Article

Neuron

Brain Mechanisms Underlying the Brief Maintenance of Seen and Unseen Sensory Information

Highlights

- The link between working memory and visual awareness has recently been challenged
- We study here the mechanism of unconscious maintenance with MEG and machine learning
- Unseen stimuli can be partially maintained within high cortical assemblies
- We show how to revise awareness theories to account for the maintenance of invisible stimuli

Authors

Jean-Rémi King, Niccolo Pescetelli, Stanislas Dehaene

Correspondence

jeanremi.king@gmail.com

In Brief

King et al. used machine learning and magnetoencephalography to track the maintenance of subliminal images from brain activity. Their results demonstrate that invisible information can be briefly maintained in high cortical regions and consequently require the revision of theories of visual awareness.

King et al., 2016, Neuron *92*, 1122–1134
December 7, 2016
http://dx.doi.org/10.1016/j.neuron.2016.10.051

CellPress

CellPress

Neuron
Article

Brain Mechanisms Underlying the Brief Maintenance of Seen and Unseen Sensory Information

Jean-Rémi King,[1,2,6,7,*] Niccolo Pescetelli,[3,6] and Stanislas Dehaene[4,5]
[1]Department of Psychology, New York University, New York, NY 10003, USA
[2]Neuroscience Department, Frankfurt Institute for Advanced Studies, 60438 Frankfurt, Germany
[3]Department of Experimental Psychology, University of Oxford, OX1 3UD Oxford, UK
[4]Cognitive Neuroimaging Unit, CEA DSV/I2BM, INSERM, Université Paris-Sud, Université Paris-Saclay, NeuroSpin Center, 91191 Gif/Yvette, France
[5]Collège de France, 11 Place Marcelin Berthelot, 75005 Paris, France
[6]Co-first author
[7]Lead Contact
*Correspondence: jeanremi.king@gmail.com
http://dx.doi.org/10.1016/j.neuron.2016.10.051

SUMMARY

Recent evidence of unconscious working memory challenges the notion that only visible stimuli can be actively maintained over time. In the present study, we investigated the neural dynamics underlying the maintenance of variably visible stimuli using magnetoencephalography. Subjects had to detect and mentally maintain the orientation of a masked grating. We show that the stimulus is fully encoded in early brain activity independently of visibility reports. However, the presence and orientation of the target are actively maintained throughout the brief retention period, even when the stimulus is reported as unseen. Source and decoding analyses revealed that perceptual maintenance recruits a hierarchical network spanning the early visual, temporal, parietal, and frontal cortices. Importantly, the representations coded in the late processing stages of this network specifically predicted visibility reports. These unexpected results challenge several theories of consciousness and suggest that invisible information can be briefly maintained within the higher processing stages of visual perception.

INTRODUCTION

Conscious perception is often associated with the ability to hold a representation in mind. Empirically, studies have repeatedly shown that the behavioral and neuronal influence of an invisible stimulus rapidly decreases with time. Consequently, several theories of visual awareness have conjectured a strong link between the visibility of a stimulus and the maintenance of its corresponding neuronal activity. For example, the Recurrence Theory predicts that an invisible stimulus may elicit a feedforward response across the cortical hierarchies, but typically fails to trigger recurrent processing within each processing stage (Lamme and Roelfsema, 2000). Similarly, the Global Neuronal Workspace Theory predicts that an invisible stimulus may be coded in peripheral cortical modules, but that this unconscious response is too weak to recruit the fronto-parietal networks responsible for the maintenance and global broadcast of the representation across the cortex (Dehaene and Changeux, 2011).

However, the association between visual awareness and information maintenance has recently been challenged. First, several groups have shown that invisible stimuli can sometimes evoke a relatively late neuronal response (Vogel et al., 1998; Sergent et al., 2005; van Gaal et al., 2011; Silverstein et al., 2015; Bernat et al., 2001; Salti et al., 2015; Charles et al., 2014). Second, Soto and collaborators have recently demonstrated that a masked Gabor patch can be maintained for several seconds, even when subjects report not seeing the stimulus (Soto et al., 2011; Soto and Silvanto, 2014; Pan et al., 2014).

Under some conditions, these challenging empirical observations may nevertheless remain compatible with current models of visual awareness. For example, perceptual representations may be transmitted in a feedforward manner across a deep cortical hierarchy without necessarily triggering locally sustained responses. This hypothesis would remain compatible with the notion that recurrent activity and metastable representations are critical to conscious perception (Lamme and Roelfsema, 2000; Schurger et al., 2015). Alternatively, some representations could be maintained over time but may nevertheless remain confined within a sensory region. This view, dissociating information maintenance and global broadcast, would therefore be consistent with the critical role of higher-order brain regions in conscious perception (Dehaene and Changeux, 2011; Baars, 1993; Lau, 2008).

Testing these alternative accounts requires us to identify the neural mechanisms responsible for the maintenance of invisible stimuli. Specifically, we need to determine, first, whether the maintenance of invisible information is confined to early sensory stages or broadcast to higher processing stages (Baars, 1993; Dehaene and Changeux, 2011; Lau, 2008), and second, whether the maintenance of such information depends on the sustained firing rate of a coding neuronal assembly (e.g., Kojima and Goldman-Rakic, 1982; Schurger et al., 2015) and/or on the dynamic transmission of information across multiple modules (e.g., Stokes et al., 2013).

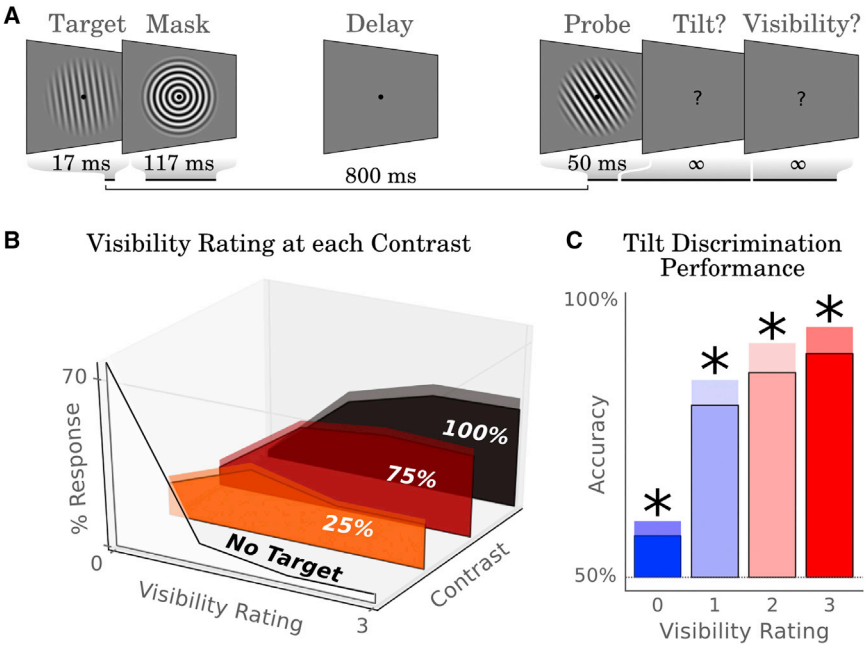

Figure 1. Visual Maintenance in Backward Masking Protocol

(A) Subjects had to mentally maintain the orientation of a masked Gabor patch to compare it to a subsequent probe (clockwise or counter-clockwise tilt). At each trial, subjects reported the visibility of the target with a four-point scale.

(B) The proportion of visibility reports for target-absent trials and each level of contrast confirms that subjects adequately used the subjective visibility ratings. Error bars indicate the SEM across subjects.

(C) Replicating Soto et al. (2011), forced-choice tilt discrimination performance correlates with visibility reports, but nevertheless remains significantly above chance in trials reported with minimum visibility (blue).

To unravel the functional architecture of perceptual maintenance and its relationship to subjective visibility, we investigated with magnetoencephalography (MEG) the neural mechanisms encoding and briefly maintaining low-level visual features, and tested how these processes varied as a function of objective stimulus features as well as subjective visibility reports.

RESULTS

Behavioral Evidence of Weak Maintenance of Unseen Stimuli

We first quantified the extent to which subjects were able to detect the masked Gabor patch (target), maintain its orientation, and compare it to a subsequent probe (Figure 1A). Subjective visibility ratings varied across the four-point visibility scale (0, completely unseen; 3, clearly seen). Subjects used the lowest visibility rating in the majority of absent trials (visibility = 0/3, 74% ± 6%). In present trials, subjects generally used one of the other three visibility ratings (visibility > 0/3, 93% ± 2%), leading to a detection (d') of 2.73 ± 0.32 (Figure 1B). This result confirms that subjects meaningfully used subjective visibility ratings. Forced-choice discrimination performance (the ability to determine whether the probe was oriented clockwise or counter-clockwise to the target) was relatively high (85% ± 5%, chance = 50%), and increased as a function of visibility ($R = 0.79 \pm 0.10$, $p < 0.001$), indicating that subjects adequately estimated their ability to detect the target. To our surprise, discrimination performance did not appear to systematically increase with the contrast of the target ($R = 0.17 \pm 0.14$, $p = 0.194$), although this effect may be underpowered by the fact that contrast varied only between three possible values. Importantly, even the targets that were reported as unseen continued to be discriminated slightly above chance level (accuracy, 58% ± 5%, $p = 0.036$; $d' = 0.20 \pm 0.09$, $p = 0.006$; Figure 1C). These behavioral results replicate the findings of Soto et al. (2011) and suggest that subjects were weakly, but significantly, able to maintain and compare the orientation of a target stimulus to that of a probe presented 800 ms later, even when they reported not seeing the target.

The Brain Automatically Encodes All Sensory Features in Parallel

We next sought to identify the neural mechanisms underlying the behavioral dissociation between visibility reports and information maintenance. We thus focused on the early brain response (100–250 ms), the delay time period (300–800 ms), and the probe time periods (900–1,150 ms) in an attempt to isolate the encoding, maintenance, and retrieval processes engaged in this task. To this end, we first compared the event-related fields (ERFs) evoked in trials with both a target and a mask to trials with a mask but no target ("absent trials"). The visual target appeared to elicit a relatively strong focal response in centro-posterior MEG channels between ~80 and 250 ms after the onset of the target (average decoding scores 100–250 ms, area under the curve [AUC] = 0.88 ± 0.01, $p < 0.001$; Figures 2A and 2B). A region of interest (ROI) analysis in the lingual gyrus suggests that this MEG activity corresponds to a sharp activity reversal between 110 (max AUC = 0.529 ± 0.005) and 169 ms (min AUC = 0.463 ± 0.006).

These early neural responses coded for the orientation of the target (Figure 2B, bottom, and Figures 3 and S2, available online). Linear circular correlations between the ERFs and the target angles revealed a focal response over posterior channels and in the lingual gyrus from ~90 ms. The corresponding decoding scores were relatively low but highly significant during this early time window (100–250 ms, 0.066 ± 0.007 radians [rad.], $p < 0.001$; Figures 3 and S2).

Task-irrelevant sensory features were also encoded in early brain responses (Figure 3). Indeed, decoding analyses demonstrated that the contrast ($R = 0.25 \pm 0.02$, $p < 0.001$), spatial frequency (AUC = 0.53 ± 0.01, $p = 0.001$), and phase of the target

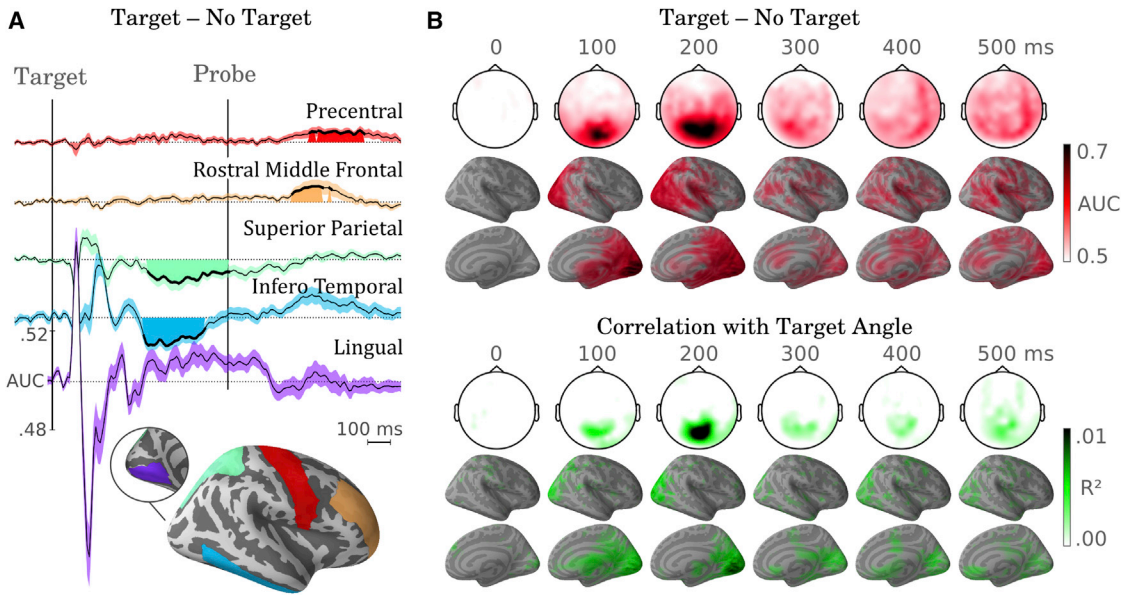

Figure 2. Weak Propagation of the Target Information across the Cortex

(A) The average area under the curve (AUC) estimates between target-present and target-absent trials within each ROI (bottom) and reveals a weak propagation of the target information from early visual cortices (purple) up to frontal regions (orange and red). Filled areas indicate significant effects after cluster correction for multiple comparisons.

(B) The exhaustive but uncorrected analyses in sensor and source spaces suggest that target information propagates to a highly distributed cortical network. Error bars indicate the SEM across subjects.

(0.024 ± 0.006 rad., $p < 0.001$ when estimators fitted on the probe) could only be decoded between ~100 and 250 ms after stimulus onset and presumably originated from occipital sources. Thus, all sensory features, whether relevant or irrelevant to the task, were simultaneously and automatically encoded in early brain responses.

The Presence, Orientation, and Visibility of the Target Are Specifically Maintained

After 250 ms, during the delay period, task-irrelevant sensory features of contrast, spatial frequency, and phase quickly dropped to chance level, but the presence, orientation, and visibility of the target remained decodable during the entire remaining time period (Figure 3). Specifically, the decoding scores of the target presence were significantly above chance during the delay period (300–800 ms, AUC = 0.73 ± 0.02, $p < 0.001$), as well as after probe onset (900–1,050 ms, AUC = 0.70 ± 0.02, $p < 0.001$; Figure 3, red), and captured a spatially distributed MEG response from ~250 ms onward, which presumably reflects a distributed set of cortical sources. This distributed pattern was confirmed by ROI-based analyses that first revealed a weak, but significant, negative response evoked in infero-temporal cortex between 410 and 690 ms (mean AUC = 0.495 ± 0.001, $p = 0.0036$), followed by a superior parietal cortex response between 430 and 800 ms (mean AUC = 0.496 ± 0.001, $p = 0.0036$; Figure 2A). A positive response could then be detected in the rostral middle frontal cortex between 1,090 and 1,270 ms (mean AUC = 0.503 ± 0.001, $p = 0.0388$) and in the precentral cortex between 1,170 and 1,420 ms after target onset (mean AUC = 0.52 ± 0.001, $p = 0.0388$). A whole-brain analysis revealed multiple weak, but significant, clusters in a relatively large number of regions, including the precuneus and cuneus, fusiform gyrus, middle and transverse temporal cortex, lateral orbitofrontal cortex and caudal middle frontal cortex, and parahippocampal area (Figures 2 and S4).

We next asked whether the neural responses identified in the delay period coded for not only the presence of the target, but also for its identity. The decoding of the target angle was significant from ~250 ms (300–800 ms, 0.032 ± 0.006 rad., $p < 0.001$). Although the sensor and source analyses suggest a similar distribution of activity pattern than the one observed for the target presence, these effects did not survive correction for multiple comparisons during the delay period. This suggests that the anatomical substrates recruited during the delay period may have been too variable across subjects to be detectable with conventional group analyses across sensors, and demonstrates the utility of decoding analyses. After the probe onset, an increased correlation between the MEG signals and the target angle could be also observed (0.055 ± 0.007 rad., $p < 0.001$) and was presumably generated by the visual cortex. Additional control analyses confirmed that the target could be decoded independently of the probe angle after probe onset ($R = 0.093 \pm 0.022$, $p = 0.001$; Figure S3). This finding suggests a recall effect reminiscent of Wolff et al. (2015) and suggests that the target information is maintained within the visual cortex in a way that is not detectable with M/EEG.

Interestingly, we observed that the neural correlates of visibility, a subjective variable, were also sustained throughout the retention period. Indeed, the decoding of visibility, restricted to target-present trials, was sustained from ~100 ms up to the

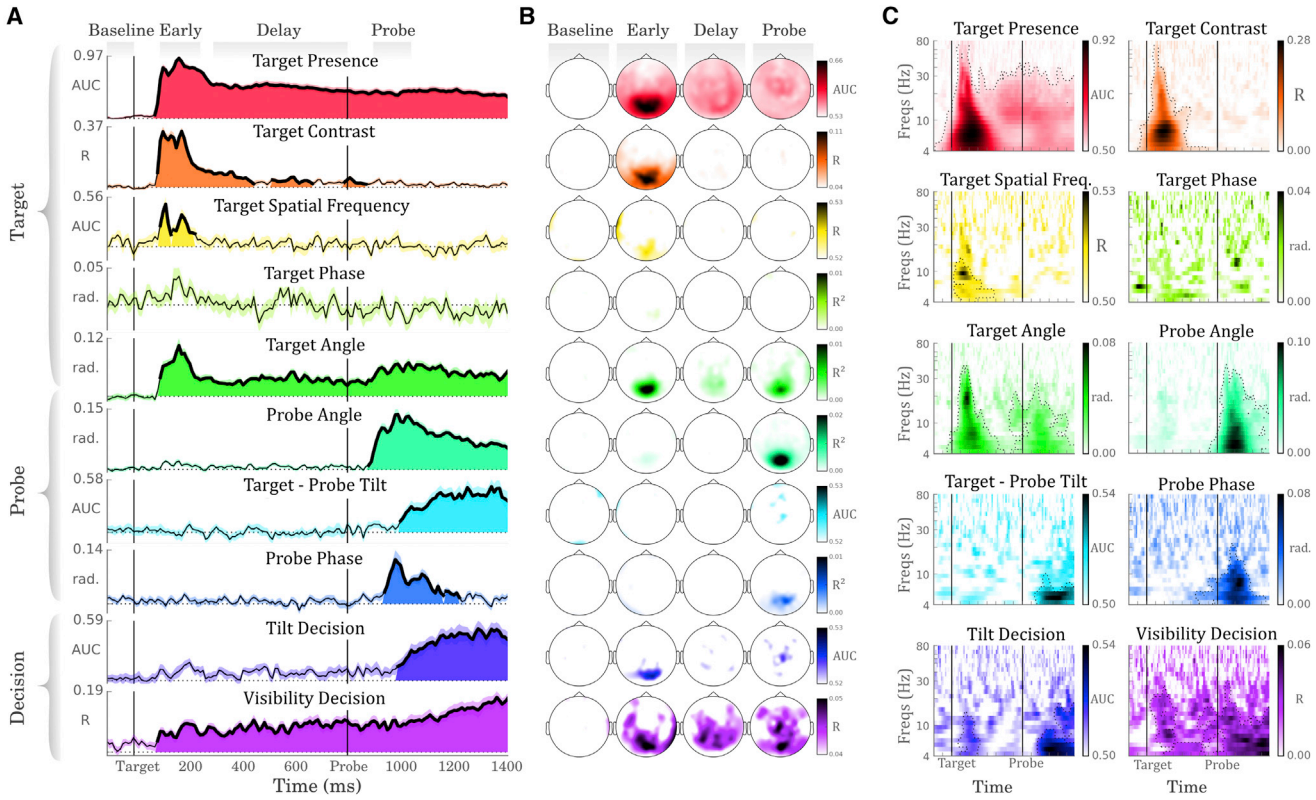

Figure 3. Decoding Reveals a Parallel Encoding of Multiple Sensory and Decisional Features

Time course of decoding performance for each sensory and decisional feature, based on the evoked MEG responses (A and B) and on their time frequency power estimates (C).

(A) Filled areas and thick lines indicate significant decoding scores (cluster corrected, $p < 0.05$) and dotted lines indicate theoretical chance level.

(B) Univariate sensor topographies depicting the average evoked response in combined gradiometers averaged within each of the four time windows of interest.

(C) Significant clusters of decoding performance are contoured with a dotted line. Overall, while all features are decodable shortly after the onset of their corresponding stimulus, only the task-relevant features of target presence, orientation, and visibility appear to be maintained in the neural activity.

Error bars indicate the SEM across subjects.

end of the epoch (100–250 ms, $R = 0.07 \pm 0.01$, $p < 0.001$; 300–800 ms, $R = 0.08 \pm 0.01$, $p < 0.001$; 900–1,050 ms, $R = 0.09 \pm 0.01$, $p < 0.001$). Both sensor and source analyses suggest that this activity is generated by a distributed set of cortical sources located in the visual, temporal, parietal, and frontal cortices. Note that although visibility ratings involved distinct finger presses within the dominant hand, neither sensor nor source analyses showed motor or premotor activity before 1,070 ms (Figures 3 and S4). This lack of detectable motor activity argues against a motor confound and fits with the fact that the median reaction time for the visibility response was 2,330 ms after the target onset.

Overall, these results suggest that while the brain first encodes, automatically and in parallel, all visual features, only those that are relevant to the task (presence, angle, and visibility) are later maintained during the delay period (Figure 3). The same conclusion was also reached on the basis of induced MEG activity. Specifically, applying similar decoding analyses after decomposing the ERFs into power estimates between 4 and 80 Hz revealed significant clusters of decoding performance of the target presence, angle, and visibility during the delay and probe periods between ~4 and 30 Hz (Figure 3C). Nevertheless, the present link between sustained neural responses and task relevance is an incidental finding. An explicit manipulation of the relevance of the sensory features would be needed to confirm that the brain selectively maintains target-relevant information and discards irrelevant information.

Unseen Sensory Information Is Maintained during the Delay and Probe Time Periods

To investigate whether the maintenance of visual information varied as a function of visibility, we separately analyzed the single-trial decoding accuracy for each visibility rating (Figure 4A).

Decoding the presence of the target was significant across all visibility conditions during the early time window (comparison of present trials at a given visibility with all absent trials, AUC = 0.83, 0.86, 0.87, and 0.88 for visibility 0–3, respectively; average SEM = ±0.02, all $p < 0.001$), as well as during the retention period (AUC = 0.64, 0.70, 0.74, and 0.75 ± 0.02, all $p < 0.001$; Figure 4A, top).

The decoding scores of the target orientations in each visibility condition were noisier than those of target presence, but showed

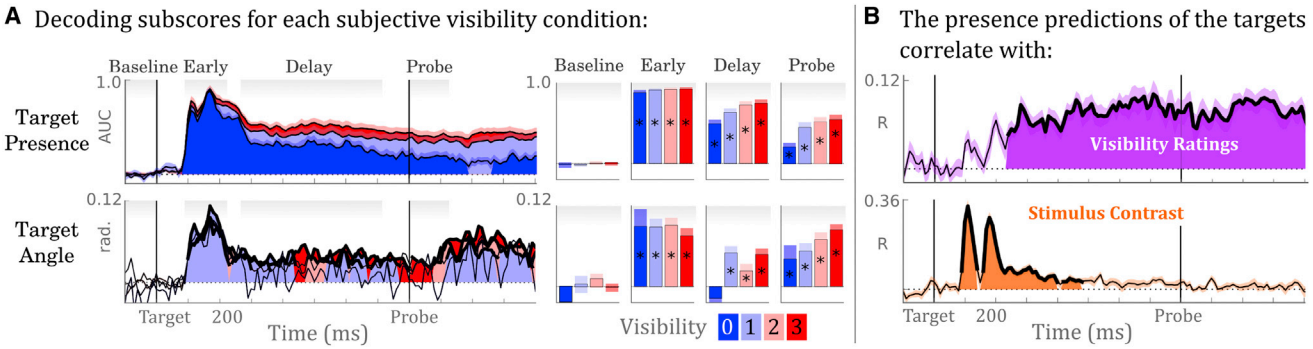

Figure 4. The Maintenance of Unseen Information Is Diminished but Remains above Chance

(A) Left: decoding subscores of the target presence (top) and target orientation (bottom) for each level of visibility (blue, 0/3; red, 3/3). Note that the estimators are trained independently of visibility ratings. For clarity purposes, the decoding scores of the target presence are estimated against all absent trials. Filled areas and thick lines indicate significant clusters, and error bars indicate the SEM across subjects. Right: decoding subscores after averaging the estimators' predictions within each time window. Overall, the results show that presence of the target can be decoded long after its presentation across multiple levels of visibility. Similar, although noisier, results are obtained when decoding the orientation of the targets.

(B) The accuracy of single-trial predictions is first modulated by the contrast of the target (orange), and then co-varies with visibility ratings, hence revealing a double dissociation between objective and subjective neural representations.

a similar overall pattern (Figure 4A, bottom). Specifically, the target orientations could be decoded shortly after the target onset in each visibility condition (0.048, 0.047, 0.049, and 0.040 ± 0.010 rad. for visibility 0–3, respectively; all $p < 0.004$), and decoding remained above chance during the delay period (0.012, 0.025, and 0.034 ± 0.010 rad. for visibility 1–3, respectively; all $p < 0.004$) and probe period (0.021, 0.028, 0.037, and 0.045 ± 0.008 rad. for visibility 0–3, respectively; all $p < 0.028$), with the notable exception of 0-visibility trials during the delay period (−0.009 ± 0.010 rad., $p = 0.100$). However, additional "bias" analyses showed that the angle decoding of 0-visibility trials was significantly biased by the target ($R = 0.083 \pm 0.065$, $p = 0.023$; Figure S5), which suggests that the orientation of unseen targets was maintained in neural activity at a barely detectable level (Figure S3).

Overall, these results show that the presence of the target and, to a lesser extent, its orientation can be partially maintained in the neural activity, even when subjects report not seeing the stimulus.

The Maintenance of Sensory Features Specifically Correlates with Visibility Ratings

Although detectable across all visibility levels, the maintained decoding of the target presence co-varied with visibility ratings from ~180 ms and until the end of the epoch (Figures 4A and 4B; $R = 0.24 \pm 0.02$, $p < 0.001$). Similar, although weaker, results were observed for target-angle analyses (Figure 4A, bottom, and Figure S3). Specifically, the accuracy of single-trial angular predictions appeared to correlate with visibility ratings during the delay (decoding $R = 0.018 \pm 0.012$, $p = 0.067$; control bias $R = 0.330 \pm 0.122$, $p = 0.0193$) and probe time periods (decoding $R = 0.030 \pm 0.011$, $p = 0.010$; control bias $R = 0.37 \pm 0.128$, $p = 0.014$).

These late decoding predictions were remarkably independent of the contrast of the target. Indeed, target contrast modulated the early decoding scores of presence ($R = 0.332 \pm 0.032$, $p < 0.001$) and orientation ($R = 0.051 \pm 0.022$, $p = 0.048$), but rapidly stopped influencing both of these codes during the delay period (Figure 4B). Finally, the correlations between the single-trial predictions and (1) visibility ratings or (2) target contrast were significantly different from one another during the early ($\Delta R = 0.248 \pm 0.034$, $p < 0.001$) and delay time windows ($\Delta R = -0.060 \pm 0.021$, $p = 0.014$), confirming the temporal specificity of these two factors.

Overall, these results show that the early encoding of the target features was performed largely independently of visibility and was mainly modulated by the contrast of the stimulus. Conversely, the late processing stages correlated with subjective visibility reports, but not with the objective stimulus contrast.

Possible Models Accounting for the Observed Sustained Activity on Unseen Trials

We next attempted to distinguish among six independent mechanisms that could account for our three main decoding findings (Figure 5): (1) decoding scores correlate with visibility reports during the delay period, but (2) not during the early period, and (3) the decoding of unseen stimuli is above chance until the probe onset.

The proposed models differ in terms of (1) their overall architecture (i.e., the number and order or the processing stages) and (2) whether they postulate that visibility correlates with the amplitude and/or the duration of a given processing stage. By design, these models fit our decoding results and relate to different aspects of neuronal theories of visual awareness. Critically, we show how each of these models can be tested by characterizing and comparing the temporal generalization (TG) analyses of each visibility condition. For concision purposes, we mainly focus on the prediction that is sufficient to invalidate a given model and incrementally increase the model complexity.

The simplest model consists of encoding and maintaining the stimulus within a unique processing stage. To account for the low but sustained decoding of unseen trials, this model implies that unseen representations are sustained over time ("metastable") and that visibility correlates with the amplitude of the

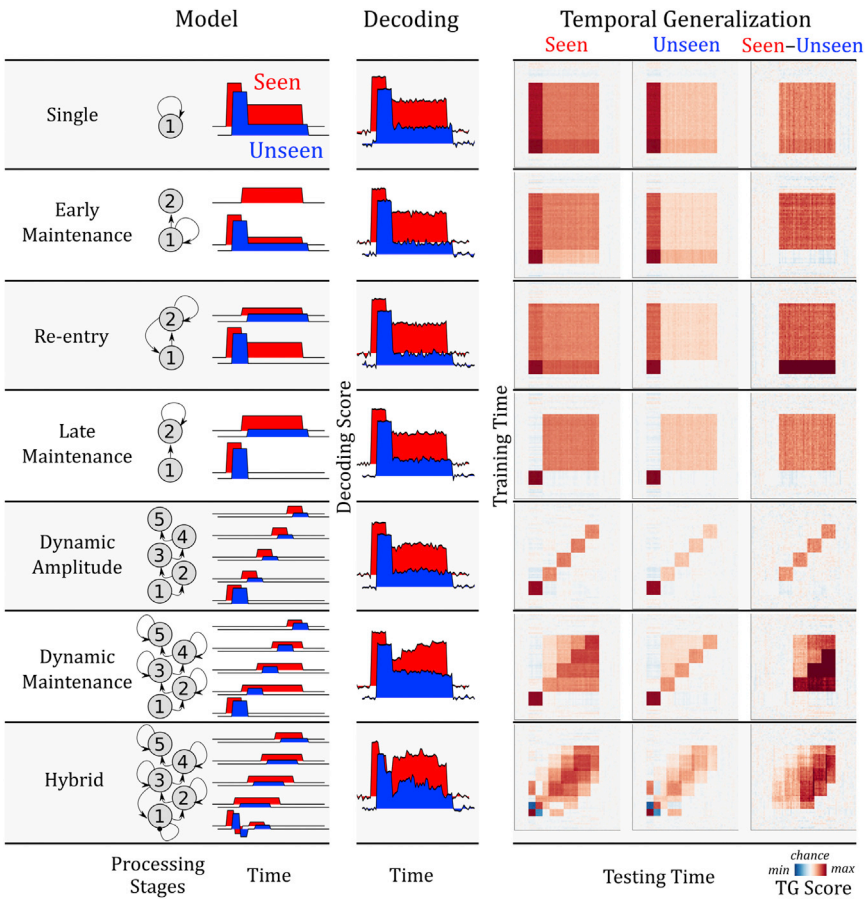

Figure 5. Hypothesized Neural Architectures of Perceptual Maintenance and Their Predicted Dynamics across Visibility Judgements

Multiple mechanisms of perceptual maintenance can account for the decoding results presented in Figure 4. The present models vary in terms of (1) the number and ordering of processing stages and (2) whether changes in visibility correlate with changes in the amplitude and/or the duration (metastability) of one or several processing stages. By design, these models generate qualitatively similar decoding performance, but may nevertheless be distinguished with TG analyses, which consists of (1) training an estimator at each time point and across all visibility conditions and (2) testing their respective ability to generalize over all other time points in each visibility condition separately (King and Dehaene, 2014a).The "single" model encodes and maintains the target information within the same neural system, yet with an amplitude that correlates with visibility during a late time period. The "early maintenance" model encodes and maintains both seen and unseen targets within the early stage, but only transmits the information to a second stage in the seen condition. The "re-entry" model maintains both seen and unseen targets in a late stage but reactivates the early stage only if the stimulus is seen. The "late-maintenance" model maintains the target in a late stage whose amplitude correlates with visibility. The "dynamic amplitude" model transmits the target information across a sequence of short-lived processing stages, whose amplitude codes for visibility. The "dynamic maintenance" model transmits the target information across a sequence of processing stages whose metastability codes for visibility. Unlike the six other models, the "hybrid" model combines multiple non-mutually exclusive mechanisms and captures the complex dynamics of our MEG results. Error bars indicate the SEM across subjects.

late neural responses. This single-stage model fits Zeki's theory of micro-consciousness (Zeki, 2003), which, contrary to many other approaches (Lamme and Roelfsema, 2000; Dehaene and Changeux, 2011; Lau, 2008; Seth, 2007), predicts a direct relationship between subjective experience and the strength of activation in sensory areas. This model predicts that a fixed pattern of neural activity should be decoded over time. To test this prediction, we estimated how the estimators trained at time t across all visibility conditions generalized to other time samples t' for each visibility condition separately. When applied systematically, this TG analysis results in a training by testing times matrix for each visibility, whose diagonals directly correspond to the decoding scores presented in the previous sections (Figure S1; King and Dehaene, 2014a). Contrary to the prediction of the single-stage model, the estimators fitted to our MEG signals were empirically better at decoding the stimulus when trained and tested at the same time point than when tested on their off-diagonal generalization scores (presence, $\Delta AUC = -0.735 \pm 0.019$, $p < 0.001$; angle, $\Delta = -0.044 \pm 0.006$ rad., $p < 0.001$). This result therefore demonstrates that a single processing stage cannot suffice to account for the maintenance of seen and unseen stimuli.

Postulating two processing stages implies that the visibility of the stimulus could relate to the amplitude and/or duration of the early and/or late processing stage. We will first consider these factors separately. The "early maintenance" model postulates that stimuli (1) are encoded and maintained within an early stage but (2) fail to be transmitted to a second processing stage in the unseen condition. This model could reconcile unconscious working memory findings (Soto and Silvanto, 2014) and the theories that dissociate unconscious sensory representations from conscious higher-order representations (e.g., Dehaene and Changeux, 2011; Lau, 2008; Baars, 1993; Seth, 2007), but goes against a strict equivalence between metastability and visibility (e.g., Schurger et al., 2015). This model predicts that in unseen trials, the late generalization scores of the estimators trained during the early period would be equal or superior to the diagonal decoding scores during the same late period. Empirically, the early estimators only weakly generalized over time in the unseen condition (presence, average $AUC = 0.526 \pm 0.011$, $p = 0.033$; cluster correction across testing time, 530 and 590 ms, $p = 0.026$; orientation, no significant cluster), and these generalization scores were lower than diagonal scores ($\Delta AUC = 0.118 \pm 0.023$, $p < 0.001$; Figures S5A and S5B). Consequently, this result invalidates the possibility that unseen representations are confined to the earliest processing stages.

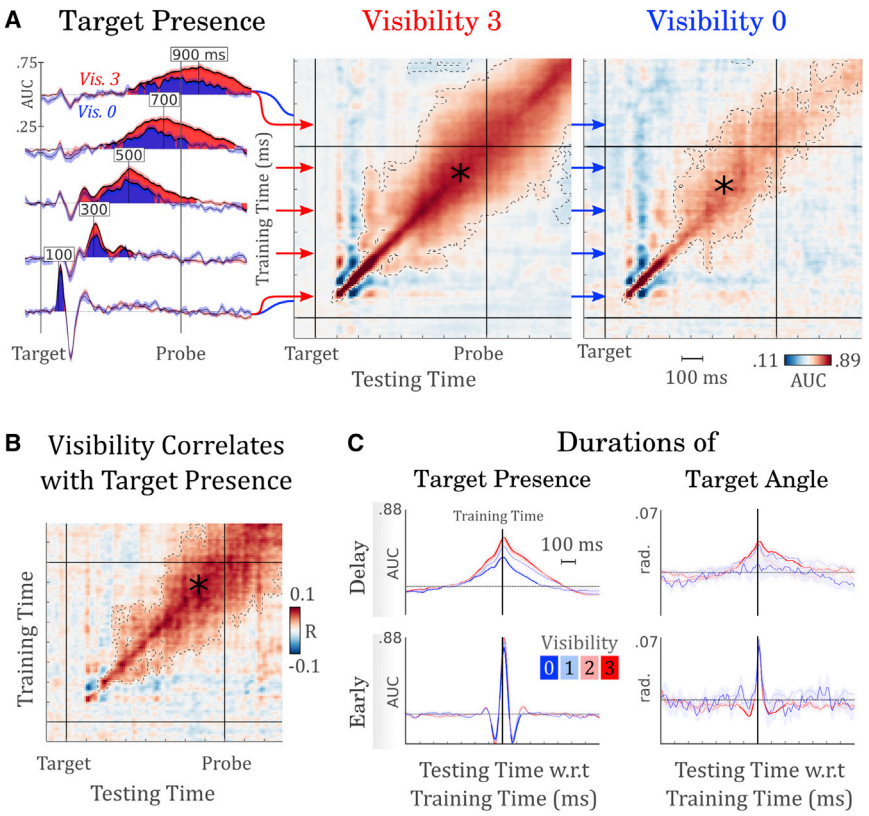

Figure 6. The Target Information Propagates to All Processing Stages and Is Partially Maintained by the Latest Stages

(A) Left: decoding time courses of five estimators trained at 100, 300, 500, 700, and 900 ms, respectively, and tested in the lowest (blue) and highest visibility conditions (red). Thick lines indicate cluster-corrected significance, and error bars indicate the SEM across subjects. Right: full TG matrices of the lowest and highest visibility conditions. Significant clusters are contoured with a dashed line. Below-chance generalizations (blue) indicate a reversal of the neural response (e.g., P1 / N1 couple). The results demonstrate that unseen stimuli propagate to all processing stages.

(B) The correlation coefficients between visibility ratings and the decoding accuracy of each estimator show that visibility selectively correlates with the late processing stages.

(C) The average decoding duration of the target presence is much briefer for the estimators trained during the early (bottom) than for the estimators trained during the delay time periods (top). The early estimators are marked by a transient bipolar response that appears virtually identical across visibility conditions whereas late estimators generalize for ~250 and 450 ms in the lowest and highest visibility conditions, respectively. Right: similar, although noisier, results were observed for target angle estimators. Overall, these results show that contrary to early processing stages, the late stages code and maintain the target information in proportion to its visibility.

The "re-entry" model postulates that the early processing stage is maintained or reactivated only if the stimulus is visible. This model requires the second stage to be metastable and identical across visibility conditions to account for the sustained decoding of unseen trials. This model fits the hypothesis that re-entrant feedback activity is critical for visual awareness (Lamme and Roelfsema, 2000; Dehaene and Changeux, 2011) but goes against the critical role of higher-order representations (Lau, 2008). This re-entry model predicts that the late generalization of early estimators should be lower in the unseen than the seen condition. Empirically, however, early estimators did not generalize differently across visibility conditions (presence, $R = 0.003 \pm 0.007$, $p = 1.000$; orientation, 0.009 ± 0.005 rad., $p = 0.101$). This result thus demonstrates that the correlation between late decoding scores and visibility ratings cannot be directly accounted for by a re-entry phenomenon.

The "late-maintenance" model postulates that unseen stimuli are weakly maintained in a late processing stage, with a lower amplitude in unseen compared to seen trials, but with a duration independent of visibility. Such a dissociation between visibility and information maintenance would argue against a tight link between visual awareness and metastability (Schurger et al., 2015) via global ignition (Dehaene and Changeux, 2011) or sustained recurrence (Lamme and Roelfsema, 2000). This model predicts that late estimators should generalize over the entire time period even when the stimulus is unseen. Empirically, however, late estimators did not generalize over the entire delay period in the unseen condition. For instance, the presence estimator trained at 500 ms only generalized between 330 and 660 ms ($p < 0.001$; Figure 6A).

In summary, these two-stage models seem unable to account for our empirical results. We therefore turn to more complex models, which postulate that unseen information can be transmitted to several processing stages. The "dynamic amplitude" model postulates that stimuli are maintained by a long sequence of processing stages, and that the amplitude of each stage correlates with the visibility of the stimulus. This dynamic architecture departs from the standard working memory architecture (Kojima and Goldman-Rakic, 1982) and extrapolates the notion of dynamic working memory (Stokes, 2015) to conditions of invisibility. It predicts that the amplitude, but not the duration, of each processing stage varies with the visibility of the stimulus. Empirically, the TG matrices revealed a diagonal pattern compatible with the predictions of this dynamic amplitude model (Figure S5), with the notable exception that the early estimators were statistically more transient (presence, 50 ± 1 ms; orientation, 77 ± 19 ms) than the late estimators (presence, 315 ± 37 ms, $p < 0.001$; orientation, 147 ± 35 ms, $p < 0.001$; Figures 6A and 6B). Furthermore, while the duration of early estimators was similar across visibility ratings (presence, 50, 50, 50, and 60 ± 1 ms; orientation, 45, 79, 55, and 50 ± 33 ms for visibility 0–3, respectively; Figure 6C, left), the duration of the late estimators was significantly shorter in the lowest visibility condition (presence, 288 ± 43 ms; orientation, 94 ± 53 ms) than in the other visible conditions (presence, 447, 477, and 442 ± 32 ms; seen – unseen, $p = 0.010$; orientation, 263, 263, and 363 ± 60 ms for

visibility 1–3, respectively; seen − unseen: $p = 0.037$). Similar, although statistically weaker, effects were also observed for target angle analyses (Figure 6C, left). Consequently, although the observed decoding performance is compatible with a model that includes a cascade of multiple successive processing stages, a simple difference in the amplitude of the late stages between seen and unseen trials cannot fully account for the neural correlates of visibility.

The "dynamic maintenance" model postulates that the stimuli are maintained by a long sequence of processing stages, and that the duration, but not the amplitude, of each stage correlates with the visibility of the stimulus. This model therefore applies a principle similar to the feedforward/recurrent dissociation (Lamme and Roelfsema, 2000) to the late processing stages. Specifically, this model predicts that in the unseen condition, the late estimators should generalize backward in time, but not forward. As mentioned above, late estimators generalized for ~288 and 94 ms forward in time for the presence and orientation estimators, respectively. Furthermore, these generalization durations were longer than the forward generalization of early estimators (all $p < 0.005$; Figure 6). These results thus show that increasing the metastability of the late stages cannot solely account for our MEG results.

To test whether the late TG depended on metastable responses at the single-trial level and were not solely due to an increased temporal variability of the late responses across trials, we replicated the TG analyses on "very high-pass-filtered" data at 2 Hz. The resulting diagonal scores remained unchanged, but the late TGs completely vanished. This result suggests that late TGs depend on slow and sustained neural responses and therefore supports the idea that late, but not early, processing stages are metastable at the single-trial level.

We therefore conclude that, separately, the two dynamic models cannot account for all empirical observations. However, these models are not mutually exclusive. Indeed, all of the observed features of the empirical TG matrices that were just reported are compatible with a hybrid architecture in which the maintenance of unseen stimuli depends on (1) a multiplicity of processing stages sequentially coding for the target information, (2) a dissociation between a transient encoding processing stage and later metastable stages, (3) an increase in the amplitude and duration of the late processing stages with subjective visibility, and (4) a re-entry of the early processing stages that is independent of visibility, supported by a weak reactivation of the early processing stage independent of visibility. Finally, we also incidentally noted a brief reversal of the early processing stage around 170 ms in most analyses (see the below chance generalization scores in Figure S5), which matches the early activity reversal identified in our sources. This reversal could correspond to an activation/inhibition response or a reversal of the direction of the electric current (King et al., 2014).

DISCUSSION

We investigated how the human brain encodes and maintains the presence, orientation, and visibility of a masked Gabor patch. Our behavioral results confirm that a stimulus subjectively rated as "completely unseen" on a four-point visibility scale can nevertheless be maintained for a sizeable duration (1.3 s), as demonstrated by the significant discrimination performance on the forced-choice task (Soto et al., 2011; Soto and Silvanto, 2014; Pan et al., 2014). This dissociation between perceptual maintenance and subjective visibility is a challenge to current theories of visual awareness (Lamme and Roelfsema, 2000; Dehaene and Changeux, 2011; Tononi and Koch, 2008; Seth, 2007; Soto et al., 2011) but may nevertheless remain compatible with the latter, depending on how the sensory information is encoded and maintained over time (Figure 5). In the present study, we therefore seek to unravel the mechanisms of perceptual maintenance and demonstrate how the decoding of MEG signals can distinguish these theoretical proposals.

Sensor analyses and estimates of the cortical sources suggest that between ~150 and 500 ms, the neural activity elicited by the target moves from the primary visual cortex to higher visual regions and finally reaches the parietal and frontal cortices after probe onset (Figure 2), in agreement with several other studies (e.g., Sergent et al., 2005; Salti et al., 2015; Cichy et al., 2014). The source estimates coding for the orientation of the target suggest similar dynamics, but fail to resist correction for multiple comparisons. This lack of robustness highlights the difficulties inherent to source analyses, including the fact that inverse modeling is ill posed in MEG, that the very high number of sources requires a drastic correction for multiple comparisons, and that neural sources may vary across subjects. This difficulty may here be increased by focusing on single-trial estimates, which are suboptimal for minimum norm estimations, because single-trial noise is modeled as neural sources.

We show here that decoding analyses overcome these limitations and reveal three main findings. First, the decoding of the presence, orientation, contrast, phase, and spatial frequency of the Gabor patch peaks early after the onset of the stimulus. Together with the source analyses, these results fit with the identification of overlapping neuronal maps in the visual cortex (Silver and Kastner, 2009) and highlight the high sensitivity of MEG recordings to subtle neural differences (Cichy et al., 2014; King and Dehaene, 2014a; Stokes et al., 2015). Importantly, these early visual codes appeared independent of visibility ratings, but correlated with the contrast of the stimulus, which strongly suggests that the encoding of the visual stimulus relates to objective rather than subjective representations.

Second, and similarly to previous findings (Salti et al., 2015; Wolff et al., 2015; Myers et al., 2015; Carlson et al., 2013; Stokes et al., 2015; Mostert et al., 2015; Cichy et al., 2014, 2016), the decoding of target presence and orientation decreased from the early time period to the delay, but then remained stable throughout the rest of the epoch, both in terms of slow evoked responses and sustained oscillatory activity (4−30 Hz; Figure 3). Critically, the decoding of the target remained above chance throughout the delay and probe time periods even in the lowest visibility condition. This finding thus confirms that unseen stimuli can be actively maintained in the neuronal activity. Interestingly, we also observed that the irrelevant sensory features of contrast, phase, and spatial frequency rapidly dropped toward chance level during the delay period. This incidental finding suggests that the brain automatically encodes all sensory features but selectively maintains those that are relevant to the task.

However, future research explicitly dissociating the task and the sensory features is necessary to confirm this selective maintenance of sensory information.

Finally, we observed that subjective visibility ratings specifically correlated with the decoding of the target presence and angle from ~250 ms and up until the probe onset. This finding confirms that the decoding of visibility does not reflect a trivial motor preparation and, importantly, suggests that both visibility and maintenance processes share a common neural substrate. Indeed, if maintenance and visibility mechanisms depended on strictly independent neural sources, then the maintained decoding scores would have been identical across visibility conditions. This finding therefore argues against the existence of a mechanism that would perfectly maintain information over time, but that would nevertheless remain inaccessible to introspection.

These three decoding findings demonstrate that we can track the active maintenance of a variably visible stimulus from MEG recordings. However, these results remain compatible with a variety of neural mechanisms, as demonstrated by our modeling (Figure 5). We therefore listed a number of elementary brain mechanisms that could conceptually underlie the observed perceptual maintenance and showed how TG analyses could be used to test each of these models. The comparison between the simulated TG and the empirical TG matrices successively invalidated a series of models and revealed three main TG results.

First, TG matrices appeared dominated by a long diagonal pattern, similar to that found in a growing number of perceptual studies (King and Dehaene, 2014a; Salti et al., 2015; Wolff et al., 2015; Myers et al., 2015; Carlson et al., 2013; Stokes et al., 2015; Mostert et al., 2015; Cichy et al., 2014; Crouzet et al., 2015; Peters et al., 2016; Meyers et al., 2008; although see, e.g., King et al., 2014; Hogendoorn et al., 2011). These diagonal patterns suggest long sequences of neural responses reminiscent of cascade models (McClelland, 1979) and strengthen a series of anatomical and functional studies revealing the deeply hierarchical organization of the cortex (Felleman and Van Essen, 1991; Rajalingham et al., 2015; Cichy et al., 2016; Chaudhuri et al., 2015). Together with our source analyses, these elements therefore strongly suggest that the target-related activity is deeply propagated across the cortex.

Second, these long TG diagonals were typically characterized by a rapid increase in TG duration after ~200 ms and revealed that early and late processing stages code the target information for ~50 and 500 ms, respectively. Such a metastability restricted to late processing stages, confirmed by very high-pass-filtered control analyses, fits with the elevated integration time of associative cortices as compared to sensory cortices (Chaudhuri et al., 2015). Interestingly, similar dynamics can be observed at a more microscopic level (Meyers et al., 2012; Stokes et al., 2013; Mante et al., 2013; King and Dehaene, 2014a), which may suggest the existence of scale-invariant computational architectures.

Finally, investigating these TG patterns in each visibility condition separately revealed that the lowest visibility condition was neither characterized by an early disruption of the diagonal pattern—as would be expected from a lack of broadcast to higher processing stages—nor by a strong or sustained (re)-activation of early processing stages—as would be expected if the maintained representations remained confined to early visual cortices (Figures 5 and 6). However, the late estimators generalized less and for a shorter duration in the lowest visibility condition than in the other visible conditions. Overall, these results therefore suggest that the perceptual maintenance of unseen stimuli is mainly accompanied by (1) a deep propagation of coding activity across all processing stages and with (2) a reduced but still significant metastable response of the late processing stages. These findings thus call for a partial revision of current theories of visual awareness in order to account for significant and metastable representations of unseen stimuli in higher processing stages.

Nevertheless, our results remain profoundly compatible with previous findings. In particular, we confirm that the visual targets first evoke a rich set of early and automatic neural responses that vary as a function of the objective properties of the stimulus, whereas later neural responses only co-vary with subjective visibility (Dehaene and Changeux, 2011; Grill-Spector et al., 2000; Fisch et al., 2009). Furthermore, the identified propagation and maintenance of "unseen" information fits with the multiple findings of late and high-level responses elicited by invisible stimuli (Vogel et al., 1998; Sergent et al., 2005; Del Cul et al., 2007; van Gaal et al., 2011; Silverstein et al., 2015; Bernat et al., 2001; Salti et al., 2015; Charles et al., 2014; Soto et al., 2011; Soto and Silvanto, 2014; Pan et al., 2014; Dutta et al., 2014). Finally, the present results also remain compatible with the idea that subjective visibility is a perceptual inference computed from ~200 ms by a highly distributed neural network (Baars, 1993; Lau, 2008; Moreno-Bote et al., 2011; King and Dehaene, 2014b; Dehaene and Changeux, 2011; Shadlen et al., 2008; Peters et al., 2016; Fisch et al., 2009). Indeed, the present data merely indicate that multiple brain regions can maintain and transmit residual sensory evidence long after an invisible stimulus is gone. However, if this information is too weak, its corresponding representation may be too similar to a noise distribution for subsequent readout processes to conclude the presence of a stimulus (Peters and Lau, 2015; King and Dehaene, 2014b).

Finally, it is important to underline that the present experimental design relies on a subjective assessment of visibility. The definition and empirical measurements of "conscious" and "unconscious" visual perception remain a topic of high controversy (e.g., Eriksen, 1960; Kouider and Dehaene, 2007; King and Dehaene, 2014b; Li et al., 2014; Pitts et al., 2014; Block, 2015; Peters and Lau, 2015; Tsuchiya et al., 2015). In particular, the current visibility metric may lead to similar ratings when subjects are confident that no target had appeared and when subjects had a very weak visual experience (Peters and Lau, 2015; King and Dehaene, 2014b). In this regard, it would be of particular interest to test how the present study generalizes to other perceptual manipulations and to different metrics of visibility and confidence. Furthermore, it remains necessary to investigate whether the present findings would generalize to maintenance periods lasting over several seconds. Longer delays may involve different neural signatures of information maintenance such as occasional stochastic bursts of activity during an otherwise silent delay period, which could account for the reduction of decoding performance between the early and delay

time periods (Mongillo et al., 2008; Stokes, 2015; Lundqvist et al., 2016; Noy et al., 2015). Beyond its empirical findings, the present study shows how such critical tests can be implemented with temporally resolved neuroimaging techniques, and thus paves the way to the identification of the processing stages that distinguish objective and subjective representations.

EXPERIMENTAL PROCEDURES

Stimuli and Protocol

Twenty young, healthy adults were scanned with MEG (22 ± 3 years old, 11 males, 18 right handed). Subjects had normal or corrected-to-normal vision. Each experiment lasted for approximately 1 hr, and participants were financially compensated ~80 euros for the study. All subjects gave written informed consent to participate in this study, which was approved by the local Ethics Committee.

Each trial started with a brief and variably contrasted target Gabor patch (17 ms), subsequently masked by a radial sinusoid (117 ms, inter-stimulus interval 50 ms; Figure 1A). A probe Gabor patch was then presented for 67 ms, 800 ms after the onset of the target. The contrast of the target was pseudo-randomly varied among 0% ("absent" trials), 25%, 75%, and 100%, whereas the contrast of the mask and the probe was fixed to 100%. Pseudo-randomization corresponds to a shuffled permutation of all conditions and was performed within each block. The orientation of the target pseudo-randomly varied among 15°, 45°, 75°, 105°, 135°, and 165°. The probe angle was tilted 30° relative to the target angle; the direction of this tilt (clockwise or counter-clockwise) was varied pseudo-randomly.

Subjects made two successive decisions. First, they performed a forced-choice discrimination task, which consisted of indicating whether the probe was tilted clockwise or counter-clockwise to the target (index and middle finger of the left hand, respectively). Subjects were then asked to report the visibility of the target (0, no experience of the target; 3, clear experience of the target, as defined by the "Perceptual Awareness Scale"; Ramsøy and Overgaard, 2004) using the index, middle, ring, and little fingers of their right hand, respectively. Subjects did 30 min of training before entering the MEG to ensure that they understood the task, and sensibly used all visibility ratings. Four subjects had to be excluded from the analysis because "unseen" or "clearly seen" reports were given less than ten times across the experiment, or because the training phase had not been completed.

The phases of the target and the probe randomly varied between −180° and 180°. The target spatial frequency pseudo-randomly varied between two possible values (30 and 35). The spatial frequency of the probe and the mask was fixed to 32.5. The target, mask, and probe stimuli had a fixed size of 16° of visual angle. Stimuli were presented on a gray background of a projector refreshed at 60 Hz, and placed 106 cm away from subject's head. Subjects were asked to keep their eyes open and to avoid eye movements by fixating on a dot continuously displayed at the center of the screen. Subjects performed a total of 840 trials, shuffled across five blocks of ~12 min each.

Circular analyses were based on the double of the Gabor angle, for the orientation only varies from 0° to 180°. Similarly, the angular errors of the decoded orientations were divided by two for consistency (i.e., $error = f(2\times\alpha)$, where f is a function using circular data, α is the stimulus orientation in radians, and *error* is the angular error). The phase of the Gabor patches was random. To facilitate the analyses and keep a consistent processing pipeline (i.e., the stratified k-folded cross-validation is only implementable with discrete values), continuous phases were digitized into six discrete, evenly separated bins.

Preprocessing

The preprocessing and statistical pipelines are available at https://github.com/kingjr/decoding_unconscious_maintenance, together with several method tutorials. MEG recordings were acquired with an ElektaNeuromag MEG system, comprising 204 planar gradiometers and 102 magnetometers in a helmet-shaped array. Subjects' head position relative to the MEG sensors was estimated with four head position coils placed on the nasion and pre-auricular points, digitized with a PolhemusIsotrak System, and triangulated before each block of trials. Six electrodes recorded electrocardiograms as well as the horizontal and vertical electro-oculograms. All signals were sampled at 1,000 Hz. Raw MEG signals were cleaned with the signal space separation (Taulu and Simola, 2006) method provided by MaxFilter to (1) suppress magnetic interferences and (2) interpolate bad MEG sensors. The signals were then high-pass filtered at 0.1 Hz (transition bandwidth = 0.01 Hz) with an overlap-add FIR filter (order = 4, length = 30 s). The raw data were then epoched between −600 ms and +1,800 ms relative to the target onset. Epochs used for evoked response analyses were further low-pass filtered at 30 Hz using MNE default parameters, cropped between −200 and 1,600, and decimated down to 100 Hz. Twenty-five Morlet wavelets with logarithmically spaced frequencies between 4 and 80 Hz (five cycles each) were used to extract the time-frequency power of the non-low-pass-filtered epochs. The resulting power estimates were then cropped between −0.200 and 1,400 ms and decimated to 125 Hz. Four large time periods of interest were used to simplify the results and maximize signal-to-noise ratio. The baseline, early, delay, and probe time windows refer to time samples between −150 and 0 ms, 100 and 250 ms, 300 and 800 ms, and 900 and 1,050 ms relative to the target onset, respectively. Indeed, several studies show that invisible stimuli can only elicit a neural response up to 250–300 ms, whereas visible stimuli typically elicit a neural ignition around this time period (Sergent et al., 2005; Gaillard et al., 2009; Fisch et al., 2009; Lamme and Roelfsema, 2000).

Sources

We retrieved the structural magnetic resonance images (MRIs) of 14 of our 20 subjects who had been previously scanned for other experiments in our lab. One of these subjects was removed from the source analyses because of a left temporal lobe anomaly. The remaining six subjects without an MRI were analyzed using the default MNI brain. Forward models were generated from the segmented and meshed MRI (decimation = 5, conductivity = [0.3, 0.006, 0.3]) using Freesurfer and MNE and manually co-registered with the head-digitized shape and fiducials (Fischl, 2012; Gramfort et al., 2013). The MNI brain was scaled to fit the head shape of subjects without an MRI. The epoch data were additionally baselined between −0.200 and 0 ms, and a shrunk co-variance (Engemann and Gramfort, 2015) was estimated across all trials from this time window. The inverse operators were generated with the default MNE parameters and applied at the single-trial level (method, dSPM, lambda2 = 0.125). The effect sizes obtained from mass-univariate analyses of estimated sources were then morphed to the MNI brain. Unless stated otherwise, source figures depict the mean of the absolute effect size across subjects in order to minimize inter-individual variability effects. Note that many source analyses failed to reach statistical significance after comparison for multiple correction over the 5,012 sources and 151 time points. They should only be interpreted with caution, and only in light of significant effects observed in sensor and decoding analyses. Additionally, we investigated five anatomical ROIs commonly activated in visual masking protocols (Dehaene and Changeux, 2011). Specifically, we used the lingual, infero-temporal, superior parietal, rostral-medial prefrontal, and precentral cortices in an attempt to isolate the sensory, perceptual, decisional, and motor responses, respectively. Each time course represents the average over all sources of the bilateral regions within each subject (e.g., mean of all sources in both left and right precentral cortices).

Statistics

Statistical analyses were based on second-level tests across subjects. Specifically, each analysis was first performed within each subject separately (across trials). We then tested the robustness of these effect sizes across subjects using, whenever possible, non-parametric statistical tests, which tend to provide more robust, although potentially less sensitive, statistical estimates. The reported effect sizes correspond to the mean effect size within subject ± SEM across subjects; the p values correspond to the second-level analyses obtained across subjects. Categorical and ordinal tests were based on two-tailed Wilcoxon and Spearman regression analyses, respectively, as provided by the Scipy package (Oliphant, 2007). Parametric circular-linear correlations were implemented from Berens (2009) and consisted of combining the linear

correlation coefficients (R) obtained between the linear data (x) and the cosine and sine of the circular data (α):

$$R_{sin} = corr(x, sin(\alpha))$$

$$R_{cos} = corr(x, cos(\alpha))$$

$$R_{norm} = corr(sin(\alpha), cos(\alpha))$$

$$R^2_{lc} = \frac{R^2_{cos} + R^2_{sin} - 2R_{cos}R_{sin}R_{norm}}{1 - R^2_{norm}},$$

where R^2_{lc} is the linear-circular correlation coefficient between x and α. Mass statistical analyses, such as those used to test the significance of each channel, each source, or each estimator at each time sample, were based on cluster-based permutation analyses (Maris and Oostenveld, 2007), using the default parameters of the MNE *spatio_temporal_cluster_1samp_test* function, which intrinsically corrects for multiple comparison issues. The default source connectivity was used for source analyses. Analyses were solely based on meaningful trials: for instance, the decoding of Gabor angle was solely based on target-present trials, and trials with missed decision responses were excluded from any analyses involving a decision factor.

Decoding

The multivariate estimators aimed at predicting a vector (y) of categorical (e.g., present versus absent), ordinal (e.g., visibility = 0, 1, 2, or 3), or circular data (e.g., Gabor angle 30°, 90°…, 330°) from a matrix of single-trial MEG data (X, shape = $n_{trials} \times (n_{chans} \times 1 \quad time \quad sample)$; Figure S1). Decoding analyses systematically consisted of (1) fitting a linear estimator (w) to a training subset of X (X_{train}), (2) predicting an estimate of y on a separate test set ($\hat{y}_{test}$), and finally (3) assessing the decoding score of these predictions as compared to the ground truth ($score(y, \hat{y})$).

Estimators

Each estimator made use of two processing steps. First, X was whitened by using a standard scaler that z scored each channel at each time point across trials. When applied onto the raw signal, this "searchlight" analysis is only able to capture effects that are consistent across trials, and is thus poor at detecting induced activity. Second, an *l2* linear model was fitted to find the hyperplane (w) that maximally predicts y from X while minimizing a log loss function. All model parameters were set to their default values as provided by the Scikit-Learn package, with the exception of setting an automatic class-weight parameter that aimed to make the analysis more robust to potential class imbalance in the dataset. Three variants of estimators were implemented to deal with categorical, ordinal, and circular data, respectively. Categorical and ordinal data were fitted with a logistic regression and a ridge regression, respectively. The logistic regression was set to generate probabilistic estimates instead of categorical predictions. Finally, a combination of two ridge regressions was used to perform circular correlations: the two ridge regressions were fitted on X to predict $sin(y)$ and $cos(y)$, respectively. The predicted angle ($\hat{y}$) was estimated from the arctangent of the resulting sine and cosine $\hat{y} = artan2(\hat{y}_{sin}, \hat{y}_{cos})$.

Cross-validation

Each estimator was fitted on each subject separately, across all MEG sensors, and at a unique time sample (sampling frequency = 100 Hz for most analyses, and 125 Hz for time-frequency decomposed analyses) using all meaningful trials. In other words, for each analysis (decoding of Gabor angle, contrast, visibility report, etc.), we fitted n_{time} estimators on an X matrix ($n_{trials} \times n_{channels} \times 1$ time sample of MEG data) to robustly predict a vector y ($n_{trials} \times 1$ categorical, ordinal, or circular data). This analysis was performed with an 8-fold stratified folding cross-validation, such that each estimator iteratively generated predictions on 1/8th of the trials (testing set) after having been fitted to the remaining 7/8th (training set) while maximizing the distribution homogeneity across training and testing sets (stratification).

Single-Trial Scores

Mass-univariate and decoding analyses generated vectors of probabilistic, ordinal, or circular data ($\hat{y}$) that could be compared to the trials' actual categorical, ordinal, or circular value (y). Categorical effects were summarized with an empirical AUC applied across all trials (AUC ranges between 0 and 1, *chance* = 0.5). Ordinal effects were summarized with a Spearman correlation R coefficient (range between −1 and 1, *chance* = 0). Circular decoding was summarized by computing the mean absolute difference between $\hat{y}$ and y (range between 0 and π, *chance* = $\pi/2$). To facilitate visualizations, this "error" metric was transformed into an "accuracy" metric (range between $-\pi/2$ and $\pi/2$, *chance* = 0). The corresponding univariate analyses, linear-circular correlations, were summarized with an R^2 value.

Correlation of Decoding Scores with Visibility and Contrast

To assess the extent to which some neural codes varied as a function of the contrast and the visibility of the target, we correlated the single-trial prediction errors with these two factors separately and within each subject: $R = corr(y, -error)$, where $error = |0.5 - \hat{y}|$ for categorical models and $error = |(\pi/2) - (y - \hat{y}) mod \pi|$ for circular models. These correlations were first applied within each visibility or contrast condition, and average across conditions. For example, to compute the correlation between decoding scores and target contrast, we computed

$$R_{contrast} = \frac{1}{n} \sum_{vis=0}^{n} corr(y(vis), -error(vis)),$$

where *vis* is the visibility condition. Only present trials were analyzed in all these analyses. The reported R values correspond to the mean correlation coefficients across subjects, and the p values reflect second-level Wilcoxon tests.

Time Periods of Interest

The decoding scores obtained for a large time window of interest were generated by (1) averaging the decoding predictions across the selected time samples at the single trial level, (2) computing the unique resulting score for each subject, and (3) performing a univariate categorical or ordinal test across subjects. The averaging of circular data (e.g., decoded angle of a target) was performed in the complex space:

$$\mu = artan2\left(\sum_i^n \frac{\sin\alpha_i \times r_i}{n}, \sum_i^n \frac{\cos\alpha_i \times r_i}{n}\right),$$

where α_i is the angle at trial *i*, n is the number of trials, r is the predicted radius, and μ the average angle.

Temporal Generalization

Time-resolved decoding analyses are a specific case of TG analyses where the estimators are fitted, tested, and scored with a unique time sample (Figure S1). Each estimator fitted across trials at time t can also be tested on its ability to accurately predict a given trial at time t', so as to estimate whether the coding pattern is similar between t and t'. When applied systematically across all pairs of time samples, this analysis results in a square TG matrix, where the y axis corresponds to the time at which the estimator was fitted, and the x axis to the time at which the estimator was evaluated.

All decoding analyses were performed with the MNE (Gramfort et al., 2013) and Scikit-Learn packages (Pedregosa et al., 2011). Most first-level decoding analyses have been integrated to the MNE package under the *TimeDecoding* and *GeneralizationAcrossTime* classes.

Simulations

Different neural architectures can account for the maintenance of a stimulus subjectively rated as unseen. To clarify how TG analyses can disentangle these models, we ran a series of exemplary simulations. Each model consisted of simulating the time courses S of shape $n_{sources}, n_{times}$ for one of two possible categories $y \in (-1, 1)$ defining the present versus absent condition. Both conditions had the same background source signal B, to simulate the presence of a common mask, probe, and response. In the present condition, an additional signal was generated, potentially with a different signal ratio depending on the visibility condition. This source space was then projected to a sensor space with a random mixing matrix A of shape n_{sensor}, n_{source}, and added to a white noise N simulating sensor-specific noise: $Xi = A(Sy_i + B) + N$, where X is the resulting sensor data and i is a given trial.

X and y were then analyzed with the same decoding and TG analyses as in the empirical sections.

Our modeling relies on three premises. First, it is based on the notion of "processing stages," here defined as a set of detectable neural sources that are

activated simultaneously. Second, it only accounts for evoked responses. Third, it assumes that invisible stimuli recruit at most a subset of the stages recruited by visible stimuli. Our modeling is therefore independent from anatomical locations and thus adequately tests signals that are not well spatially resolved, such as MEG signals. We focus here son six independent models that isolate elementary properties distinguishing the neuronal theories of visual awareness.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and five figures and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2016.10.051.

AUTHOR CONTRIBUTIONS

Conceptualization, J.-R.K., N.P., and S.D.; Methodology, J.-R.K.; Software, J.-R.K.; Investigation, N.P. and J.-R.K.; Formal Analysis: J.-R.K.; Resources, S.D.; Data Curation, J.-R.K.; Writing – Original Draft, J.-R.K.; Writing – Review & Editing, J.-R.K., S.D., and N.P.; Visualization, J.-R.K.; Supervision, J.-R.K.; Project Administration, S.D.; Funding Acquisition, J.-R.K. and S.D.

ACKNOWLEDGMENTS

This project received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 660086, INSERM, CEA, Collège de France, the Direction Générale de l'Armement, the Bettencourt-Schueller Foundation, the Fondation Roger de Spoelberch, and the Philippe Foundation. We are grateful to our anonymous reviewers and to Denis Engemann, Alex Gramfort, Eric Larson, Sébastien Marti, Virginie Van Wassenhove and her group, and Valentin Wyart, as well as to the MNE and Scikit-Learn communities for their invaluable comments and support.

Received: February 23, 2016
Revised: July 5, 2016
Accepted: October 21, 2016
Published: December 7, 2016

REFERENCES

Baars, B.J. (1993). A Cognitive Theory of Consciousness (Cambridge University Press).

Berens, P. (2009). Circstat: a matlab toolbox for circular statistics. J. Stat. Softw. *31*, 1–21.

Bernat, E., Bunce, S., and Shevrin, H. (2001). Event-related brain potentials differentiate positive and negative mood adjectives during both supraliminal and subliminal visual processing. Int. J. Psychophysiol. *42*, 11–34.

Block, N. (2015). Consciousness, big science, and conceptual clarity. In The Future of the Brain, J. Freeman and G. Marcus, eds. (Princeton University Press), pp. 161–176.

Carlson, T., Tovar, D.A., Alink, A., and Kriegeskorte, N. (2013). Representational dynamics of object vision: the first 1000 ms. J. Vis. *13*, 1.

Charles, L., King, J.-R., and Dehaene, S. (2014). Decoding the dynamics of action, intention, and error detection for conscious and subliminal stimuli. J. Neurosci. *34*, 1158–1170.

Chaudhuri, R., Knoblauch, K., Gariel, M.-A., Kennedy, H., and Wang, X.-J. (2015). A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. Neuron *88*, 419–431.

Cichy, R.M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. Nat. Neurosci. *17*, 455–462.

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. arXiv, arXiv:1601.02970, https://arxiv.org/abs/1601.02970.

Crouzet, S.M., Busch, N.A., and Ohla, K. (2015). Taste quality decoding parallels taste sensations. Curr. Biol. *25*, 890–896.

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

Del Cul, A., Baillet, S., and Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. PLoS Biol. *5*, e260.

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

Engemann, D.A., and Gramfort, A. (2015). Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals. Neuroimage *108*, 328–342.

Eriksen, C.W. (1960). Discrimination and learning without awareness: a methodological survey and evaluation. Psychol. Rev. *67*, 279–300.

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

Fisch, L., Privman, E., Ramot, M., Harel, M., Nir, Y., Kipervasser, S., Andelman, F., Neufeld, M.Y., Kramer, U., Fried, I., and Malach, R. (2009). Neural "ignition": enhanced activation linked to perceptual awareness in human ventral stream visual cortex. Neuron *64*, 562–574.

Fischl, B. (2012). FreeSurfer. Neuroimage *62*, 774–781.

Gaillard, R., Dehaene, S., Adam, C., Clémenceau, S., Hasboun, D., Baulac, M., Cohen, L., and Naccache, L. (2009). Converging intracranial markers of conscious access. PLoS Biol. *7*, e61.

Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. Front. Neurosci. *7*, 267.

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

Hogendoorn, H., Carlson, T.A., and Verstraten, F.A.J. (2011). Mapping the route to visual awareness. J. Vis. *11*, 1–10.

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

King, J.-R., and Dehaene, S. (2014b). A model of subjective report and objective discrimination as categorical decisions in a vast representational space. Philos. Trans. R. Soc. Lond. B Biol. Sci. *369*, 20130204.

King, J.-R., Gramfort, A., Schurger, A., Naccache, L., and Dehaene, S. (2014). Two distinct dynamic modes subtend the detection of unexpected sounds. PLoS ONE *9*, e85791.

Kojima, S., and Goldman-Rakic, P.S. (1982). Delay-related activity of prefrontal neurons in rhesus monkeys performing delayed response. Brain Res. *248*, 43–49.

Kouider, S., and Dehaene, S. (2007). Levels of processing during nonconscious perception: a critical review of visual masking. Philos. Trans. R. Soc. Lond. B Biol. Sci. *362*, 857–875.

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

Lau, H.C. (2008). A higher order Bayesian decision theory of consciousness. Prog. Brain Res. *168*, 35–48.

Li, Q., Hill, Z., and He, B.J. (2014). Spatiotemporal dissociation of brain activity underlying subjective awareness, objective performance and confidence. J. Neurosci. *34*, 4382–4395.

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

Mante, V., Sussillo, D., Shenoy, K.V., and Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature *503*, 78–84.

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods *164*, 177–190.

McClelland, J.L. (1979). On the time relations of mental processes: an examination of systems of processes in cascade. Psychol. Rev. *86*, 287–330.

Meyers, E.M., Freedman, D.J., Kreiman, G., Miller, E.K., and Poggio, T. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. J. Neurophysiol. *100*, 1407–1419.

Meyers, E.M., Qi, X.-L., and Constantinidis, C. (2012). Incorporation of new information into prefrontal cortical activity after learning working memory tasks. Proc. Natl. Acad. Sci. USA *109*, 4651–4656.

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

Moreno-Bote, R., Knill, D.C., and Pouget, A. (2011). Bayesian sampling in visual perception. Proc. Natl. Acad. Sci. USA *108*, 12491–12496.

Mostert, P., Kok, P., and de Lange, F.P. (2015). Dissociating sensory from decision processes in human perceptual decision making. Sci. Rep. *5*, 18253.

Myers, N.E., Walther, L., Wallis, G., Stokes, M.G., and Nobre, A.C. (2015). Temporal dynamics of attention during encoding versus maintenance of working memory: complementary views from event-related potentials and alpha-band oscillations. J. Cogn. Neurosci. *27*, 492–508.

Noy, N., Bickel, S., Zion-Golumbic, E., Harel, M., Golan, T., Davidesco, I., Schevon, C.A., McKhann, G.M., Goodman, R.R., Schroeder, C.E., et al. (2015). Intracranial recordings reveal transient response dynamics during information maintenance in human cerebral cortex. Hum. Brain Mapp. *36*, 3988–4003.

Oliphant, T.E. (2007). Python for scientific computing. Comput. Sci. Eng. *9*, 10–20.

Pan, Y., Lin, B., Zhao, Y., and Soto, D. (2014). Working memory biasing of visual perception without awareness. Atten. Percept. Psychophys. *76*, 2051–2062.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in python. J. Mach. Learn. Res. *12*, 2825–2830.

Peters, M.A.K., and Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. eLife *4*, e09651.

Peters, B., Bledowski, C., Rieder, M., and Kaiser, J. (2016). Recurrence of task set-related MEG signal patterns during auditory working memory. Brain Res. *1640* (Pt B), 232–242.

Pitts, M.A., Metzler, S., and Hillyard, S.A. (2014). Isolating neural correlates of conscious perception from neural correlates of reporting one's perception. Front. Psychol. *5*, 1078.

Rajalingham, R., Schmidt, K., and DiCarlo, J.J. (2015). Comparison of object recognition behavior in human and monkey. J. Neurosci. *35*, 12127–12136.

Ramsøy, T.Z., and Overgaard, M. (2004). Introspection and subliminal perception. Phenomenol. Cogn. Sci. *3*, 1–23.

Salti, M., Monto, S., Charles, L., King, J.-R., Parkkonen, L., and Dehaene, S. (2015). Distinct cortical codes and temporal dynamics for conscious and unconscious percepts. eLife *4*, http://dx.doi.org/10.7554/eLife.05652.

Schurger, A., Sarigiannidis, I., Naccache, L., Sitt, J.D., and Dehaene, S. (2015). Cortical activity is more stable when sensory stimuli are consciously perceived. Proc. Natl. Acad. Sci. USA *112*, E2083–E2092.

Sergent, C., Baillet, S., and Dehaene, S. (2005). Timing of the brain events underlying access to consciousness during the attentional blink. Nat. Neurosci. *8*, 1391–1400.

Seth, A. (2007). Models of consciousness. Scholarpedia *2*, 1328.

Shadlen, M.N., Kiani, R., Hanks, T.D., and Churchland, A.K. (2008). Neurobiology of decision making: an intentional framework. In Better Than Conscious? Decision Making, the Human Mind, and Implications for Institutions, C. Engel and W. Singer, eds. (MIT Press), pp. 71–101.

Silver, M.A., and Kastner, S. (2009). Topographic maps in human frontal and parietal cortex. Trends Cogn. Sci. *13*, 488–495.

Silverstein, B.H., Snodgrass, M., Shevrin, H., and Kushwaha, R. (2015). P3b, consciousness, and complex unconscious processing. Cortex *73*, 216–227.

Soto, D., and Silvanto, J. (2014). Reappraising the relationship between working memory and conscious awareness. Trends Cogn. Sci. *18*, 520–525.

Soto, D., Mäntylä, T., and Silvanto, J. (2011). Working memory without consciousness. Curr. Biol. *21*, R912–R913.

Stokes, M.G. (2015). 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. Trends Cogn. Sci. *19*, 394–405.

Stokes, M.G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., and Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. Neuron *78*, 364–375.

Stokes, M.G., Wolff, M.J., and Spaak, E. (2015). Decoding rich spatial information with high temporal resolution. Trends Cogn. Sci. *19*, 636–638.

Taulu, S., and Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. Phys. Med. Biol. *51*, 1759–1768.

Tononi, G., and Koch, C. (2008). The neural correlates of consciousness: an update. Ann. N Y Acad. Sci. *1124*, 239–261.

Tsuchiya, N., Wilke, M., Frässle, S., and Lamme, V.A.F. (2015). No-report paradigms: extracting the true neural correlates of consciousness. Trends Cogn. Sci. *19*, 757–770.

van Gaal, S., Lamme, V.A.F., Fahrenfort, J.J., and Ridderinkhof, K.R. (2011). Dissociable brain mechanisms underlying the conscious and unconscious control of behavior. J. Cogn. Neurosci. *23*, 91–105.

Vogel, E.K., Luck, S.J., and Shapiro, K.L. (1998). Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. J. Exp. Psychol. Hum. Percept. Perform. *24*, 1656–1674.

Wolff, M.J., Ding, J., Myers, N.E., and Stokes, M.G. (2015). Revealing hidden states in visual working memory using electroencephalography. Front. Syst. Neurosci. *9*, 123.

Zeki, S. (2003). The disunity of consciousness. Trends Cogn. Sci. *7*, 214–218.