Probing the limits of activity-silent non-conscious working memory

Darinka Trübutschek[a,b,c,1], Sébastien Marti[c,2], Henrik Ueberschär[d], and Stanislas Dehaene[c,e,1]

[a]Cognitive Neuroimaging Unit, Commissariat à l'Energie Atomique DSV/I2BM, INSERM, NeuroSpin Center, Université Paris-Sud, Université Paris-Saclay, 91191 Gif/Yvette, France; [b]Ecole des Neurosciences de Paris Ile-de-France, 75006 Paris, France; [c]Ecole Doctorale Cerveau-Cognition-Comportement, Sorbonne Université, 75005 Paris, France; [d]Institut de Mathématiques de Jussieu, Sorbonne Université, 75005 Paris, France; and [e]Collège de France, 75005 Paris, France

Contributed by Stanislas Dehaene, May 21, 2019 (sent for review December 7, 2018; reviewed by Floris P. de Lange, Lucia Melloni, and David Soto)

Two types of working memory (WM) have recently been proposed: (*i*) active WM, relying on sustained neural firing, and (*ii*) activity-silent WM, for which firing returns to baseline, yet memories may be retained by short-term synaptic changes. Activity-silent WM in particular might also underlie the recently discovered phenomenon of non-conscious WM, which permits even subliminal stimuli to be stored for several seconds. However, whether both states support identical forms of information processing is unknown. Theory predicts that activity-silent states are confined to passive storage and cannot operate on stored information. To determine whether an explicit reactivation is required before the manipulation of information in WM, we evaluated whether participants could mentally rotate brief visual stimuli of variable subjective visibility. Behaviorally, even for unseen targets, subjects reported the rotated location above chance after several seconds. As predicted, however, at the time of mental rotation, such blindsight performance was accompanied by (*i*) neural signatures of consciousness in the form of a sustained desynchronization in alpha/beta frequency and (*ii*) a reactivation of the memorized information as indicated by decodable representations of participants' guess and response. Our findings challenge the concept of genuine non-conscious "working" memory, argue that activity-silent states merely support passive short-term memory, and provide a cautionary note for purely behavioral studies of non-conscious information processing.

working memory | conscious perception | activity-silent brain states | non-conscious working memory | magnetoencephalography

Working memory (WM) is critical to store information for rapid access, transformation, and flexible use. Until recently, it was thought to depend on conscious processing (1, 2) and persistent neural activity (3, 4). However, a growing body of work suggests that successful WM maintenance may be dissociated from both. Subjectively unseen items may still be retrieved above chance level after several seconds (i.e., non-conscious WM; refs. 5–10). Likewise, content-specific delay-period activity may be disrupted (11) or even vanish entirely when maintaining non-conscious or unattended information (i.e., activity-silent WM; refs. 9 and 12–14).

Theories and simulations predict that such "activity-silent" maintenance without accompanying neural activity may be supported by transient, functional changes in synapses, temporarily linking neural populations coding for the stored items (15, 16). Later, a nonspecific stimulation of these very neurons may reinstate the original firing pattern, an effect that was recently observed experimentally (12, 13). Short-term synaptic changes may thus effectively allow networks to go silent for several seconds while still supporting delayed information readout. Although not always explicitly stated as such, the bulk of the available evidence opens up the possibility that these activity-silent representations may be preferentially reserved for information not currently accessible to conscious awareness (e.g., refs. 9 and 17).

Critically, it remains unknown whether active vs. activity-silent forms of WM support identical forms of information processing. Beyond maintenance, a key feature of WM is the ability to manipulate information, for example during mental rotation (18). If non-conscious WM representations are indeed stored via activity-silent short-term synaptic changes, it is unclear whether they might be altered without first being reactivated (i.e., coded via active neural firing). Neural network models operate by exchanging patterns of spiking activity, and there exists no theory of how computations could unfold solely via transient synaptic changes. In fact, modifications or transformations of activity-silent representations or states are thought to be activity-dependent (16). Thus, we predicted that for activity-silent WM to enter into an information-processing stream it would first have to be reinstated into an active form.

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

Results

We first collected behavioral measures (n = 23), then recorded magnetoencephalography (MEG) signals in a second sample

Significance

Our work tackles a current debate regarding working memory (WM). Traditionally, this ability to maintain and manipulate information has been thought to require conscious processing and persistent neural activity. Recent evidence challenges this assumption: Information may be briefly stored without conscious awareness and sustained neural activity. Here, we reconcile these competing views. We recorded brain activity while adults remembered or mentally rotated subjectively visible or invisible stimuli. We showed that the mere short-term storage of information may proceed without consciousness or persistent neural activity. However, manipulating information in WM during rotation required both. Thus, non-conscious, activity-silent maintenance is a genuine phenomenon but should be termed "activity-silent short-term memory"; when using a memory, a reactivation, associated with conscious reportability, is necessary.

Author contributions: D.T., S.M., and S.D. designed research; D.T. performed research; H.U. contributed new reagents/analytic tools; D.T. analyzed data; and D.T., S.M., H.U., and S.D. wrote the paper.

Reviewers: F.P.d.L., Radboud University Nijmegen; L.M., Max Planck Institute for Empirical Aesthetics; and D.S., Basque Center on Cognition, Brain and Language.

The authors declare no conflict of interest.

Published under the PNAS license.

[1]To whom correspondence may be addressed. Email: darinkat87@gmail.com or stanislas.dehaene@cea.fr.

[2]Deceased January 11, 2019.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1820730116/-/DCSupplemental.

($n = 30$), always employing the same task (Fig. 1). On each trial, a target square in gray (barely visible target-present trials, 80%) or black ink (target-absent control condition, 20%) was flashed in 1 of 24 locations, then masked. Halfway through a 3-s delay period, a cue instructed participants to maintain the original target location (no-rotation condition) or to mentally rotate it 120° clockwise or counterclockwise (rotation condition). Subjects had to comply with these instructions even if they had not seen the target: They were asked to guess the correct final response location if necessary. Participants then rated their subjective visibility of the target using the classical perceptual awareness scale (19), ranging from 1 (no perception whatsoever) to 4 (clearly seen).

Behavioral Evidence for Mental Rotation of Non-conscious Stimuli. We first quantified the extent to which subjects could detect, maintain, and manipulate targets in the behavioral experiment. Participants varied their visibility ratings as a function of target presence, reporting the majority of target-absent trials as unseen (visibility = 1; 88.1 ± 3.1%) and approximately two-thirds of the target-present trials as seen (visibility > 1; 67.7% ± 3.5%). Target detection d' exceeded chance [2.0 ± 0.1; t (22) = 13.2, $P < 0.001$]. Spatial position of the target also influenced subjective visibility reports: More targets were detected when they were presented along the horizontal (73.5 ± 3.5%; i.e., positions 4 through 9 and 16 through 21 in *SI Appendix*, Fig. S1*A*) rather than along the vertical axis [60.0 ± 4.1%; t (22) = 3.8, $P < 0.001$; i.e., positions 10 through 15 and 22 through 3 in *SI Appendix*, Fig. S1*A*]. Crucially, task (no-rotation vs. rotation) did not modulate subjects' visibility [task × target presence × visibility interaction: $F(1, 22) = 3.2$, $P = 0.088$], suggesting that participants used the rating scale similarly in both tasks.

Forced-choice localization performance corroborated this interpretation. On seen trials in the no-rotation condition, accuracy was relatively high (65.8 ± 2.5%; chance = 1/24 = 4.17%) and increased monotonically from glimpsed (visibility = 2) to clearly seen targets (visibility = 4; pairwise comparisons: $P < 0.05$, except for the comparison between visibility 2 and 3, where $P = 0.296$; *SI Appendix*, Fig. S2 *A*, *Top*). Accuracy remained high on seen rotation trials (30.1 ± 1.9%), albeit lower than on no-rotation trials [t (22) = 12.3, $P < 0.001$] and without a clear increase as a function of visibility (pairwise comparisons: $P > 0.180$; *SI Appendix*, Fig. S2 *A*, *Bottom*). Most crucially, even on the unseen trials, performance was well above chance for the no-rotation and rotation task, irrespective of rotation direction (*SI Appendix*, Table S1).

Subjects' responses surrounded the correct location, yet with greater spread after rotation than no-rotation trials (Fig. 2*A* and *SI Appendix*, Fig. S2*A*). We quantified the rate of approximately correct responding (i.e., correct location ± 30°) and estimated the precision of genuine representations (as opposed to random guesses) held in WM (i.e., SD within this tolerance interval, having accounted for random guessing; see *Methods* and refs. 9 and 10). Spatial position influenced participants' ability to identify the correct response location, with subjects' responses falling in the region of correct responding more frequently when the target had been presented along the horizontal (83.5 ± 1.7%) rather than along the vertical axis [74.3 ± 3.6%; t (22) = 2.9, $P = 0.008$; *SI Appendix*, Fig. S1*B*]. Moreover, both task [$F(1, 22) = 9.9$, $P = 0.005$] and visibility [$F(1, 22) = 151.1$, $P < 0.001$] affected the rate of correct responding. Participants' responses fell near the correct location more often in the no-rotation (76.5 ± 2.4%) than in the rotation condition (69.4 ± 2.4%), and when having seen (94.1 ± 1.0%) rather than when not having seen the target square (51.9 ± 3.8%). These factors did not interact [$F(1, 22) = 0.2$, $P = 0.657$; Fig. 2*A*, *Top Inset*], indicating that decrements in performance following a mental rotation were comparable for seen and unseen targets.

Analysis of precision reinforced this conclusion: Out of 23 subjects, 19 displayed above-chance blindsight (i.e., genuine maintenance in WM compared with random guessing) across both rotation directions (chance = 20.83%; $P < 0.05$ in a χ^2 test) and were included here. Task [$F(1, 18) = 34.9$, $P < 0.001$] and visibility [$F(1, 18) = 10.3$, $P = 0.005$] again influenced localization performance but this time also interacted [$F(1, 18) = 8.9$, $P = 0.008$]. Rotating the target decreased the precision of participants' responses for seen [t (18) = −11.9, $P < 0.001$] and unseen targets [t (18) = −2.3, $P = 0.031$], but this reduction was stronger for seen than unseen trials [t (18) = −3.0, $P = 0.008$; Fig. 2 *A*, *Bottom Inset*]. There was therefore no observable detriment to rotating an unseen location.

We replicated these observations in the MEG experiment. Subjects employed the visibility scale meaningfully, rating target-present trials primarily as seen (64.6 ± 3.2%) and target-absent trials as unseen [83.6 ± 2.5%; detection d': 1.7 ± 0.1, t (29) = 14.2, $P < 0.001$] in both tasks [task × target presence × visibility interaction: $F(1, 29) = 2.1$, $P = 0.159$]. There was again a slight bias toward preferentially seeing targets on the horizontal (69.1± 3.5%) rather than the vertical axis [51.5 ± 3.5%; t (29) = 6.8, $P < 0.001$; *SI Appendix*, Fig. S1*A*]. Localization accuracy for seen targets was modestly high in the no-rotation condition (57.5 ± 2.2%; *SI Appendix*, Fig. S2 *B*, *Top*) and reduced following a mental rotation [27.1 ± 1.6%, t (29) = 14.3, $P < 0.001$; *SI Appendix*, Fig. S2 *B*, *Bottom*]. We again observed a long-lasting blindsight effect in both tasks and for all rotation directions (*SI Appendix*, Table S1). Targets presented along the horizontal axis (73.4 ± 2.7%) were identified correctly more frequently than their counterparts along the vertical axis [62.43 ± 3.8%; t (29) = 4.5, $P < 0.001$; *SI Appendix*, Fig. S1*B*]. Task and visibility influenced the rate of correct responding [main and interaction effects: all Fs (1, 29) > 4.8, all Ps < 0.036] and precision [$n = 27$; main and interaction effects: all Fs (1, 26) > 8.3, all Ps < 0.008]. Mental rotation decreased participants' performance on seen trials [t (29) = 5.0, $P < 0.001$], but this effect did not reach significance on unseen trials [t (29) = 1.8, $P = 0.090$; Fig. 2 *B*, *Top Inset*]. Moreover, it also reduced precision more for seen [t (26) = −15.9, $P < 0.001$] than for unseen targets [t (26) = −3.9, $P < 0.001$; Fig. 2 *B*, *Bottom Inset*].

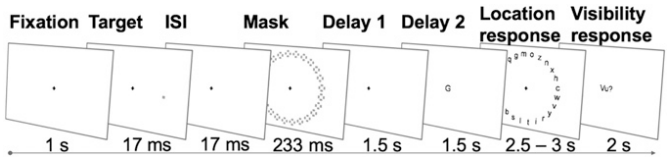

Fig. 1. Experimental design. Participants completed a spatial delayed-response task. On each trial, a faint target was flashed in 1 of 24 locations and masked. A letter cue presented halfway through a 3-s delay period signaled the specific task. (*i*) Following an equal sign (=), subjects were to report the exact location in which the target had appeared. (*ii*) The letter *D* indicated a 120° clockwise and (*iii*) the letter *G* a 120° counterclockwise rotation with respect to the target position. Last, participants rated their subjective visibility of the target on a four-point scale.

Combining the data from both experiments confirmed the above conclusions (*SI Appendix*, *Supplementary Results* and Fig. S3): Even when they failed to perceive the target, subjects succeeded in manipulating it. However, there exist at least three possible explanations for this long-lasting blindsight effect. First, it may have been the product of a genuine non-conscious manipulation. Second, it may have resulted from a fraction of seen trials miscategorized as unseen; this interpretation, although rejected in our previous experiment without rotation (9), needs to be reexamined here. Third, subjects may have recovered the information from non-conscious WM at or before the time of the cue, transformed it into a conscious, active representation, and thereafter consciously manipulated this early guess. To resolve these possibilities, we turned to our MEG data, focusing on five a priori time windows: early brain responses (0.1 to 0.3 s), the P3b time window critical for conscious perception (0.3 to 0.6 s), the delay period before (0.6 to 1.76 s) and after (1.76 to 3.26 s) the rotation cue, and the response period (3.26 to 3.5 s).

Long-Lasting Blindsight Does Not Arise from Miscategorization of Seen Trials. Above-chance objective performance for unseen targets could have resulted from the erroneous mislabeling of some seen targets as unseen. In this case, the unseen correct trials should display the same neural signatures of conscious processing as seen trials (9). There should be an amplification of brain activity during the P3b time window, and a classifier trained to distinguish accuracy on the unseen trials should resemble a standard visibility decoder (i.e., seen vs. unseen). By contrast, the classification of seen vs. unseen correct trials should produce a different pattern of results or fail entirely.

To evaluate this miscategorization hypothesis, we first examined which patterns of brain activity distinguished seen (visibility > 1) from unseen (visibility = 1) trials. This analysis revealed typical markers of conscious processing (9, 20). Seen targets elicited a strong positive response between ∼300 and 600 ms in right-lateralized centroparietal sensors, corresponding to activations in occipital, temporal, parietal, and dorsolateral prefrontal brain areas ($p_{clust} = 0.011$; Fig. 3*A*). Brain activity was amplified during the P3b time

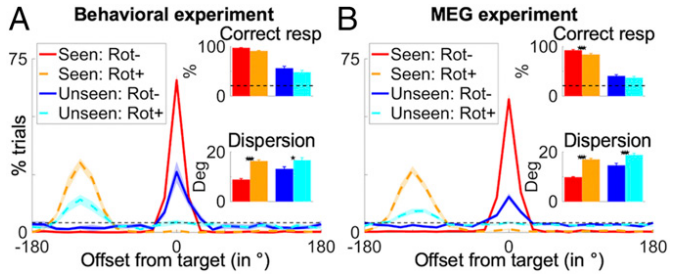

Fig. 2. Behavioral evidence for manipulation of non-conscious information in the behavioral (*A*) and MEG (*B*) experiment. Panels depict distributions of participants' localization responses with respect to target location (0°; positive displacement = counterclockwise offset) as a function of task (no-rotation = solid line, rotation = dotted line) and visibility (seen = warm colors, unseen = cool colors). (*Top Inset*) The rate of correct responding and (*Bottom Inset*) the precision of WM representations in all subjects with sufficient blindsight. Horizontal dotted lines index chance at 4.17% (for single locations) and 20.83% (for the region of correct responding), respectively. Shaded area and error bars represent the SEM across participants. $*P < 0.05$, $**P < 0.01$, and $***P < 0.001$ in a post hoc paired samples *t* test following a significant interaction.

window (i.e., ∼292 and 576 ms; $p_{uncorrected} < 0.05$), although differences with unseen targets also persisted between ∼964 and 1,320 ms ($p_{uncorrected} < 0.05$; Fig. 3*B*). Rotation and no-rotation trials did not differ (task × visibility interaction: $p_{clust} > 0.280$).

When contrasting the unseen correct (i.e., within ±30° of the correct response location) with the unseen incorrect epochs, we observed no evidence for a miscategorization. No significant differences emerged ($p_{clust} > 0.221$) and there was no sign of amplified brain activity (*SI Appendix*, Fig. S4*A*), even when considering the time courses in channels most sensitive to divergences for seen and unseen targets (*SI Appendix*, Fig. S4*B*). Bayesian statistics provided substantial evidence in favor of the null hypothesis (i.e., no difference in MEG amplitude between unseen correct and incorrect trials) for all time windows (all Bayes' factors < 0.24).

Because chance corresponded to 20.83% (i.e., 5/24 positions), a non-negligible portion of the unseen correct trials might have resulted from guessing, potentially obscuring differences between unseen correct and incorrect epochs. To address this possibility, we next estimated neural activity for unseen correct epochs after correction for chance responding (21). If these chance-free unseen correct trials resulted from a miscategorization of seen epochs, we should now observe clear signatures of conscious processing. This was not the case. Chance-free brain activity was still indistinguishable from the one on unseen incorrect and unseen correct trials (whole brain: all $p_{clust} > 0.252$; critical time courses: all Bayes' factors < 0.76). Moreover, it remained strikingly different from a synthetic waveform, derived by proportionally mixing the signals from seen and unseen incorrect trials (as would be expected under the miscategorization hypothesis; *SI Appendix*, Fig. S4*B*). The miscategorization hypothesis can therefore be rejected.

Decoding analyses confirmed this conclusion. Training a linear multivariate pattern classifier to discriminate seen from unseen trials resulted in above-chance diagonal decoding from ∼120 ms onwards [all $p_{clust} < 0.05$; time bins: areas under the curve (AUCs) > 0.54, $ps_{corr} < 0.005$; Fig. 3 *C*, *Top*]. The temporal generalization of each time-specific estimator to all other time points supported this picture (Fig. 3 *C*, *Bottom*): Visibility decoding was primarily confined to a thick diagonal, indicating that conscious perception was associated with a chain of metastable patterns of brain activity. Similar findings emerged when training and testing a visibility classifier separately in the no-rotation and rotation condition, or when generalizing from one task to the other (*SI Appendix*, Fig. S5). Multivariate neural signatures of conscious perception were thus stable across experimental tasks and in line with previous observations (9, 22).

Crucially, we found no discernable pattern when classifying unseen correct vs. unseen incorrect trials (all $p_{clust} > 0.05$; time bins: AUCs < 0.51, $ps_{corr} > 0.05$; Bayes' factors < 0.28; *SI Appendix*, Fig. S4*C*). However, training a classifier to distinguish the seen from the unseen correct epochs resulted in a similar, albeit weaker, decoding time course and generalization matrix as when directly training on all unseen or even just the unseen incorrect trials (time bins: AUCs > 0.52, all $ps_{corr} < 0.05$; Bayes' factors > 2.07; *SI Appendix*, Fig. S6). As such, this pattern of results is exactly opposite to what one would have expected in the case of a miscategorization and suggests that information was genuinely encoded in non-conscious WM.

Long-Lasting Blindsight Effect Results from an Active, Conscious Process. What process allowed participants to mentally rotate an unseen target? Was it a genuine non-conscious manipulation, or did subjects first reinstate an active, conscious representation of the estimated target position around the time of the cue and then rotate this conscious guess? Disambiguating between these alternatives first requires the identification of a neural marker of active, conscious processing. Prior work has identified a rhythmic signal—a suppression of power in the alpha (8 to 12 Hz) and low (13 to 20 Hz) and high beta frequency bands (20 to 27 Hz)—as a possible reflection of such a cognitive state (9, 23, 24).

We observed this signature in the current task. Across all trials (both no-rotation and rotation), there was a prominent desynchronization in alpha/beta frequencies over an extensive set of central sensors, emanating primarily from parietal brain sources (Fig. 4*A*). This desynchronization was reliably modulated by visibility. Power decreased more strongly on seen than on unseen trials between ∼580 and 1,320 ms in the alpha ($p_{clust} = 0.032$) and between ∼460 and 1,300 ms in the low beta band ($p_{clust} = 0.046$; Fig. 4 *B*, *Top*). Similarly, desynchronizations were more pronounced for seen than for target-absent epochs in the low ($p_{clust} = 0.015$) and high beta bands ($p_{clust} = 0.030$) between ∼280 and 940 and ∼820 and

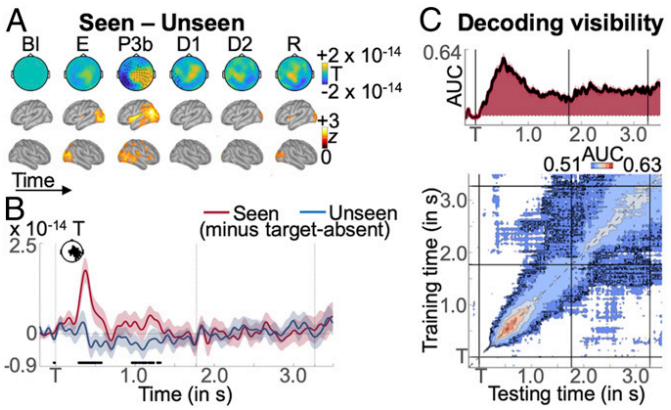

Fig. 3. Typical neural signatures and dynamics of conscious processing for seen targets. (*A*) Sequence of brain activations evoked by seen targets in sensor (*Top*) and source space (*Bottom*). Each topography depicts the difference in amplitude between seen and unseen trials averaged over the time window shown (i.e., BI = –0.2 to 0 s, E = 0.1 to 0.3 s, P3b = 0.3 to 0.6 s, D1 = 0.6 to 1.76 s, D2 = 1.76 to 3.26 s, *R* = 3.26 to 3.5 s; magnetometers only). Sources reflect *z*-scores of absolute difference with respect to a pre-stimulus baseline. Black asterisks indicate sensors showing a significant difference between seen and unseen trials at any point during the respective time window as assessed by a Monte Carlo permutation test. (*B*) Average time courses of seen (red) and unseen (blue) trials in that subset of magnetometers having shown a significant effect in *A*. Shaded area illustrates SEM across subjects. Significant differences between conditions are depicted with thick black line (two-tailed Wilcoxon signed-rank test, uncorrected). Vertical dotted lines index onset of the target (T), cue, and response screens. For display purposes only, data were low-pass-filtered at 8 Hz. (*C*, *Top*) Average time course of diagonal decoding of visibility (i.e., seen vs. unseen). Shaded area denotes above-chance decoding as assessed by a one-tailed cluster-based permutation analysis. Horizontal, dotted line represents chance level at 50%. (*C*, *Bottom*) Temporal generalization matrix of the same visibility decoder. Each horizontal row in the matrix corresponds to an estimator trained at time *t* and tested on all other time points *t*′. The diagonal gray line demarks classifiers trained and tested on the same time points (i.e., the diagonal estimator shown at the top). The thick black outline indexes above-chance decoding as evaluated by a two-tailed cluster-based permutation test. In both plots, vertical lines mark onset of the target (T), cue, and response screens. For display purposes, data were smoothed with a moving average of five samples (i.e., 40 ms).

2,000 ms. There were no differences in the power profiles between (*i*) unseen and target-absent trials (all $p_{clust} > 0.250$) and (*ii*) unseen correct and incorrect epochs (all $p_{clust} > 0.280$; Fig. 4 *B*, *Bottom*), suggesting that, irrespective of accuracy, all unseen trials resembled those without a target. In line with our previous results (9), we thus interpret the observed alpha/beta desynchronization as a potential correlate of active, conscious processing.

We are now in a position to evaluate the remaining alternatives. If the long-lasting blindsight effect resulted from a genuine, nonconscious (activity-silent) form of rotation, alpha/beta desynchronization for unseen targets should remain lower than its counterpart on seen trials throughout the entire epoch. By contrast, if participants consciously rotated a guess, using a normal process of active mental rotation, a more pronounced surge of desynchronization should be seen after the cue for unseen than for seen targets, rendering seen and unseen trials indistinguishable from this point forward. Differences in desynchronization between seen and unseen/target-absent trials should only exist during the pre-cue phase. Note that, whichever of these post-cue patterns is to be observed, it may be present on both no-rotation and rotation trials: No-rotation cues occurred on only a minority of one-third of trials. As such, subjects could have capitalized on the predictable temporal structure of the task with a fixed delay between the target stimulus and the rotation cue to make a conscious guess in anticipation of the upcoming cue, without yet knowing whether it would signal a rotation or no-rotation. We do thus not necessarily expect there to be any differences in the post-cue effects between no-rotation and rotation trials and, as such, conducted the following analysis on all trials.

Our results support the active, conscious manipulation hypothesis (Fig. 4*C*). Following an initial divergence during the early pre-cue maintenance phase (*SI Appendix*, Fig. S7 *A–C*), differences in spectral profiles between seen, unseen, and target-absent trials vanished by ∼1 s. All epochs in the no-rotation and rotation condition were characterized by a prominent, sustained desynchronization in the alpha and low and high beta frequencies. This suppression in power varied as a function of subjective visibility (i.e., seen vs. unseen) and time (i.e., pre-cue vs. post-cue delay). It was much more pronounced during the post-cue than during the pre-cue maintenance period (i.e., main effect of time: all Fs > 18.6, all Ps < 0.001). Crucially, this difference between pre- and post-cue power was also larger for unseen than for seen targets in the alpha and low beta bands (interaction: all Fs > 4.01, all Ps ≤ 0.05), and marginally so in the high beta band [interaction: $F(1, 29) = 2.95$, $P = 0.097$; Fig. 4*D*]. No such interaction emerged when contrasting the unseen correct with the unseen incorrect trials (i.e., interaction: all Fs < 2.83, all Ps > 0.103; Fig. 4*D*), as these conditions displayed largely similar power profiles throughout the entire epoch (*SI Appendix*, Fig. S7 *D–F*). Note that all of these observations also held when restricting our analysis just to the rotation trials [seen vs. unseen: all interaction Fs (1, 29) > 6.5, all Ps < 0.016; unseen correct vs. unseen incorrect: all interaction Fs (1, 29) < 0.3, all Ps > 0.592].

We thus observed a reliable distinction between seen and unseen brain states only during the maintenance period preceding the rotation cue (up until at least 1 s). Unseen targets were accompanied by a significantly smaller desynchronization in the alpha and low and high beta frequencies, but this difference disappeared ∼500 ms before the presentation of the symbolic rotation cue (at 1.5 s) and persisted throughout the post-cue delay. At this critical point in the task then, when on the majority of the trials a mental rotation was required, unseen trials appeared to be indistinguishable from seen trials. The fact that this alpha/beta desynchronization following unseen targets appeared even before the symbolic rotation cue might either have resulted from temporal smoothing inherent to time-frequency analyses, or, as speculated before, from an active process of temporal anticipation of the cue encouraged by the fixed delay between target and cue.

Inasmuch as our analysis is sensitive enough to pick up potentially weak differences in power, these results thus suggest that, irrespective of whether they were simply maintaining or rotating the location of the target, subjects were engaged in a comparable and active mental process for seen and unseen targets after the presentation of the cue. Together with (*i*) the observed pre-cue characteristics of this alpha/beta desynchronization rendering it a plausible signature of conscious processing and (*ii*) our previous observation that, in the context of pure maintenance, no such alpha/beta desynchronization emerges for unseen targets (9), the present findings are compatible

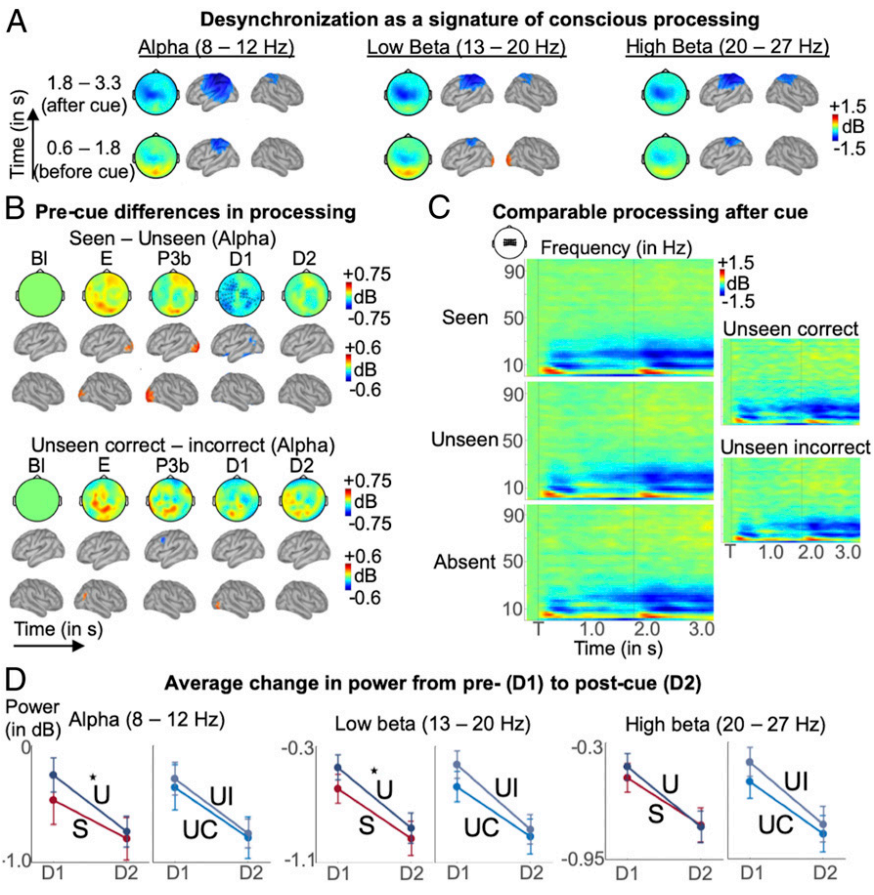

Fig. 4. Time-frequency markers of conscious processing emerge around the time of the cue on unseen trials. (*A*) Average pre-cue (0.6 to 1.8 s; *Bottom*) and post-cue (1.8 to 3.3 s; *Top*) desynchronization in the alpha (8 to 12 Hz; *Left*), low beta (13 to 20 Hz; *Middle*), and high beta (20 to 27 Hz; *Right*) frequency bands in magnetometers and source space (in decibels, relative to pre-stimulus baseline). (*B*, *Top*) Alpha band activity (8 to 12 Hz) related to consciously perceiving the target square (i.e., seen vs. unseen) is shown in magnetometers and source space (in decibels, relative to pre-stimulus baseline) as a function of time bin (i.e., Bl = −0.2 to 0 s, E = 0.1 to 0.3 s, P3b = 0.3 to 0.6 s, D1 = 0.6 to 1.76 s, D2 = 1.76 to 3.26 s). Black asterisks denote cluster of sensors displaying a significant difference at any point in time during the respective time window (as evaluated by a Monte Carlo permutation test). (*B*, *Bottom*) Same as *B*, *Top*, but for the contrast between unseen correct and unseen incorrect trials. (*C*) Average time-frequency power relative to baseline as a function of visibility and target presence in a subset of central magnetometers. Vertical lines demark onset of target (T) and cue presentation. (*D*) Plots depict average pre-cue (D1) and post-cue (D2) power in the same group of sensors as in *C* as a function of frequency (i.e., alpha, low beta, and high beta) and visibility [i.e., seen (S), unseen (U), unseen correct (UC), and unseen incorrect (UI)]. Error bars represent SEM across subjects. Asterisks denote significant interaction in a repeated-measures ANOVA at $P < 0.05$.

with the notion that participants solved the mental rotation task by reinstating an active, conscious estimate of the memorized target location. Additional multivariate analyses further support this interpretation (*SI Appendix*, *Supplementary Results* and Fig. S8).

The Location of Unseen Targets Can Only Be Tracked Transiently. To test this conclusion further, we used multivariate decoding to track the neural representations of the (estimated) target and response locations during the encoding, maintenance, manipulation, and retrieval of seen and unseen targets. We first trained a regression model to predict target angle from participants' whole-brain activity separately for each time point. To avoid spatial confounds and maximize statistical power, we fitted estimators on target-present trials across all rotation and visibility conditions. Given our choice of rotation angles (−120°, 0°, and +120°), the correct response location was strongly decorrelated from the target location only when pooling across all trials. We then evaluated model performance on left-out subsets of epochs (*Methods*). Note that none of the findings changed qualitatively when testing separately on the rotation and no-rotation task (*SI Appendix*, Fig. S9) and that eye movements did not contribute significantly to these results (*SI Appendix*, *Supplementary Results* and Fig. S10).

Starting at ∼80 ms, estimator performance for seen targets steadily rose until ∼264 ms and then slowly decayed toward chance at ∼1.46 s (Fig. 5*A*). Following the cue, a rebound of position-selective activity was observed and was then fairly sustained for the remainder of the trial ($p_{clust} < 0.05$; time bins: Ws > 417.0, $ps_{corr} < 0.005$, Bayes' factors > 77.93). Thus, in line with previous findings (9), seen targets were initially encoded via active firing, then this activity decayed and was reactivated throughout most of the post-cue delay period.

A different picture emerged for unseen targets. While target location was again actively stored early on, this representation was weaker than the one for seen targets (paired-samples Wilcoxon signed rank test: pre-cue time bins: Ws > 370.0, $ps_{corr} < 0.02$, Bayes' factors > 3.42) and decayed quickly, vanishing entirely by ∼920 ms ($p_{clust} < 0.05$; pre-cue time bins: Ws > 351.0, $ps_{corr} < 0.035$, Bayes' factors > 7.34). During the post-cue delay, although not evident in the actual decoding time course ($p_{clust} > 0.05$; Fig. 5*A*), the estimator's performance over the entire time window remained above chance (rads = 0.03 ± 0.01, $W = 355.0$, $p_{corr} = 0.025$, Bayes' factor = 6.41) and at levels comparable to those on seen trials ($W = 315.0$, $p_{corr} = 0.460$, Bayes' factor = 0.86). A more fine-grained analysis with a moving average of 100 ms revealed that this effect was driven primarily by the initial phase of the delay, up to ∼2.6 s. We observed no modulation of this pattern of findings by accuracy (time bins: Ws < 279.0, all $ps_{corr} > 0.950$, Bayes' factors < 0.41; Fig. 5 *A*, *Insets*).

In summary, whereas seen targets were maintained with persistent, albeit decaying, activity, unseen targets elicited weaker position-related activity that quickly decayed to baseline level (activity-silent WM). During the post-cue phase, once participants actively maintained or manipulated their WM contents, the representation of seen targets was reactivated and sustained. Unseen targets may also have benefitted from a short-lived revival, but their decoding was much weaker, perhaps because participants made more errors and therefore reinstated the actual target location on only a subset of unseen trials.

An Estimate of the Location of Unseen Targets Is Reinstated before the Rotation Cue. On more than half of unseen trials (62.0 ± 2.8%), subjects chose an incorrect location. What determined participants' final response on those trials? As laid out in the introduction, if target locations really had been stored in an activity-silent format, we hypothesize that, around the time of mental rotation, these should have been reactivated (i.e., reinstated into active neural firing), albeit with occasional errors, and subjects should then have attempted to consciously rotate this guess. According to this hypothesis, around the time of the cue, brain signals should contain a decodable representation of the "pre-rotation location", that is, the spatial location that, given the subjects' response, would have been the location guessed and rotated. On no-rotation trials, this location coincided with response location, whereas on rotation trials it corresponded to the position of participants' response rotated 120° in the direction opposite to what the cue had instructed. Detecting the presence of such a pre-rotation representation on unseen trials would support our time-frequency analyses and the hypothesis that, around the time of the cue, subjects attempted to recover a conscious representation of the target and then consciously rotated this guess. If, however, unseen performance was based on a genuine manipulation of activity-silent WM without the intermediate step of reactivation, then such decoding should fail.

On seen trials, decoding the pre-rotation location was possible, and the time course was similar to the one for the true position of the target (unsurprisingly given that most seen trials were correct; Fig. 5*B*). From ∼56 ms onwards, the pre-rotation location was coded in activity-based brain states ($p_{clust} < 0.05$; time bins: Ws > 408.0, $ps_{corr} < 0.005$, Bayes' factors > 517.26), first peaking at ∼264 ms (rad = 0.18 ± 0.02) and then slowly decaying before being revived by the rotation cue.

Crucially, pre-rotation location could also be decoded on unseen trials. Shortly after the target, the estimator's performance began to rise and first exceeded chance at ∼376 ms (rad = 0.052 ± 0.015). Decoding persisted until ∼1.8 s ($p_{clust} < 0.05$; P3b time window and pre-cue delay: Ws > 382, $ps_{corr} < 0.005$, Bayes' factors > 78.83), although estimator performance itself did not drop until ∼2.5 s. Indeed, a follow-up analysis with narrower 100-ms time windows suggested that the pre-rotation location may have been maintained until ∼2.2 s ($P < 0.05$, uncorrected). There was again no evidence for a modulation of this pattern as a function of accuracy (time bins: Ws > 120.0, $ps_{corr} > 0.600$, Bayes' factors < 1.44; Fig. 5 *B*, *Insets*).

As predicted, while the representation of the pre-rotation location was stronger for seen than for unseen targets early on (early and P3b time window: Ws > 450.0, $ps_{corr} < 0.005$, Bayes' factors > 124,688.30), this difference started to diminish during the pre-cue maintenance phase ($W = 347.0$, $p_{corr} = 0.085$, Bayes' factor = 1.76) and vanished entirely by the last second before the rotation cue (moving average of 100 ms: Ws < 359.0, $ps_{corr} > 0.05$, Bayes' factor < 1.32). Participants' location estimates were therefore similarly represented on both seen and unseen trials during the last part of the pre-cue maintenance period. In conjunction with the results from the time-frequency analyses (i.e., emergence of signatures of conscious processing for unseen targets), these findings are compatible with the proposal that, even on unseen trials, subjects mentally rotated an active, conscious guess of a target location.

An Active Representation of Target Location Is Mentally Rotated. We last trained and tested a regression model to decode response location. On seen trials, response location emerged reliably only in the second half of the post-cue delay period (Fig. 5*C*). Starting at ∼2.38 s, decoding performance gradually built up until its peak at the very end of the epoch ($p_{clust} < 0.05$; post-cue time bins: Ws > 440.0, $Ps_{corr} < 0.005$, Bayes' factors > 21,997.68). There was substantial temporal overlap between the decoding of the target/pre-rotation location and the response position: As the former started to decay around ∼2.5 s, the latter slowly began to rise.

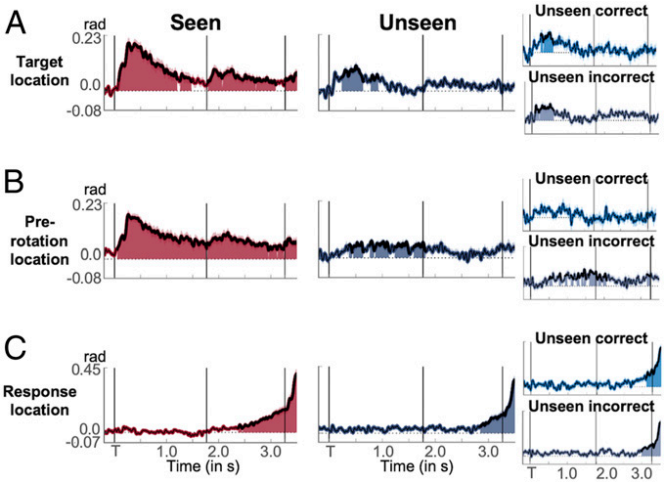

Fig. 5. Tracking a mental rotation. (*A*) Time courses of average decoding of target location on seen (red), unseen (dark blue), unseen correct (light blue), and unseen incorrect (blue) trials. Shaded areas represent above-chance performance as assessed by a one-tailed cluster-based permutation test.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

Fig. 6 further shows the probability density distributions for decoded target and response locations. On seen trials, before the cue, decoder estimates for target angle were strongly concentrated around the target location, irrespective of rotation condition and direction (resultant vector lengths > 0.41; Rayleigh tests for nonuniformity: zs > 5.09, Ps < 0.005; nonparametric multisample test for equal medians: Ps > 0.302). This picture changed following the rotation cue. While angle estimates on no-rotation trials stayed fairly centered on the original target location (resultant vector lengths > 0.37; Rayleigh test: z > 4.01, P < 0.017), their counterparts for clockwise and counterclockwise rotations began to shift toward the respective correct response positions (response period: clockwise rotation: M_{circ} = 37.3°; resultant vector length = 0.49; one-sample test against a mean direction of 0°: $P < 0.05$; counterclockwise rotation: M_{circ} = 95.6°; resultant vector length = 0.31; one-sample test against a mean direction of 0°: $P < 0.05$). During the response period, all three distributions were characterized by a different center of mass (nonparametric multisample test for equal medians: Ps < 0.05), located in close proximity to the expected final position. Depending on the direction of the rotation, the representation of the original target location was progressively transformed into a representation of the response position. On average, a mental rotation following seen targets was reflected by an active transition period, during which the stimulus code was progressively replaced by the response code. Note, however, that while such a smooth transition was visible in the mean we cannot determine here whether continuous or discrete transitions occurred on individual trials (25).

We next considered the unseen trials. If subjects performed a conscious rotation of (an estimate of) unseen locations, then one would predict the response estimator to perform comparably on seen and unseen targets. This was indeed the case (Fig. 5*C*). Decoding response location on unseen trials yielded consistent above-chance performance from ∼2.84 s onwards ($p_{clust} < 0.05$; post-cue time bins: Ws > 410.0, $ps_{corr} < 0.005$, Bayes' factors > 594.74), again beginning to rise around the time the model for the pre-rotation location had faded (cf. time courses in Fig. 5 *B* and *C*). As would be expected if the same underlying process were responsible for the generation of responses across all experimental conditions, we observed no differences as a function of accuracy (time bins: Ws < 314.0, $ps_{corr} > 0.480$, Bayes' factors < 0.81) or visibility (time bins: Ws < 334.0, $ps_{corr} > 0.600$, Bayes' factors < 2.45). Pre-rotation and response locations could also be tracked on unseen trials, albeit, as expected, with reduced accuracy (*SI Appendix*, Fig. S11). The transformation from one representation into another therefore appeared to have been comparable for seen and unseen targets, in both cases relying on decodable activity patterns rather than on activity-silent brain states.

Discussion

Recent work has challenged classical views of WM as a purely conscious process based on persistent neural firing. Information may also be stored in non-conscious and/or activity-silent WM, without any accompanying neural activity, via slowly decaying changes in synaptic weights (9, 12–16), and in the absence of subjective awareness (8–10, 26). However, in previous research, only the short-term maintenance of information has been explored, while its mental manipulation, a key feature of WM, has been ignored.

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

Manipulation as a Limit for Non-conscious Silent Processes. It has proven difficult to put upper bounds on the depth of non-conscious processing. Non-conscious signals tend to affect a wide range of behaviors and trigger activity in many different brain areas, including the prefrontal cortex (27, 28). Recent work on non-conscious WM has even called into question basic assumptions regarding the nature of non-conscious processes, suggesting that non-conscious stimuli may be maintained much longer than previously thought (7–10, 26).

Our behavioral results support this conclusion, as they provide evidence for non-conscious mental rotation. On unseen trials, subjects reported the correct response position much better than chance after several seconds, irrespective of whether they had to maintain the original target or rotate its position. We replicated this long-lasting blindsight effect in two independent experiments and, as such, seemingly expanded the range of possible non-conscious WM processes (6, 8–10).

Our neural data indicated that visibility reports were genuine. Before the cue, we observed typical markers of conscious processing almost exclusively for seen targets. Brain activity was amplified during the P3b time window (20) and participants' visibility was reliably decodable (7, 9). Moreover, there was a sustained desynchronization of alpha/beta frequency, which became even more pronounced after the rotation cue, thereby coinciding with the most demanding task phase (9, 29, 30). By contrast, for unseen targets, signatures of conscious processing were absent or markedly reduced in comparison with the ones on seen trials early on. There was neither an ignition of brain activity during the P3b time window nor a comparably strong alpha/beta desynchronization. These findings, in line with our previous work (9), show that unseen trials were genuine and did not correspond to a subset of miscategorized seen trials.

However, those neural signatures changed drastically around the time of the cue, suggesting that an estimate of target location was reactivated on all trials in anticipation of the likely following rotation. Slightly before the cue, around ∼1 s, an alpha/beta power desynchronization, previously not present, appeared even for unseen targets and reached similar levels as on seen trials during the post-cue maintenance period. Starting at more or less the same time (i.e., around ∼500 ms), a decodable representation of the pre-rotation location emerged. In conjunction with prior work (9), both findings suggest that participants reinstated an active representation of their best guess about target location, in anticipation of the upcoming rotation cue. Given that alpha/beta desynchronization, in the present and past work (e.g., ref. 9), also appeared as a plausible signature of conscious processing, robustly distinguishing seen from unseen/target-absent trials shortly after target onset, we further hypothesize that, around the same time, this reactivated estimate regained access to consciousness and no longer relied on non-conscious WM processes.

On unseen trials, the weak activity-silent representation of the target may have competed against other ongoing noise fluctuations in the brain, resulting in a mixture of trials where decision was solely based on stochastic events (31) and others biased toward the correct target location. Variability across trials and participants and the temporal smoothing inherent to time-frequency analyses precludes a definitive determination of the exact onset of the alpha/beta desynchronization on unseen trials, but the results indicate that this transition already occurred shortly before the symbolic rotation cue. To save time, given the fixed delay between the presentation of the target and the cue, participants may have guessed the identity of the unseen target shortly before the moment they knew the rotation cue would appear.

In conjunction with previous work (8, 9, 26), these findings highlight the limits of non-conscious WM. While information may be temporarily stored nonconsciously, manipulating items is associated with a reinstatement of a conscious representation. Our results may help circumscribe the boundaries of non-conscious processing. Consciousness has been theorized and empirically demonstrated to be a prerequisite for serial tasks, such as the chaining of mental operations (32). We here observed that such chaining may remain possible even if the initial input was not represented consciously, but only inasmuch as subjects willfully operate on previously non-conscious information. Future research might expand on this work and attempt to more strongly encourage the reliance on non-conscious processing by, for instance, rendering the task cues subliminal.

The Complementarity of Active and Silent Processes in WM. Our data speak to the current debate on the nature of WM representations in the brain. Traditional models emphasize stable, persistent activity as the main candidate mechanism supporting WM (3, 4). Recent investigations point toward a more dynamic view, with the contents of WM being maintained in changing patterns of neural activity or activity-silent brain states (9, 12–14, 16, 33).

Together with our previous work (9), our current results suggest that sustained neural activity and activity-silent mechanisms may accommodate different processes. Storage of information in WM need not require neural activity. Without the manipulation requirement in our task, delay-period activity vanished for unseen and was intermittent for seen targets (9). Such prolonged activity-silent periods occurred less frequently here, perhaps because participants tried to more actively retain information about the target in preparation for the required mental rotation. However, even in the present setting, target-related neural activity first decayed toward chance before being reactivated by the cue.

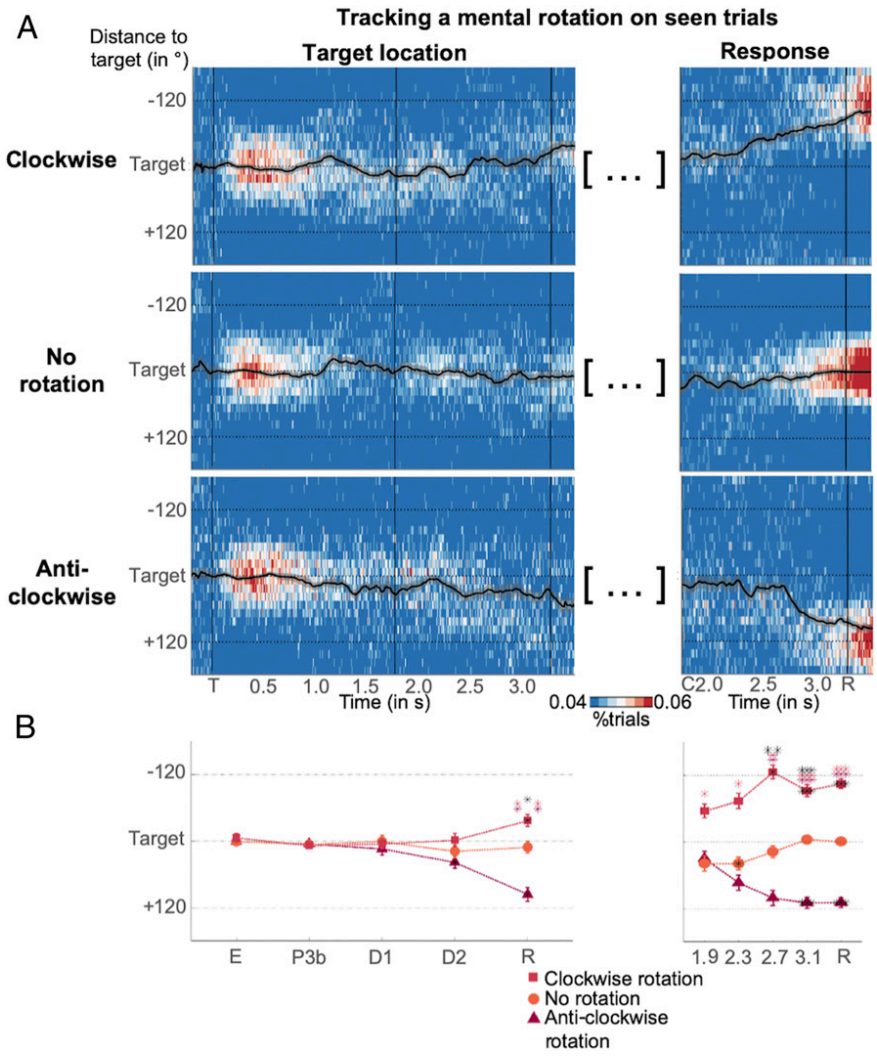

Fig. 6. Tracking a mental rotation on seen trials. (*A*, *Left*) Time courses of probability density distributions of the angular distance between the estimates of a decoder trained with target angle and actual target location are shown as a function of rotation condition. For display purposes, data were smoothed with a moving average of 12 samples (i.e., 96 ms). The overlaid black line illustrates the evolution of the circular mean of the individual distributions (also smoothed). The shaded area reflects circular standard variation across subjects. Vertical event markers denote the onset of the target (T), cue (C), and response (R) screens, and horizontal markers index correct response positions after rotation. *A*, *Right*, same as in *A*, *Left*, except for angular distance between the estimates of a decoder trained with response angle and actual target location. (*B*) Circular means of the above distributions as a function of rotation condition and time bin (i.e., E = 100 to 300 ms, P3b = 300 to 600 ms, D1 = 0.6 to 1.76 s, D2 = 1.76 to 3.26 s, *R* = 3.26 to 3.5 s). Error bars reflect circular SD. Asterisks inside markers denote significant deviation from mean direction of 0 (as assessed by a circular equivalent of a one-sample *t* test) and asterisks on top significant differences in median direction between conditions (as assessed by a circular equivalent to the Kruskal–Wallis test; black = clockwise vs. counterclockwise; red = clockwise vs. no-rotation; violet = counterclockwise vs. no-rotation). $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

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

Taken together, we propose that active and activity-silent processes make distinct contributions to WM. WM maintenance (just like long-term memory storage) can be achieved without accompanying neural activity via activity-silent mechanisms. WM manipulation appears to depend on a reactivation of active neural firing (similar to long-term memory retrieval). Recent evidence from a computational model corroborates this conclusion by demonstrating that, while short-term synaptic plasticity may support short-term maintenance, persistent neuronal activity automatically emerges from learning during active manipulation (34). Moreover, similar divisions of labor between activity-silent and activity-based brain states have recently been observed for the active selection vs. maintenance of WM contents (35). All of these data lend support to the emerging view that WM is best conceptualized as an activity-induced temporary and flexible shift in the functionality of a network (16).

Tracking Intermediate Representations during a Mental Rotation. A last aspect of our work that deserves attention concerns the act of mental rotation itself. Numerous studies support the idea that mental rotation depends on analog spatial representations, with the initial representation progressively being rotated through intermediate steps. Reaction times have been found to increase in near-linear fashion with the size of the rotation angle (36), and activity in spatially mapped brain areas has been reported to be modulated parametrically by angular distance (37). Recordings of single-neuron activity from the motor cortex during a motor rotation task also suggest a gradual rotation of a neural population vector (38).

Our results indicate that such a transformation of neural representations is now decodable from human MEG recordings. On seen trials, following the rotation cue, average decoder estimates of target and response angle progressively moved away from the original target location toward the expected response position, seemingly passing through a series of intermediate locations. A similar transformation may also have been present for the pre-rotation location for unseen targets, although data were too noisy to support any definitive conclusions. These findings are compatible with the view that locations intermediate between the target/pre-rotation position and the response location were coded and represented in the brain. However, this interpretation is based on an analysis of multivariate estimates averaged across trials and participants. Isolated bursts of activity, occurring at different time points and coding for discrete spatial positions, if averaged over many events, might also result in the apparent smooth transition we observed here (39, 40). Future research relying on single-trial analyses will be needed to disambiguate between these alternatives.

Conclusion

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

Methods

Participants. All experimental procedures had been approved by the Ethics Committee on Human Research at NeuroSpin (Gif-sur-Yvette, France). Twenty-three healthy volunteers (4 men; M_{age} = 23 y, SD_{age} = 2.5 y) were included in the behavioral experiment. Another 30 participants (14 men; M_{age} = 25.4 y, SD_{age} = 3.8 y) were entered in the MEG analyses. All subjects gave written informed consent and received up to 80€ as compensation.

WM Task. We adapted our previous paradigm (9) to probe participants' ability to manipulate WM representations under varying subjective visibility (Fig. 1). After a 1-s fixation period, a target square was flashed for 17 ms in 1 of 24 circular locations and then masked (233 ms). Mask contrast was calibrated separately for each subject to yield approximately equal proportions of seen and unseen trials (*SI Appendix, Supplementary Methods*). Halfway through a 3-s delay, a central cue instructed participants as to the specific task. A third of the trials, indexed by an equal sign, served as a control, requiring subjects to maintain the target position. On the remainder of the trials, participants were to mentally rotate the original target location. While the uppercase letter *D* necessitated a 120° clockwise rotation (one-third of the trials), the letter *G* indicated a 120° counterclockwise rotation (1/3 of the trials). Subjects responded by either speaking (MEG experiment; 2.5 s) or typing on a standard AZERTY keyboard (behavioral experiment; 3 s) the letter—out of a set of 24 randomly presented in all possible locations— corresponding to the desired position. A location response was required even when participants had not seen the target; in that case, they were instructed to guess the correct final position. Subjects then rated target visibility on the 4-point Perceptual Awareness Scale (19), using either the number-pad keys of the computer keyboard (behavioral experiment; 2 s) or the buttons of a non-magnetic response box (fiber optic response pad; Cambridge Research Systems Ltd; MEG experiment; 2 s). To qualify as unseen (visibility = 1), participants were neither to have any visual experience of the target nor a hunch concerning its location. All other subjective impressions were to be categorized as seen (visibility 2, 3, or 4). Intertrial intervals ranged between 333 and 666 ms (MEG experiment) or between 1 and 2 s (behavioral experiment). A central fixation cross was shown throughout the entire trial, and 20% target-absent catch trials were included to allow for the isolation of brain activity specific to the target.

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

Behavioral Analyses. We followed our previous approach (9) to evaluate WM performance. Repeated-measures ANOVA was applied to three indices of objective performance. (*i*) Accuracy refers to that proportion of trials that falls onto the correct response location and serves as a measure of the amount of information stored in WM (chance = 1/24, i.e., 4.17%). (*ii*) The rate of correct responding also reflects the quantity of information held in WM but is more refined than accuracy alone, as it accounts for small errors in subjects' ability to identify the correct response location. It was defined as the proportion of trials within ± two positions of the correct response location (i.e., ±30°; chance = 5/24, i.e., 20.83%). (*iii*) As an estimate of the precision of WM representations, we computed the SD of that part of the distribution of participants' spatial responses that corresponded to genuine WM (as opposed to random guessing within the region of correct responding; ref. 9). Only subjects with sufficient blindsight (i.e., $P < 0.05$ in a χ^2 test against chance) when collapsing across all experimental conditions were included in this analysis.

MEG Acquisition and Preprocessing. We recorded participants' brain activity continuously during the WM paradigm with a 306-channel, whole-head magnetometer by Elekta Neuromag. MEG sensors were arranged in 102 triplets, comprised of one magnetometer and two orthogonal planar gradiometers, and MEG signals were acquired at a sampling rate of 1,000 Hz with a hardware band-pass filter between 0.1 and 330 Hz. To allow for offline rejection of artifacts induced by eye movements and heartbeat, we monitored these functions with vertical and horizontal electrooculograms and electrocardiograms. Subjects' head position inside the MEG helmet was inferred at the beginning of each run with an isotrack Polhemus Inc. system from the location of four coils placed over frontal and mastoïdian skull areas.

We adapted Marti et al.'s (22) preprocessing pipeline. First, we identified bad MEG channels visually in the raw signal and then employed MaxFilter software (ElektaNeuromag) to compensate for head movements between experimental blocks, suppress magnetic interference from outside the sensor helmet, and interpolate bad channels (41). We then switched to Fieldtrip for further preprocessing (42). Continuous data were first epoched with respect to target onset (i.e., −0.5 to 3.5 s). The resulting trials were downsampled to 250 Hz and any artifacted epoch was removed by means of a semiautomatic procedure: We visually inspected variance of the MEG signals to identify and reject contaminated epochs. In a last step, we performed independent component analysis separately for each channel type to remove any residual artifacts related to eye movements or cardiac activity.

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

Estimating Chance-Free Brain Activity for Unseen Correct Trials. To account for chance responding on unseen correct trials, we employed a strategy developed by Lamy et al. (21) and first calculated the proportion of unseen correct trials correctly responded to by chance separately for each subject:

$$P_{UC} = ((1-r)/(19/24)) * (5/24), \text{where } P_{UC} = \%\text{UnseenCorrect}_{Chance} \text{ and } r = \text{rate of correct responding.} \quad [1]$$

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

$$A(UC_{Observed}) = P_{UC} * A(\%UC_{Chance}) + (1 - P_{UC}) * A(\%UC_{ChanceFree}) \quad [2]$$

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

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

$$A(UC_{Miscategorized}) = (1 - P_{UC}) * A(Seen_{Observed}) + P_{UC} * A(\%UI_{Observed}). \quad [4]$$

Source Reconstruction. Structural magnetic resonance scans were available for 29 of our 30 subjects, having been acquired with a 3D T1-weighted spoiled gradient recalled pulse sequence (voxel size: 1 × 1 × 1 mm; repetition time: 2,300 ms; echo time: 2.98 ms; field of view : 256 × 240 × 176 mm; 160 slices).

To identify the anatomical locations of the MEG signals in these participants, we first segmented subjects' T1 images into gray/white matter using FreeSurfer and then reconstructed the cortical, scalp, and head surfaces in Brainstorm (43). Coregistration between the anatomical scans and the MEG data were based on participants' head position in the MEG helmet. Subject-specific forward models relied on analytical models with overlapping spheres. Separately for each condition and participant, we modeled neuronal current sources with a constrained weighted minimum-norm current estimate (depth-weighting factor: 0.5). Noise covariance matrices were computed from ~5-min-long empty-room recordings, measured immediately after each subject. Before group analysis, single-trial source estimates were either (*i*) averaged within each subject and condition, transformed into *z*-scores relative to our pre-stimulus baseline (–0.2 to 0 s), rectified, and spatially smoothed over 5 mm or (*ii*) in the case of time-frequency decompositions, transformed into average power in the alpha (8 to 12 Hz) and low (13 to 20 Hz) as well as high beta (20 to 27 Hz) bands with complex Morlet wavelets. We then computed the contrasts of interests and projected the resulting participant-specific source estimates on a generic brain model built from the standard template of the Montreal Neurological Institute. Group averages for spatial clusters of at least 50 vertices and thresholded at 50% of the maximum amplitude are shown for each time window under consideration (cortex smoothed at 60%).

Multivariate Pattern Analysis. We aimed at predicting the identity and/or value of a specific categorical (i.e., visibility or accuracy) or circular (i.e., target, pre-rotation, or response location) variable (y) from single-trial brain activity (X) separately for each participant and time point. Relying on the Scikit-Learn package for MNE 0.15 (44), we (*i*) fitted a linear estimator w to a training subset of X (X_{train}) to isolate the topographical patterns best differentiating our experimental conditions, (*ii*) predicted an estimate of y ($\hat{y}$) from a test set (X_{test}), and (*iii*) compared the resulting predictions to the true value of y either for the entire set of labels (*score*(y, $\hat{y}$)) or a specific subset (*subscore*(y, $\hat{y}$)).

For analyses based on circular data, models were always trained on all available target-present trials. That is, estimators were fitted on all three rotation conditions (i.e., clockwise rotation, no-rotation, and counterclockwise rotation) and both visibilities (i.e., seen and unseen). Performance was then evaluated only for that subset of test trials currently under investigation. For instance, decoding scores for response location on unseen trials were obtained by first training the estimator with data from all target-present trials and then applying this model only to trials with unseen targets. In contrast to an analysis, in which the train and test sets come from the very same subset of trials, this approach offers two main advantages. First, it augments the number of trials available to train the model and, as such, maximizes statistical power and increases the ability to detect small effects. Second, keeping all rotation and visibility conditions also decorrelates the individual representations of the target, pre-rotation, and response locations.

Two main classes of estimators were used: a linear support vector machine (SVM) for categorical and a combination of two ridge regressions for circular data. Whereas the former generated a continuous output in the form of the distance between the hyperplane (w) and the respective sample of y, the latter first separately fit the sine (sin(y)) and cosine (cos(y)) of the spatial position in question and then estimated an angle from the arctangent of the individual predictions [$\hat{y} = \arctan2(\hat{y}_{\sin}, \hat{y}_{\cos})$]. To increase the number of instances available for each circular label, we averaged neighboring spatial locations. Before model fitting, all channel-time features (X) were *z*-score-normalized, and, for any analysis involving SVMs, a weighting procedure was applied to counteract the effects of potential class imbalances.

To avoid overfitting, we embedded this sequence of analysis steps in a 5-fold, stratified cross-validation procedure: For nonindependent training and test sets, estimators were iteratively fitted on four-fifth of the data (X_{train}) and generated predictions for the remaining one-fifth (X_{test}). By contrast, when generalizing from one task to the other (i.e., no-rotation to rotation condition), estimators from each training set were directly applied to the entire test set and the respective predictions averaged. Within the same cross-validation loop, we also evaluated time generalization: Each estimator was first trained at time t and then tested at all other time points, resulting in a square matrix of training time × testing time.

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

Statistical Analysis. All statistics reported refer to group-level analyses. In the case of ERF and frequency data, we (*i*) performed cluster-based, nonparametric t tests with 1,000 Monte Carlo permutations to identify significant spatiotemporal differences between experimental conditions, while simultaneously correcting for multiple comparisons, and (*ii*) additionally present uncorrected outcomes of nonparametric signed-rank tests for follow-up analyses ($P_{\text{uncorrected}} < 0.05$). We relied on the above cluster-based permutation analysis to assess multivariate decoding performance (i.e., categorical data: AUC > 0.5; circular data: rad > 0; 5,000 permutations). Temporal averages over five a priori time bins, corresponding to an early perceptual period (0.1 to 0.3 s), the P3b time window (0.3 to 0.6 s), the maintenance period before (0.6 to 1.76 s) and after the cue (1.76 to 3.26 s), as well as the response (3.26 to 3.5 s), are also provided. Bonferonni correction was applied to these a priori analyses to correct for multiple comparisons ($P_{\text{corr}} < 0.05/5$). When appropriate, we present circular statistics and computed Bayesian statistics based on two- or one-sided t tests ($r = 0.707$).

ACKNOWLEDGMENTS. This work was funded by INSERM, Commissariat à l'énergie atomique (CEA), Collège de France, European Research Council (ERC), and Fondation Roger de Spoelberch. D.T. was funded by a graduate fellowship from the Ecole des Neurosciences de Paris and Fondation Schneider Electric. We gratefully acknowledge Valentina Borghesani, Pedro Pinheiro Chagas, and Fosca Al Roumi for their invaluable daily support and stimulating discussion, and specifically thank Theofanis I. Panagiotaropoulos for helpful comments on a previous version of this manuscript.

1. B. J. Baars, S. Franklin, How conscious experience and working memory interact. *Trends Cogn. Sci. (Regul. Ed.)* **7**, 166–172 (2003).
2. A. Baddeley, Working memory: Looking back and looking forward. *Nat. Rev. Neurosci.* **4**, 829–839 (2003).
3. J. M. Fuster, G. E. Alexander, Neuron activity related to short-term memory. *Science* **173**, 652–654 (1971).
4. J. Kamiński *et al*., Persistently active neurons in human medial frontal and medial temporal lobe support working memory. *Nat. Neurosci.* **20**, 590–601 (2017).
5. F. Bergström, J. Eriksson, Maintenance of non-consciously presented information engages the prefrontal cortex. *Front. Hum. Neurosci.* **8**, 938 (2014).
6. F. Bergström, J. Eriksson, The conjunction of non-consciously perceived object identity and spatial position can be retained during a visual short-term memory task. *Front. Psychol.* **6**, 1470 (2015).
7. J.-R. King, N. Pescetelli, S. Dehaene, Brain mechanisms underlying the brief maintenance of seen and unseen sensory information. *Neuron* **92**, 1122–1134 (2016).
8. D. Soto, T. Mäntylä, J. Silvanto, Working memory without consciousness. *Curr. Biol.* **21**, R912–R913 (2011).
9. D. Trübutschek *et al*., A theory of working memory without consciousness or sustained activity. *eLife* **6**, e23871 (2017).
10. D. Trübutschek, S. Marti, S. Dehaene, Temporal-order information can be maintained in non-conscious working memory. *Sci. Rep.* **9**, 6484 (2019).
11. K. Watanabe, S. Funahashi, Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. *Nat. Neurosci.* **17**, 601–611 (2014).
12. N. S. Rose *et al*., Reactivation of latent working memories with transcranial magnetic stimulation. *Science* **354**, 1136–1139 (2016).
13. M. J. Wolff, J. Jochim, E. G. Akyürek, M. G. Stokes, Dynamic hidden states underlying working-memory-guided behavior. *Nat. Neurosci.* **20**, 864–871 (2017).
14. M. J. Wolff, J. Ding, N. E. Myers, M. G. Stokes, Revealing hidden states in visual working memory using electroencephalography. *Front. Syst. Neurosci.* **9**, 123 (2015).
15. G. Mongillo, O. Barak, M. Tsodyks, Synaptic theory of working memory. *Science* **319**, 1543–1546 (2008).
16. M. G. Stokes, 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends Cogn. Sci. (Regul. Ed.)* **19**, 394–405 (2015).
17. J. Silvanto, Working memory maintenance: Sustained firing or synaptic mechanisms? *Trends Cogn. Sci. (Regul. Ed.)* **21**, 152–154 (2017).
18. S. J. Luck, E. K. Vogel, Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends Cogn. Sci. (Regul. Ed.)* **17**, 391–400 (2013).
19. T. Z. Ramsøy, M. Overgaard, Introspection and subliminal perception. *Phenomenol. Cogn. Sci.* **3**, 1–23 (2004).
20. R. Gaillard *et al*., Converging intracranial markers of conscious access. *PLoS Biol.* **7**, e61 (2009).

21. D. Lamy, M. Salti, Y. Bar-Haim, Neural correlates of subjective awareness and unconscious processing: An ERP study. *J. Cogn. Neurosci.* **21**, 1435–1446 (2009).
22. S. Marti, J.-R. King, S. Dehaene, Time-resolved decoding of two processing chains during dual-task interference. *Neuron* **88**, 1297–1307 (2015).
23. K. C. Backer, M. A. Binns, C. Alain, Neural dynamics underlying attentional orienting to auditory representations in short-term memory. *J. Neurosci.* **35**, 1307–1318 (2015).
24. F. Meyniel, M. Pessiglione, Better get back to work: A role for motor beta desynchronization in incentive motivation. *J. Neurosci.* **34**, 1–9 (2014).
25. K. W. Latimer, J. L. Yates, M. L. R. Meister, A. C. Huk, J. W. Pillow, NEURONAL MODELING. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science* **349**, 184–187 (2015).
26. F. Bergström, J. Eriksson, Neural evidence for non-conscious working memory. *Cereb. Cortex* **28**, 3217–3228 (2018).
27. K. Nakamura *et al*., Neural capacity limits during unconscious semantic processing. *Eur. J. Neurosci.* **47**, 929–937 (2018).
28. B. van Vugt *et al*., The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. *Science* **360**, 537–542 (2018).
29. M. Pessiglione *et al*., How the brain translates money into force: A neuroimaging study of subliminal motivation. *Science* **316**, 904–906 (2007).
30. V. Wyart, C. Tallon-Baudry, How ongoing fluctuations in human visual cortex predict perceptual awareness: Baseline shift versus decision bias. *J. Neurosci.* **29**, 8715–8725 (2009).
31. E. Vul, D. Hanus, N. Kanwisher, Attention as inference: Selection is probabilistic; responses are all-or-none samples. *J. Exp. Psychol. Gen.* **138**, 546–560 (2009).
32. J. Sackur, S. Dehaene, The cognitive architecture for chaining of two mental operations. *Cognition* **111**, 187–211 (2009).
33. M. G. Stokes *et al*., Dynamic coding for cognitive control in prefrontal cortex. *Neuron* **78**, 364–375 (2013).
34. N. Y. Masse, G. R. Yang, H. F. Song, X.-J. Wang, D. J. Freedman, Circuit mechanisms for the maintenance and manipulation of information in working memory. bioRxiv: 10.1101/305714 (22 April 2018).
35. R. Quentin *et al*., Differential brain mechanisms of selection and maintenance of information during working memory. *J. Neurosci.* **39**, 3728–3740 (2019).
36. R. N. Shepard, J. Metzler, Mental rotation of three-dimensional objects. *Science* **171**, 701–703 (1971).
37. T. D. Wager, E. E. Smith, Neuroimaging studies of working memory: A meta-analysis. *Cogn. Affect. Behav. Neurosci.* **3**, 255–274 (2003).
38. A. P. Georgopoulos, J. T. Lurito, M. Petrides, A. B. Schwartz, J. T. Massey, Mental rotation of the neuronal population vector. *Science* **243**, 234–236 (1989).
39. M. Lundqvist *et al*., Gamma and beta bursts underlie working memory. *Neuron* **90**, 152–164 (2016).
40. M. Stokes, E. Spaak, The importance of single-trial analyses in cognitive neuroscience. *Trends Cogn. Sci. (Regul. Ed.)* **20**, 483–486 (2016).
41. S. Taulu, M. Kajola, J. Simola, Suppression of interference and artifacts by the signal space separation method. *Brain Topogr.* **16**, 269–275 (2004).
42. R. Oostenveld, P. Fries, E. Maris, J.-M. Schoffelen, FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* **2011**, 156869 (2011).
43. F. Tadel, S. Baillet, J. C. Mosher, D. Pantazis, R. M. Leahy, Brainstorm: A user-friendly application for MEG/EEG analysis. *Comput. Intell. Neurosci.* **2011**, 879716 (2011).
44. A. Gramfort *et al*., MNE software for processing MEG and EEG data. *Neuroimage* **86**, 446–460 (2014).