de Lange FP (2015) Expectation Suppression in Early Visual Cortex Depends on Task Set. *PLOS ONE* 10:e0131172.
- Stocco A, Lebiere C, Anderson JR (2010) Conditional Routing of Information to the Cortex: A Model of the Basal Ganglia’s Role in Cognitive Coordination. *Psychol Rev* 117:541–574.
- Stokes MG (2015) ‘Activity-silent’ working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn Sci* 19:394–405.
- Stokes MG, Kusunoki M, Sigala N, Nili H, Gaffan D, Duncan J (2013) Dynamic Coding for Cognitive Control in Prefrontal Cortex. *Neuron* 78:364–375.
- Stokes M, Thompson R, Cusack R, Duncan J (2009a) Top-Down Activation of Shape-Specific Population Codes in Visual Cortex during Mental Imagery. *J Neurosci* 29(5):1565–1572.
- Stokes M, Thompson R, Nobre AC, Duncan J (2009b) Shape-specific preparatory activity mediates attention to targets in human visual cortex. *Proc Natl Acad Sci* 106(46):19569–19574.
- Stolk A, Todorovic A, Schoffelen J-M, Oostenveld R (2013) Online and offline tools for head movement compensation in MEG. *NeuroImage* 68:39–48.
- Summerfield C, Egner T (2009) Expectation (and attention) in visual cognition. *Trends Cogn Sci* 13(9):403–409.
- Summerfield C, Egner T (2016) Feature-Based Attention and Feature-Based Expectation. *Trends Cogn Sci* 20(6):401–404.
- Summerfield C, de Lange FP (2014) Expectation in perceptual decision making: neural and computational mechanisms. *Nat Rev Neurosci* 15:745–756.
- Summerfield C, Trittschuh EH, Monti JM, Mesulam M-M, Egner T (2008) Neural repetition suppression reflects fulfilled perceptual expectations. *Nat Neurosci* 11(9):1004–1006.
- Supér H, van der Togt C, Spekreijse H, Lamme VAF (2003) Internal State of Monkey Primary Visual Cortex (V1) Predicts Figure–Ground Perception. *J. Neurosci.* 23, 3407–3414.

- Swets JA (2014) Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers. (Psychology Press).
- Todorovic A, van Ede F, Maris E, de Lange FP (2011) Prior Expectation Mediates Neural Adaptation to Repeated Sounds in the Auditory Cortex: An MEG Study. *J Neurosci* 31(25):9118–9123.
- Todorovic A, de Lange FP (2012) Repetition Suppression and Expectation Suppression Are Dissociable in Time in Early Auditory Evoked Fields. *J Neurosci* 32:13389–13395.
- Utzerath C, St. John-Saaltink E, Buitelaar J, de Lange FP (2017) Repetition suppression to objects is modulated by stimulus-specific expectations. *Sci Rep* 7:8781.
- van Veen BD, van Drongelen W, Yuchtman M, Suzuki A (1997) Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Trans. Biomed. Eng.* 44, 867–880.
- Wacongne C, Labyt E, Wassenhove V van, Bekinschtein T, Naccache L, Dehaene S (2011) Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proc Natl Acad Sci* 108:20754–20759.
- Wallenstein GV, Hasselmo ME, Eichenbaum H (1998) The hippocampus as an associator of discontiguous events. *Trends Neurosci* 21(8):317–323.
- Watson AB, Pelli DG (1983) Quest: A Bayesian adaptive psychometric method. *Percept. Psychophys.* 33, 113–120.
- Wimmer K, Compte A, Roxin A, Peixoto D, Renart A, de la Rocha J (2015) Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area MT. *Nat Comm* 6: 6177.
- Wolff MJ, Ding J, Myers NE, Stokes MG (2015) Revealing hidden states in visual working memory using electroencephalography. *Front Syst Neurosci*:123.
- Wolff MJ, Jochim J, Akyürek EG, Stokes MG (2017) Dynamic hidden states underlying working-memory-guided behavior. *Nat Neurosci* 20(6): 864-871.
- Wyart V, de Gardelle V, Scholl J, Summerfield C (2012) Rhythmic Fluctuations in Evidence Accumulation during Decision Making in the Human Brain. *Neuron* 76, 847–858.
- Wyart V, Nobre AC, Summerfield C (2012) Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proc Natl Acad Sci* 109(9):3593–3598.
- Xu Y (2018) Sensory Cortex Is Nonessential in Working Memory Storage. *Trends Cogn Sci* 22:192-193.
- Zhang H, Liu J, Huber DE, Rieth CA, Tian J, Lee, K (2008) Detecting faces in pure noise images: a functional MRI study on top-down perception. *Neuroreport* 19, 229–233.

Nederlandse samenvatting

Zicht is een van onze belangrijkste zintuigen. Elke dag vertrouwen de meesten van ons op onze ogen om ons een accurate weergave van de wereld te geven. Onze ogen zetten informatie uit licht om in neurale signalen, die vervolgens worden doorgestuurd naar de hersenen. Hier wordt de informatie verder verwerkt, wat normaliter leidt tot bewuste waarneming. Hoe vindt deze verwerking plaats? En is onze waarneming inderdaad zo waarheidsgetrouw als we doorgaans geloven?

Het antwoord op deze laatste vraag is: nee, niet altijd. Een overtuigende illustratie waarbij onze subjectieve waarneming niet strookt met de objectieve werkelijkheid is weergegeven in Fig. 1.

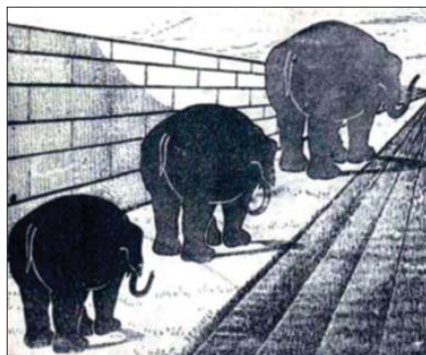


Fig 1. De olifant rechtsboven lijkt groter dan die linksonder, terwijl hun grootte in werkelijkheid identiek is.

Velen zullen beweren dat de olifant rechtsboven groter is dan die linksonder, terwijl ze in werkelijkheid dezelfde grootte hebben. De illusie wordt verklaard door de context waarin de olifanten zijn weergegeven in beschouwing te nemen. De lijnen suggereren dat de olifant rechtsboven zich op grotere afstand bevindt dan de olifant linksonder, dus de conclusie is dat de olifant rechtsboven groter moet zijn. Onze hersenen beschikken over dit soort kennis van de wereld, en maken er gebruik van bij de verwerking van de informatie afkomstig uit de ogen.

Het is duidelijk dat onze waarneming afhankelijk is van, maar niet volledig wordt bepaald door wat er daadwerkelijk om ons heen gebeurt. Behalve de externe informatie afkomstig uit licht, vinden er in onze hersenen ook processen plaats die gebruikmaken van interne informatie. We noemen zulke processen *top-down* factors, omdat ze als het ware van bovenaf beïnvloeden hoe binnenkomende, *bottom-up* informatie wordt verwerkt. In dit proefschrift richt ik mij op deze *top-down* modulatie, en met name hoe dit zich uit over een tijdsbestek van tientallen tot honderden milliseconden nadat een visuele stimulus wordt aangeboden.

Ik ben specifiek geïnteresseerd in drie vormen van *top-down* modulatie: besluitvorming, perceptuele verwachtingen en mentale inbeelding. Bij besluitvorming wordt een proefpersoon verzocht om de juiste knop in te drukken afhankelijk van de stimulus die wordt aangeboden. Perceptuele verwachtingen worden gevormd wanneer twee stimuli, die vlak na elkaar getoond worden, aan elkaar gerelateerd zijn. Na het zien van de eerste stimulus kunnen de hersenen voorspellen wat de tweede stimulus wordt en zich daarop instellen, zodat de tweede stimulus optimaal verwerkt wordt. Bij mentale inbeelding vragen we proefpersonen om zich zo levendig mogelijk een bepaalde visuele stimulus in te beelden, terwijl we onderzoeken welke hersenactiviteit daarmee gepaard gaat.

In mijn onderzoek maak ik gebruik van magnetoencephalografie (MEG). MEG is een techniek om de magnetische velden te meten die worden gegenereerd door activiteit in de hersenen. MEG beschikt over een zeer hoge temporele resolutie: we kunnen activiteit tot op de milliseconde nauwkeurig meten. Daarentegen is de spatiële resolutie relatief laag: we kunnen niet goed bepalen wáár in de hersenen de gemeten activiteit vandaan komt.

In plaats daarvan richt ik mij op de *informatie* die het neurale signaal bevat over een stimulus middels zogenaamde *multivariate decoding* analyses. We bieden een aantal stimuli aan aan een proefpersoon en we stellen vast welke patronen van hersenactiviteit gepaard gaan met specifiek die stimuli. Deze patronen noemen we de *sensorische representaties* van die stimuli, omdat ze beschrijven hoe de informatie omtrent de gepresenteerde stimulus is gerepresenteerd in het neurale signaal. In mijn experimenten onderzoek ik of en hoe deze sensorische representaties beïnvloed worden door top-down factors.

Dit proefschrift omvat vier experimenten. In het eerste toonden we proefpersonen moeilijk zichtbare gestreepte patronen, waarover ze vervolgens beslisten of ze het patroon al dan niet hadden gezien. Ik vroeg mij af hoe een stimulus in de hersenactiviteit gerepresenteerd wordt in het geval dat de persoon een verkeerde beslissing maakt (aangeven dat er een patroon was, terwijl dit er in werkelijkheid niet was en vice versa). Ik vond dat de hersenen de stimulus correct representeerden, in overeenstemming met de fysieke buitenwereld. De incorrecte waarneming moet dus voortkomen uit fouten in de verdere verwerking, waarbij de bottom-up informatie wordt omgezet in een beslissing en respons. Daarnaast deed ik nog een ontdekking: ondanks dat de stimuli slechts 50 ms werden getoond, was de sensorische representatie veel langer aanwezig, tot wel 400 ms. Dit toont aan dat de hersenen in staat zijn om middels interne, top-down processen de stimulusinformatie vast te houden, zelfs als deze fysiek niet meer aanwezig is.

In het tweede experimenten lieten we proefpersonen twee gestreepte patronen snel na elkaar zien. De strepen in deze patronen hadden een bepaalde orientatie. Belangrijk voor dit experiment was dat de orientatie van het tweede patroonafhankelijk was van het eerste patroon. Daardoor konden de hersenen, na het verwerken van de eerste stimulus, een voorspelling maken over de tweede stimulus. Onze onderzoeksvraag was hoe deze voorspelling eruit zag in termen van de sensorische representatie. We vonden dat de hersenen intrinsiek activiteit genereerde die eruitzag alsof de tweede stimulus al was getoond, nog voordat deze daadwerkelijk gepresenteerd werd. Bovendien vonden we een verband met waarneming: des te sterker deze perceptuele voorspelling in het neurale signaal, des te accurater nam de persoon de stimulus waar. Dit experiment laat zien dat de hersenen in staat zijn om intern sensorische representaties op te wekken, om de waarneming te faciliteren van stimuli die op korte termijn aangetroffen zullen worden.

Het derde experiment betrof ook perceptuele verwachtingen, maar nu was ik voornamelijk geïnteresseerd in schendingen daarvan. Eerder onderzoek heeft uitgewezen dat het tonen van een verwachte stimulus gepaard gaat met verminderde hersenactiviteit, vergeleken met het tonen van dezelfde stimulus wanneer deze niet conformeert aan de verwachting. Dit fenomeen heet *expectation suppression*. Het is echter niet duidelijk of dit effect kan worden toegeschreven aan een relatieve verlaging van activiteit bij een verwachte stimulus of juist aan een relatieve verhoging bij een onverwachte stimulus. Om deze vraag te beantwoorden ontwierp ik een experiment met nog een derde conditie, waarbij er géén specifieke verwachting kon worden opgebouwd over de te tonen stimulus. De resultaten waren verrassend: we vonden überhaupt geen expectation suppression. Aangezien expectation suppression veelvuldig is beschreven in de literatuur, draagt mijn experiment bij aan het vaststellen van de randvoorwaarden waaronder dit fenomeen zich manifesteert.

In het vierde experiment waren we geïnteresseerd in visuele inbeelding en mentale rotatie van het ingebeelde beeld. Eerder onderzoek heeft laten zien dat de neurale representatie van een ingebeeld beeld sterk lijkt op de sensorische representatie die wordt opgewerkt wanneer ditzelfde beeld daadwerkelijk wordt getoond. Dit onderzoek was echter, vanwege beperkte temporele resolutie, niet in staat om te kijken hoe de mentale rotatie de sensorische representatie veranderde. In het huidige experiment maakten we daarom gebruik van MEG. We vonden dat na een korte aanbidding (217 ms) van de in te beelden stimulus, de sensorische representatie daarvan ongeveer een seconde lang aanwezig was in het neurale signaal. Dit laat wederom zien dat de hersenen in staat zijn om relevante stimulus informatie intern vast te houden. Daarnaast vonden we dat de proefpersonen hun ogen tijdens het inbeelden systematisch bewogen in relatie tot het mentale beeld. Dit leidde tot artifacten in het neurale signaal en de multivariate decoding analyse, wat verdere interpretatie van de data belemmerde. Dit experiment leert een belangrijke les, namelijk dat oogbewegingen een ernstige ongewenste invloed kunnen hebben op neurale data en serieus ter overweging genomen dienen te worden tijdens de ontwerp- en analysefase van een experiment.

Dit proefschrift draagt bij aan de cognitieve neurowetenschap op zowel inhoudelijk als methodologisch vlak. Ik heb laten zien dat de hersenen in staat zijn om de sensorische representatie van een relevante stimulus intrinsiek op te wekken dan wel vast te houden. Het is belangrijk om op te merken dat deze sensorische representatie gedefinieerd is als de neurale activiteit die opgewekt wordt wanneer de stimulus fysiek wordt aangeboden. Dit betekent dat interne, top-down processen de verwerking van visuele informatie zodanig kunnen beïnvloeden, dat het net lijkt alsof de stimulus daadwerkelijk aanwezig is. Methodologisch is dit proefschrift innovatief, omdat ik kijk naar specifieke sensorische representaties gecombineerd met de hoge temporele resolutie van MEG. Dit stelde me in staat om experimentele vragen te beantwoorden die met andere technieken niet te beantwoorden waren. De resultaten uit deze onderzoeken dragen bij aan het begrijpen van de verwerking van visuele informatie in de hersenen.

Acknowledgements

Throughout the past four and a half years, I've had the pleasure of meeting and working with a large number of fantastic people. I'm grateful to all of those who have supported, helped, inspired or educated me, as well as to those with whom I simply had great fun. Choosing this path in my life was an educated gamble, but I can only say that I've hit jackpot.

Floris, I couldn't have wished for a better supervisor. You never failed to amaze me with your knowledge. I could always rely on you for guidance on what to do next. But what I especially appreciated is your positive and pleasant personality. You are engaged, you care for your students and you are just a fun guy to hang out with. And all of this while at the same time dealing with an ever-growing and busy career. A true inspiration.

Peter, I couldn't have wished for a better co-supervisor, either. Though we started out as fellow PhD students, you soon graduated and continued to grow as a successful researcher. But not only did you develop as a knowledgeable neuroscientist, you also knew how to be a good, caring and always-approachable mentor. I'm very grateful for all of your support and guidance, especially during the laatste loodjes.

To all of the former and current predators (I'm not spelling out individual names - I won't risk forgetting someone), you are a fantastic bunch of people, that know not only how to do science, but also how to have fun at work. I've always felt comfortable in and part of the group. It is thanks to your suggestions and involvement that my work came to what it is today (to be clear: I'm referring to this thesis, not my quitting science). I hope the Sinterklaas evening, group lunch and all the other traditions will continue (has someone already took up coffee duty again, by the way?).

Speaking of coffee (not tea, Matthias), my PhD wouldn't have been the same without Maarten, Matthias, Erik, Erik, Lieke and Marisha. Just imagine what I could have accomplished if I weren't due for a coffee break every other hour.

Thanks to Daniel, Maarten (so random) and Matthias for all the lovely board game nights. You guys ensured that I wouldn't fall into the trap of believing that one shouldn't drink during weekdays, or that one should be at work early the next morning, or that there is such a thing as "too much crisps".

Matthias and Marisha, I'm proud to have been your daily supervisor during your master's theses. Being a supervisor was a new challenge for me, but thanks to you it turned out a rewarding experience that contributed to my PhD as well as personal development.

I would also like to thank my former roommates Ruud, Tim and Maarten for making me feel welcome at the Donders right from the first day. Thanks to my later roommates,

Claudia, Annelies, Renee and Christienne too, for providing a peaceful environment where I could happily ~~facebook~~-do science.

Thanks to the Donders Institute as a whole, including the directorate, the administrative staff, the technical group and Leon, Pim and Betty&Mora. I hope the institute will continue to be that place of which every young, aspiring scientist is proud to be a member.

Jan-Matthijs and Robert, thanks for all your input and suggestions on MEG-related issues. I couldn't have done it without your help.

I'm also grateful to my parents, Wil and Chris, for being there and for moral support. You always stimulated me to choose and pursue what *I* wanted. And of course for designing and producing this booklet - a task that I dreaded ever since I realized what a PhD thesis actually is.

Last but not least: Marleen, what a sacrifice you made to leave behind beautiful Groningen, to start a new life with me in Nijmegen - all because I wanted to do neuroscience. Nevertheless, you readily managed to find your own way in this new city, while at the same time creating a warm and safe haven for me to come home to every day. Thank you for everything (and especially for marrying me).

List of publications

Peer-reviewed scientific journals

Dijkstra N, **Mostert P**, de Lange FP, Bosch S, van Gerven MAJ (2018) Differential temporal dynamics during visual imagery and perception. *eLife* 7:e33904

Fritsche M, **Mostert P**, de Lange FP (2017) Opposite Effects of Recent History on Perception and Decision. *Curr Bio* 27(4): 590-595.

Kok P, **Mostert P**, de Lange FP (2017) Prior expectations induce prestimulus sensory templates. *Proc Natl Acad Sci*:201705652.

Manahova ME, **Mostert P**, Kok P, Schoffelen J-M, de Lange FP (2017) Stimulus familiarity and expectation jointly modulate neural activity in the visual ventral stream. *J Cogn Neurosci* 30(9): 1366-1377.

Mostert P, Albers AM, Brinkman L, Todorova L, Kok P, de Lange FP (2018) Eye Movement-Related Confounds in Neural Decoding of Visual Working Memory Representations. *eNeuro* 0401-17.2018

Mostert P, Kok P, De Lange FP (2015) Dissociating sensory from decision processes in human perceptual decision making. *Sci Rep* 5:18253.

In preparation

Han B, **Mostert P**, de Lange FP (in preparation) Predictable tones elicit stimulus-specific suppression of neural activity in auditory cortex.

Meijs EL, **Mostert P**, Slagter HA, de Lange FP, van Gaal S (in preparation) Exploring the role of expectations in sensory and decision processes in the attentional blink.

Other publications

Mostert P (2016) Decoding brain activity. *Donders Wonders* 11/01/2016. Available at: <http://blog.donders.ru.nl/?p=4361&lang=en>

Mostert P (2018) Opening the black decoding box. Available at: <https://www.predictivebrainlab.com/opening-the-black-decoding-box/>

Biography

Pim was born on November 5th, 1989 in Groningen. He grew up in Assen, but moved to Groningen to study Chemistry at the Rijksuniversiteit Groningen. He soon realized Chemistry was not for him, and he switched to Psychology. He obtained the Bachelor degree in Psychology in 2011 (cum laude), with a minor in Artificial Intelligence. It was during this time that his interest in the brain was sparked and he decided to apply for the research master Behavioral and Cognitive Neuroscience in Groningen. For his first year's minor project, he investigated a possible relation between visual perception and dopamine levels under the supervision of Dr. Jacob Jolij. For his graduation project, he did a five month's internship in the lab of Prof. Gustavo Deco at the Universitat Pompeu Fabra in Barcelona. The work focused on explaining spontaneous neural activity in the visual cortex using neurocomputational models. After graduating in 2013 (summa cum laude), Pim obtained an NWO Research Talent grant to pursue a PhD under supervision of Prof. Floris de Lange and co-supervision of Dr. Peter Kok at the Donders Institute in Nijmegen. Having his thesis nearly finished in 2018, Pim decided to make a career switch. He is currently employed as a software developer at Isatis Health B.V., in Nijmegen.

Donders Graduate School for Cognitive Neuroscience

For a successful research Institute, it is vital to train the next generation of young scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School for Cognitive Neuroscience (DGCN), which was officially recognised as a national graduate school in 2009. The Graduate School covers training at both Master's and PhD level and provides an excellent educational context fully aligned with the research programme of the Donders Institute.

The school successfully attracts highly talented national and international students in biology, physics, psycholinguistics, psychology, behavioral science, medicine and related disciplines. Selective admission and assessment centers guarantee the enrolment of the best and most motivated students.

The DGCN tracks the career of PhD graduates carefully. More than 50% of PhD alumni show a continuation in academia with postdoc positions at top institutes worldwide, e.g. Stanford University, University of Oxford, University of Cambridge, UCL London, MPI Leipzig, Hanyang University in South Korea, NTNU Norway, University of Illinois, North Western University, Northeastern University in Boston, ETH Zürich, University of Vienna etc.. Positions outside academia spread among the following sectors: specialists in a medical environment, mainly in genetics, geriatrics, psychiatry and neurology. Specialists in a psychological environment, e.g. as specialist in neuropsychology, psychological diagnostics or therapy. Positions in higher education as coordinators or lecturers. A smaller percentage enters business as research consultants, analysts or head of research and development. Fewer graduates stay in a research environment as lab coordinators, technical support or policy advisors. Upcoming possibilities are positions in the IT sector and management position in pharmaceutical industry. In general, the PhDs graduates almost invariably continue with high-quality positions that play an important role in our knowledge economy.

For more information on the DGCN as well as past and upcoming defenses please visit:

<http://www.ru.nl/donders/graduate-school/phd/>

DONDERS

I N S T I T U T E



Max Planck Institute
for Psycholinguistics

ISBN 978-94-6284-172-7

Radboud University  Radboudumc