Behavioral/Cognitive

Targeted Memory Reactivation during Sleep Elicits Neural Signals Related to Learning Content

Boyu Wang,[1,2]* James W. Antony,[1]* Sarah Lurie,[3] Paula P. Brooks,[1] Ken A. Paller,[3,4] and Kenneth A. Norman[1]

[1]Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey 08544, [2]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104, [3]Northwestern University Interdepartmental Neuroscience Program, and [4]Department of Psychology, Northwestern University, Evanston, Illinois 60208

Retrieval of learning-related neural activity patterns is thought to drive memory stabilization. However, finding reliable, noninvasive, content-specific indicators of memory retrieval remains a central challenge. Here, we attempted to decode the content of retrieved memories in the EEG during sleep. During encoding, male and female human subjects learned to associate spatial locations of visual objects with left- or right-hand movements, and each object was accompanied by an inherently related sound. During subsequent slow-wave sleep within an afternoon nap, we presented half of the sound cues that were associated (during wake) with left- and right-hand movements before bringing subjects back for a final postnap test. We trained a classifier on sleep EEG data (focusing on lateralized EEG features that discriminated left- vs right-sided trials during wake) to predict learning content when we cued the memories during sleep. Discrimination performance was significantly above chance and predicted subsequent memory, supporting the idea that retrieval leads to memory stabilization. Moreover, these lateralized signals increased with postcue sleep spindle power, demonstrating that retrieval has a strong relationship with spindles. These results show that lateralized activity related to individual memories can be decoded from sleep EEG, providing an effective indicator of offline retrieval.

Key words: **episodic memory; memory consolidation; memory reactivation; multivariate pattern analysis; sleep**

Significance Statement

Memories are thought to be retrieved during sleep, leading to their long-term stabilization. However, there has been relatively little work in humans linking neural measures of retrieval of individual memories during sleep to subsequent memory performance. This work leverages the prominent electrophysiological signal triggered by lateralized movements to robustly demonstrate the retrieval of specific cued memories during sleep. Moreover, these signals predict subsequent memory and are correlated with sleep spindles, neural oscillations that have previously been implicated in memory stabilization. Together, these findings link memory retrieval to stabilization and provide a powerful tool for investigating memory in a wide range of learning contexts and human populations.

Introduction

In recent decades, evidence has increasingly converged on the idea that newly formed memory traces are spontaneously retrieved during sleep, promoting their stabilization (Wilson and McNaughton, 1994; Ji and Wilson, 2007). Decoding the content of replayed information from animal studies (using invasive electrophysiology) has massively advanced our understanding of how sleep contributes to learning and memory. Detecting noninvasive, content-specific signals of retrieval in humans is, therefore, of utmost interest for understanding brain mechanisms of memory.

Several recent studies have brought us closer to this goal. Multivariate classification methods have been used to decode (during sleep) the category of information (faces or scenes) learned before sleep in the EEG (Schönauer et al., 2017) and stimulus-specific information in fMRI (Deuker et al., 2013). Another approach that allows for more temporal precision over which memory is being retrieved is to play learning-related cues during sleep. This approach, termed targeted memory reactivation

Received Oct. 30, 2018; revised June 11, 2019; accepted June 14, 2019.

Author contributions: B.W., J.W.A., and K.A.N. analyzed data; B.W., J.W.A., S.L., P.P.B., K.A.P., and K.A.N. edited the paper; B.W., J.W.A., and K.A.N. wrote the paper; J.W.A., S.L., K.A.P., and K.A.N. designed research; J.W.A., S.L., and P.P.B. performed research; J.W.A. wrote the first draft of the paper.

Parts of this paper were adapted from S.L.'s undergraduate thesis. This work was supported by a CV Starr Fellowship to J.W.A. and National Science Foundation BCS Grants 1533511 and 1461088 to K.A.P. and K.A.N. We thank Elizabeth McDevitt and Monika Schönauer for helpful comments.

The authors declare no competing financial interests.

*B.W. and J.W.A. contributed equally to this work as co-first authors.

Correspondence should be addressed to James W. Antony at jantony@princeton.edu.

https://doi.org/10.1523/JNEUROSCI.2798-18.2019

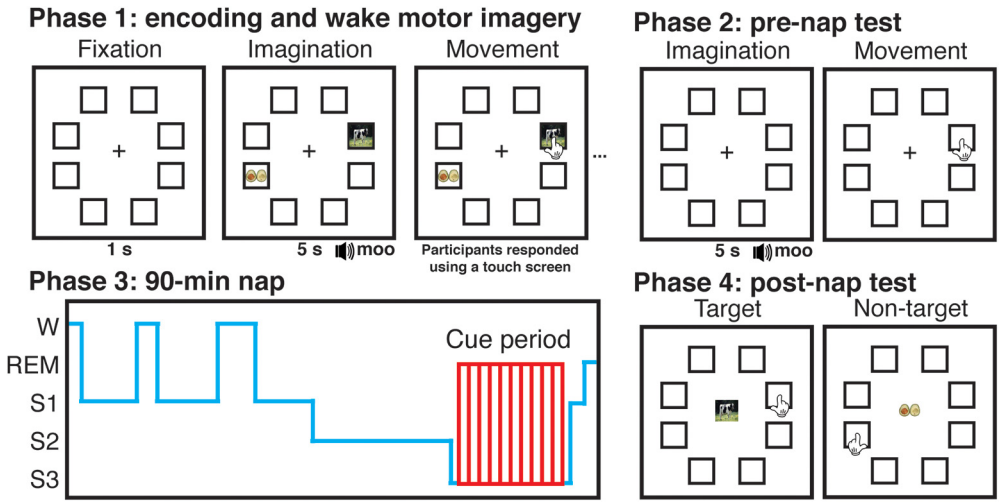

Figure 1. Experimental task and design. Subjects imagined reaching out and touching targets that were associated with auditory cues before executing their movement. Subjects used either their left or right hand depending on the location of the target on the screen (left hand for 32 left-side targets, right hand for 32 right-side targets). Subjects then took a prenap test on the location of each target. During periods of SWS in an afternoon nap, subjects were presented with half of the sounds associated with left-side and half associated with right-side targets. After the nap, subjects returned to the laboratory and took a test on all target and nontarget locations.

(TMR) (Oudiette and Paller, 2013), biases neuronal replay in the hippocampus in rodents (Bendor and Wilson, 2012), and benefits memory in humans (Rasch et al., 2007; Rudoy et al., 2009), suggesting it resembles retrieval that occurs spontaneously. Recently, Cairney et al. (2018) linked TMR cues (auditory words) with visual images and found that the image category (face or scene) could be decoded during post-TMR cue activity. Additionally, they found that the presence of sleep spindles, brief (0.5–3 s) bursts of 11–16 Hz oscillatory brain activity during NREM sleep that frequently correlate with memory retention (Antony and Paller, 2017) overlapped with time intervals when this decoding was possible. Similarly, Shanahan et al. (2018) showed decoding of TMR-evoked category information using fMRI. Finally, multivariate features from EEG activity after TMR cues could be used to successfully decode different parts of a procedural memory sequence (Belal et al., 2018).

Other prior findings suggest that neural signals resulting from lateralized neural activity can be robustly tracked with EEG during sleep. After training subjects to perform lateralized judgments on words during wake (e.g., "press left if word, press right if nonword"), corresponding lateralized responses were elicited by those stimuli during sleep (Kouider et al., 2014; Andrillon et al., 2016). Additionally, lateralized differences in the number and amplitude of fast sleep spindles emerged after cueing learning contexts using learning-related odors that were predominantly associated with left- or right-sided visual stimuli (Cox et al., 2014).

Here, we sought to extend the aforementioned results by "tagging" individual declarative memories with lateralized movements and activating them during sleep via auditory TMR cues. In so doing, we could constrain our search for evidence of memory retrieval to these robust lateralization signals. Subjects first encoded visual objects on the left or right side of a computer touch screen (Fig. 1). Each target object (e.g., cow) was accompanied by an inherently related sound (e.g., "moo") and was presented concurrently with a visual nontarget (e.g., avocado) on the opposite side of the screen. During encoding, subjects imagined moving their corresponding hand toward the target (e.g., right hand for right-side stimuli) for 5 s, after which they executed the movement. Subjects then took an item-location test for each target without feedback before an afternoon nap. During indications of slow-wave sleep (SWS), we presented half of the TMR cues that were previously associated with both left- and right-side targets. After a 90-min break, subjects returned for a final memory test.

Behaviorally, we predicted subjects would remember cued targets better than uncued ones. Neurally, we predicted the emergence of lateralized, learning-related EEG activity during sleep following TMR cues associated with left- versus right-side targets. We predicted that these signals would be larger for targets with intact prenap memory and would also predict postnap memory. Another key goal of our study was to relate lateralized evidence of retrieval to other sleep physiological predictors of postnap memory. TMR benefits have been intimately linked with postcue spindles (Schreiner et al., 2015; Antony et al., 2018a,b; Cairney et al., 2018). Therefore, we anticipated that postcue spindles would positively predict lateralized evidence of retrieval during sleep, as well as subsequent memory during the postnap test. Furthermore, prior work has shown that spindles have a refractory period of 3–6 s, during which TMR cues are less effective (Antony et al., 2018b); as such, we expected that precue spindles would negatively predict both lateralized evidence of retrieval and subsequent memory.

Materials and Methods

Subjects. Twenty-four subjects (14 female, 18–33 years old, mean 22.3) with normal or corrected-to-normal vision and fluent in English were recruited via campus flyers and online scheduling software. Twenty-nine other subjects were excluded for not sleeping long enough to receive at least one round of sleep cues. Subjects were given hourly monetary compensation for participating ($20/h) and small additional increases based on good performance (up to $5). To increase the likelihood of a successful nap, we requested that subjects go to bed at their normal sleeping time but wake up 2 h earlier than normal. We also requested that they abstain from alcohol the night before and coffee the morning of the experiment. Written informed consent was obtained in a manner approved by the Princeton University Institutional Review Boards.

Experimental setup. Visual stimuli were presented on a 120 Hz LCD monitor in the testing room outfitted with an infrared touch screen frame. Touches on this display were registered as mouse input. Audio

cues (during both wake and sleep) were played through a speaker system in the testing room.

Stimuli. Subjects viewed 128 images of common objects. Each image was shown in one of eight stimulus locations with a 300 × 300 pixel resolution. Sixty-four of these were associated with inherently related sounds (e.g., cow-"moo") lasting up to 500 ms; the sounds were adapted from a previous study (Oudiette et al., 2013) and an online repository at www.freesound.org. Shorter sounds were looped, and longer sounds were trimmed, sped up, and pitch-corrected to keep them recognizable. During the nap, sleep cues were embedded in constant white noise (~ 44 dB), resulting in increases of no larger than 5 dB.

Design and procedure. The experiment included four phases: learning, prenap test, nap, and postnap test (Fig. 1). Subjects arrived in the laboratory at 11:00 A.M., and EEG electrodes were attached. They completed the learning phase task, took a 15-min break, and then took the prenap test. Following the test, subjects napped in the testing room for 90 min from ~13:00 to 14:30. After the nap, subjects were given a 90-min break during which they were allowed to leave the laboratory. Subjects returned at 16:00 to complete the postnap test.

In the learning phase, subjects were trained to execute a lateralized motor action corresponding to the locations of 64 image-pairs. Each pair consisted of one target and one nontarget image; the pair was presented along with an audio cue that was inherently associated with the target (e.g., cow-"moo"). Thus, the content of the audio cue ("moo") specified to the subject which image was the target and (by exclusion) which image was the nontarget. Nontargets had no immediately obvious sounds (e.g., avocado) and were included along with the target to ensure subjects paid attention to the semantic content of the images.

Target–nontarget pairs and image locations were assigned pseudorandomly for each subject. Images appeared in one of eight possible locations, each visually designed on the screen by a black, square outline. The target and nontarget for a given sound were always on opposite sides of the screen. We assigned image locations such that each location was associated with 16 images, an equal number of which were targets and nontargets.

Subjects underwent two blocks of image-location learning, each with 64 trials, so that each target–nontarget pair was seen twice (once per image-pair). Trial order was randomized in each block. A single trial comprised a fixation period, a stimulus presentation, an imagined selection, a physical selection, and a rest period. During the fixation period, subjects were presented with a red central fixation cross and the eight empty image frames. After 1 s, an image-pair (target and nontarget) appeared and the sound associated with the target was played. Subjects had 5 s to imagine reaching out and touching the target with their corresponding hand (e.g., right hand for right side of screen). Then the fixation crosshair changed from red to green, signaling that the subject could execute their movement. If the subject touched the nontarget or any other location, on-screen text instructed them to try again until they touched the target. They were accurate at doing this on the first attempt 98.3 ± 0.3% of the time.

In the prenap test, each trial comprised a fixation period, a sound stimulus presentation, an imagined location selection, a physical location selection, and a rest period. After 1 s fixation, the target-related sound was played, and subjects had 5 s in which to imagine touching the frame where the associated target had appeared during the learning phase, after which the crosshair turned green and they executed their movements. Each target was tested once, and subjects received no feedback on response accuracy.

Following the test, subjects napped for 90 min. All naps began between 12:00 and 13:00. Upon indications of SWS, we played half of the sounds associated with left targets (16) and half of the sounds associated with right targets (16) to cue their associated memories. Cues were separated by 4.5 s, and cue order was randomized within loops. If the subject was asleep for 60 min and had not received cues, the experimenter loosened the criteria to include Stage 2 sleep. The EEG was monitored continuously during cueing, and cues were stopped immediately upon indications of arousal; cues were resumed if the subject reentered SWS. Subjects were cued between 1 and 7 times per cue (mean = 3.32 ± 0.42) depending on how long they spent in SWS and Stage 2 sleep. If by 90 min the subject was still asleep and had not completed a loop of cues, they were kept for an additional 20 min to try to reach this criterion. Table 1 shows a full breakdown of the stages and where cues fell with respect to those stages. Sleep stages are assigned based on whichever stage is more prevalent within each 30 s epoch, meaning that sounds can occur during unintended stages. Only sounds occurring during Stage 2 sleep and SWS were included in our analyses.

Table 1. Time in each sleep stage and number of sounds per stage

	Wake	S1	S2	S3	REM
Time/stage (min)					
Mean	27.35	6.37	27.43	26.65	3.78
SEM	3.09	0.93	2.54	2.92	1.13
Cues/stage					
Mean	2.09	0.13	9.52	97.04	0.13
SEM	0.75	0.07	4.25	14.68	0.07

After the nap, subjects left the laboratory and returned again 90 min later for the final phase. They then completed a postnap item-location test on both target and nontarget locations (128 trials total). On each trial, an image appeared in the center of the screen surrounded by the 8 empty frames as before. The subject was instructed to touch the frame where they believed the image appeared during the learning phase. After this test, subjects were compensated for their time and left the laboratory.

EEG data acquisition and preprocessing. Continuous EEG data were recorded during the learning phase, prenap test, and nap using Ag/AgCl active electrodes (BioSemi ActiveTwo). Recordings were made at 256 Hz from 64 scalp EEG electrodes from locations on the standard 10/20 layout plus intermediate 10% electrodes. Additionally, two mastoid channels were placed behind the left and right ears for offline rereferencing, a horizontal EOG electrode was placed next to the right eye, a vertical EOG electrode under the left eye, and an EMG electrode on the chin.

EEG data were processed using a combination of internal functions in EEGLAB (Delorme and Makeig, 2004) and custom-written scripts. Data were rereferenced offline to the average signal of the left and right mastoids; then they were high-pass filtered at 0.1 Hz and low-pass filtered at 60 Hz in successive steps. For each wake recording, we removed artifactual segments in the continuous EEG of the learning phase. We ran spatial independent components analysis (ICA) (Delorme and Makeig, 2004) for each subject to remove eyeblink components. Any electrodes with artifacts for long stretches of time were marked and interpolated within-subject after ICA.

Next, we created across-subject components. The learning phase data were segmented into epochs, including the 5 s imagination period with 2 s of baseline (i.e., −2 to 5 s relative to the imagination cue). We subtracted voltage values from lateralized channels (e.g., C3–C4) and omitted channels along the midline to reduce the number of features from 64 channels to 27 channel pairs (Kouider et al., 2014; Andrillon et al., 2016) (Fig. 2). We then concatenated data from all 24 subjects to create one across-subject dataset from which we ran across-subject spatial ICA. Component activations (analogous to channel voltage, but in component space) from this ICA were then used as our features of interest. This procedure allowed us to reduce the dimensionality of the data while maintaining consistent features across individuals.

To assess spindle power, we first bandpass-filtered the raw signal between 11 and 16 Hz. Then we calculated the root-mean-square (RMS) value for every time point using a moving window of ± 100 ms (Antony et al., 2018a,b). To calculate spindles in a quantized fashion, a threshold was determined by multiplying the SD of the entire channel's signal for all nonartifactual segments of NREM sleep by 1.5. Any RMS signal that crossed this threshold consecutively for 0.5–3 s was considered a spindle. Times for the start of each spindle were recorded for alignment with sleep cues.

Sleep physiological analyses. Sleep stages were determined by an expert scorer according to standard American Academy of Sleep Medicine criteria (Rechtschaffen and Kales, 1968). Artifacts (large movements, blinks, arousals, and rare, large deflections in single channels) during sleep were identified visually and rejected in 5 s chunks following sleep staging.

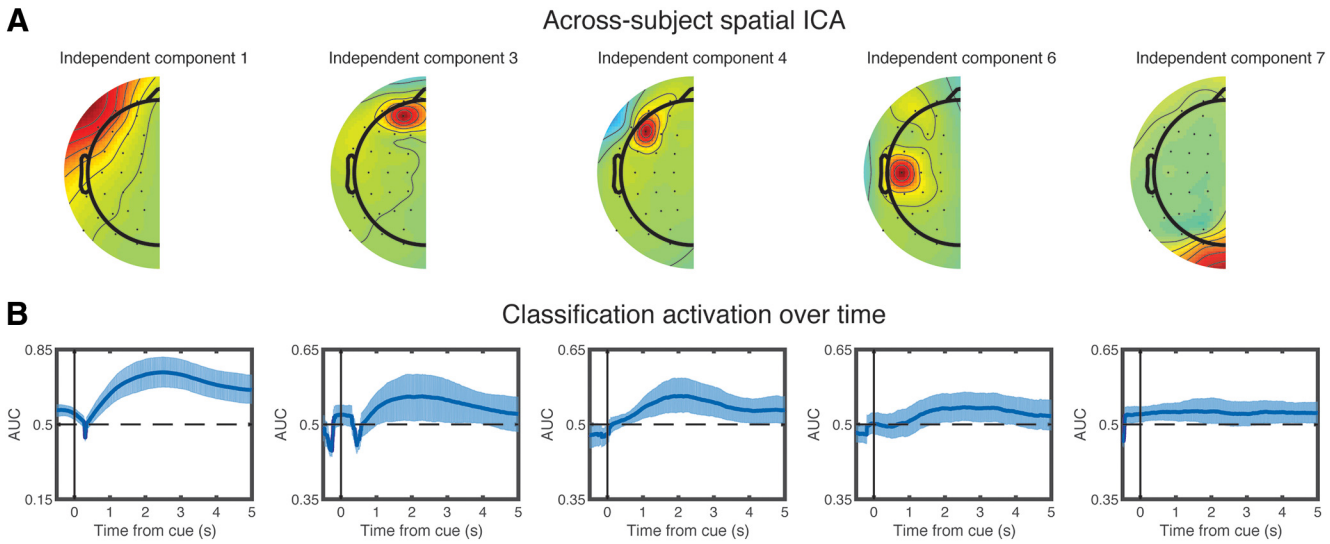

Figure 2. Wake classification and feature selection for sleep. Voltages from lateralized EEG electrodes were subtracted, and midline electrodes were eliminated to create 27 electrode pairs. Next, data from all subjects were combined. ICA was run on the combined data, and we assessed which of the resulting components discriminated between left and right motor imagery during wake. ***A***, The five components (1, 3, 4, 6, and 7) that significantly discriminated between left and right motor imagery. ***B***, Classification accuracy (operationalized as AUC) as a function of time bin for those five components. Error bars indicate the edges of the middle 95th percentile of classification accuracies. *x* axis values indicate the center time point for each 1000 ms time bin. There is a wider *y* axis range for component 1 than other components.

Experimental design and statistical analysis. All behavioral contrasts for within-subject measures were performed using paired, two-tailed *t*-tests with a significance level of $p = 0.05$; these tests contrasted the mean proportions of items that were correct under various conditions. To calculate the mean cueing effect, we regressed out the effect of prenap memory from postnap − prenap differences for cued and uncued groups. Specifically, we computed (for each subject) the average level of prenap memory (separately for cued and uncued) and the average level of postnap − prenap memory change (separately for cued and uncued). We then put the subject-wise cued and uncued memory scores into a single regression (i.e., with number-of-subjects × 2 data points), using prenap memory scores to predict memory change scores. The residuals from this regression reflect the cued and uncued memory change scores for each subject, controlling for prenap memory.

To identify lateralized features that might be useful for decoding memory retrieval during sleep, we first asked how well each ICA component discriminated between left and right imagination conditions during wake in two successive steps. First, we performed baseline correction on each trial by subtracting out the average activity from −1000 to 0 ms relative to the onset of the visual stimulus. We divided each trial's data into overlapping time bins of length 1000 ms; for each time bin, we trained a single-feature, logistic regression classifier (that component's average value across the time bin) to discriminate between left and right trials, using a leave-one-subject-out cross-validation procedure. This procedure yielded an area under the curve (AUC) value for each subject and each time bin. We then used a bootstrap procedure to generate 95% CIs for the average AUC value (across subjects) for each time bin. For this procedure, we resampled subjects' AUC scores with replacement 1000 times. For each resampling, we computed the average AUC across (resampled) subjects, yielding a bootstrap distribution of 1000 average AUC scores for each time bin. After sorting the AUC values at each time bin from low to high, we counted the number of time bins for which the middle 95th percentile of bootstrapped trials passed above or below chance (at the 2.5th or 97.5th percentile); we refer to this measure as "number of significant time bins."

Second, to correct for multiple comparisons (i.e., looking at multiple time bins), we scrambled the left and right condition labels 200 times and repeated the bootstrap procedure above for each of the scrambles. This allowed us to compute a null distribution on the "number of significant time bins" measure described above. If, for a given component, the number of individually significant time bins exceeded 95% of the null distribution, the component was deemed to significantly discriminate between left and right trials (corrected for multiple comparisons).

To compute classification accuracy during sleep (in response to TMR cues), we computed the time series of activity (on each trial, in response to the TMR cue) of each of the five components that were identified as significant using the wake-classification procedure outlined above. We again performed baseline correction from −1000 to 0 ms relative to cue onset. As with the wake-classification procedure, we divided each trial's data (time-locked to the TMR cue) into overlapping time bins of length 1000 ms, and we trained a separate classifier for each time bin. Putting all of this together, each of these time bin-specific classifiers was fed a 5-dimensional vector for each TMR cue (reflecting the average activity of each of the five components during the relevant time bin, in response to that cue); the job of the classifier was to determine whether the TMR cue that was just played was associated with a left- or right-sided movement. As with wake classification, we used a leave-one-subject-out cross-validation procedure that yielded an AUC score for each time bin for each subject. For the sleep data, we used an L1-regularized logistic regression classifier ($\lambda = 0.1$). We used regularization here, despite the relatively low number of feature dimensions, as a hedge against the noisy nature of EEG data, which can otherwise lead to spurious classification results (Jamalabadi et al., 2016). We chose the largest lambda (in powers of 10) that still led to some nonzero weights for the time points of interest. Multiple-comparisons-corrected significance was assessed using the same bootstrapping-and-scrambling procedure that we used for the wake classifier. We report the *p* value based on this analysis such that, for example, if the true number exceeds 98% of the null distribution, it was reported as $p = 0.02$. In some cases, the regularization was strong enough such that all of the feature weights were zero, resulting in a classification score that was exactly at chance. Also, for the sleep-EEG analyses, we omitted 1 subject who did not have any items that were initially remembered and later forgotten (making it impossible to include them in subsequent-memory analyses). We did this to ensure that all of our sleep-EEG analyses included the same subjects, thereby maximizing comparability across these analyses. After this 24th subject was dropped, we were left with 23 subjects for most of the sleep-EEG analyses (importantly, the qualitative nature of our results was the same regardless of whether the 24th subject was included; for example, redoing the classifier analysis shown in Fig. 3*A* with all 24 subjects still produced a significant result of $p < 0.01$). Nine other subjects did not have trials in both left- and right-

cue conditions that were initially remembered and later forgotten, leaving 14 subjects for analyses related to this condition.

First, we ran this sleep-classification analysis on all trials. Second, we conditionalized the data based on whether items were correct before the nap or not, hypothesizing that only items correct before the nap could be retrieved during sleep. For this and all other conditionalized analyses, we trained and tested on the same subset of trials that we used to conditionalize the data (here, prenap accuracy). We also ran an analysis to look at items that were remembered correctly prenap, and additionally conditionalized by whether items were remembered (or not) postnap, hypothesizing that the content-related signal may predict subsequent memory. For the latter two ways of splitting the data, we also conducted subtraction analyses, whereby we calculated differences in the AUC (e.g., between prenap-correct and prenap-incorrect) after each iteration of resampling the data, and then we assessed whether these AUC differences were significantly different from zero using the same bootstrapping-and-scrambling procedure as above. The mean number of trials per subject that entered our analysis (omitting those falling during rejected artifacts) was as follows: items remembered on both the prenap and postnap tests, 48.83 ± 15.4; items remembered on the prenap test, but not the postnap test, 9.39 ± 4.14; items not remembered on the prenap test, but remembered on the postnap test, 20.17 ± 7.34; items not remembered on either test, 27.61 ± 10.82. This counts the same cue presented to the same subject (e.g., "moo" on the first and second round of cues) as separate trials. The mean number of items in each memory category was as follows: items remembered on both the prenap and postnap tests, 29.62 ± 1.58; items remembered on the prenap test, but not the postnap test, 5.58 ± 0.56; items not remembered on the prenap test, but remembered on the postnap test, 12.04 ± 0.90; items not remembered on either test, 16.75 ± 1.31.

To calculate the effects of spindle power on subsequent memory, we contrasted items that were correct both prenap and postnap with items that were only correct prenap. We averaged RMS values across specified time windows for each subject and calculated differences across conditions using within-subject *t*-tests. Our specified precue window was −2000 to 0, based on a previous study (Antony et al., 2018b). Our *a priori* postcue window was 1000 to 1500 ms (Antony et al., 2018a), although we eventually widened this window to 1000 to 4500 ms after observing that left/right decoding accuracy stayed above chance up to 4.5 s after cue onset. The postcue analyses were baseline-corrected, whereas the precue analyses were not.

We additionally investigated whether precue and postcue spindle power modulates this content-related signal. For these analyses, we calculated AUC as above after splitting trials by each subject's median RMS power value in the prespecified time windows (pre: −2000 to 0 ms; post: 1000 to 4500 ms). Similar to the above approaches, we also calculated subtracted differences between AUC values for trials above and below their corresponding median RMS values. For these analyses, we only used trials that were correct before the nap.

For comparison with the results of a previous study (Cox et al., 2014), we calculated spindle incidence and amplitude measures in electrode pairs chosen *a priori* (TP7–TP8, and P7–P8). Spindle amplitude was measured as the absolute value (in μV) of the maximum negative peak. These contrasts were double subtractions, such that we subtracted these spindle measures between the two electrodes and the two conditions (left vs right targets). We considered spindles that started between 0 and 4 s after the cue. Based on the previous results, we initially restricted our analyses to fast spindles (13.5–16 Hz), but we also later tested slow and fast spindles together (11–16 Hz). We calculated these double subtraction measures for each subject and asked whether they differed from zero using paired, two-tailed *t*-tests with a significance level of $p = 0.05$.

Data availability. The datasets generated during and/or analyzed during the current study and the analysis code will be made available within a year of publication on the Open Science Framework website.

Results

Behavior

Before sleep, subjects took an item-location memory test on all targets, with 8 possible locations (chance proportion correct = 0.125). Performance did not differ on the prenap test between targets that were later cued (mean proportion correct and SE of measurement: 0.55 ± 0.024) and uncued (0.55 ± 0.027; $t_{(23)} = 0.06$, $d_z = 0.01$, $p = 0.95$). After the nap, subjects returned to the laboratory to take another item-location memory test on all targets and nontargets. We measured memory retention for targets by subtracting prenap performance from postnap performance while regressing out prenap performance, similar to previous reports (Antony et al., 2018a,b). We measured retention for nontargets by assessing postnap accuracy because these were not tested before the nap. We predicted that cued targets would show better retention than uncued targets. Contrary to our prediction, this difference was not significant (cued: 0.32 ± 0.02; uncued: 0.29 ± 0.02; $t_{(23)} = 1.4$, $d_z = 0.29$, $p = 0.17$). However, we reasoned that retrieval during sleep may only be possible for items remembered prenap because feedback was not given after the prenap test. Therefore, we ran another analysis looking at the proportion of items still remembered postnap that were remembered prenap. We found that initially remembered cued targets were remembered marginally better than uncued targets (cued: 0.85 ± 0.02; uncued: 0.81 ± 0.03; $t_{(23)} = 1.8$, $d_z = 0.38$, $p = 0.07$). Finally, there was no difference between the proportion of remembered nontargets whose sound was associated with cued targets versus uncued targets (cued: 0.28 ± 0.02; uncued: 0.29 ± 0.02; $t_{(23)} = 0.63$, $d_z = 0.13$, $p = 0.53$).

Feature selection via decoding during wake

Our primary goal in analyzing the wake EEG data was to find neural features that discriminated between left and right conditions during motor imagery at encoding, with the aim of using those features for classification during sleep. To accomplish this, we relied on the prominent signal produced by the lateralized readiness potential as a result of imagining motor activity from opposite limbs (Kouider et al., 2014). We subtracted each lateralized electrode on the right side of the head from its left side counterpart, eliminated 10 midline electrodes, and ran across-subject spatial ICA to obtain 27 components (Fig. 2*A*; for more details, see Materials and Methods). We then asked which components significantly discriminated between the conditions (left vs right) during wake (for more details, see Materials and Methods). Plots of significant components are shown locked to the visual stimulus onset when subjects were directed to begin motor imagery (Fig. 2*B*). These components thus provided neural features of learning-related activity by which we could assess memory retrieval during sleep.

Content-related activity during sleep differs by prenap and postnap target memory

We next sought to obtain evidence of learning-content-related activity during sleep. We investigated neural activity after playing TMR cues that were paired with targets associated with left- versus right-side movements (hereafter, left-targets and right-targets). We trained a multivariate classifier on sleep EEG data, using (as input features) the same ICA components that were significantly discriminable during wake (for details, see Materials and Methods). This approach differs from using a classifier trained on wake EEG data; when we did this, we found no evidence of reinstatement (see No reinstatement of specific, learning-related pattern from wake imagery).

We conducted our sleep-classification analyses in three steps. First, we asked whether left-targets and right-targets were discriminable when we allow all items to enter in the analysis. We found that the classifier significantly predicted lateralized cate-

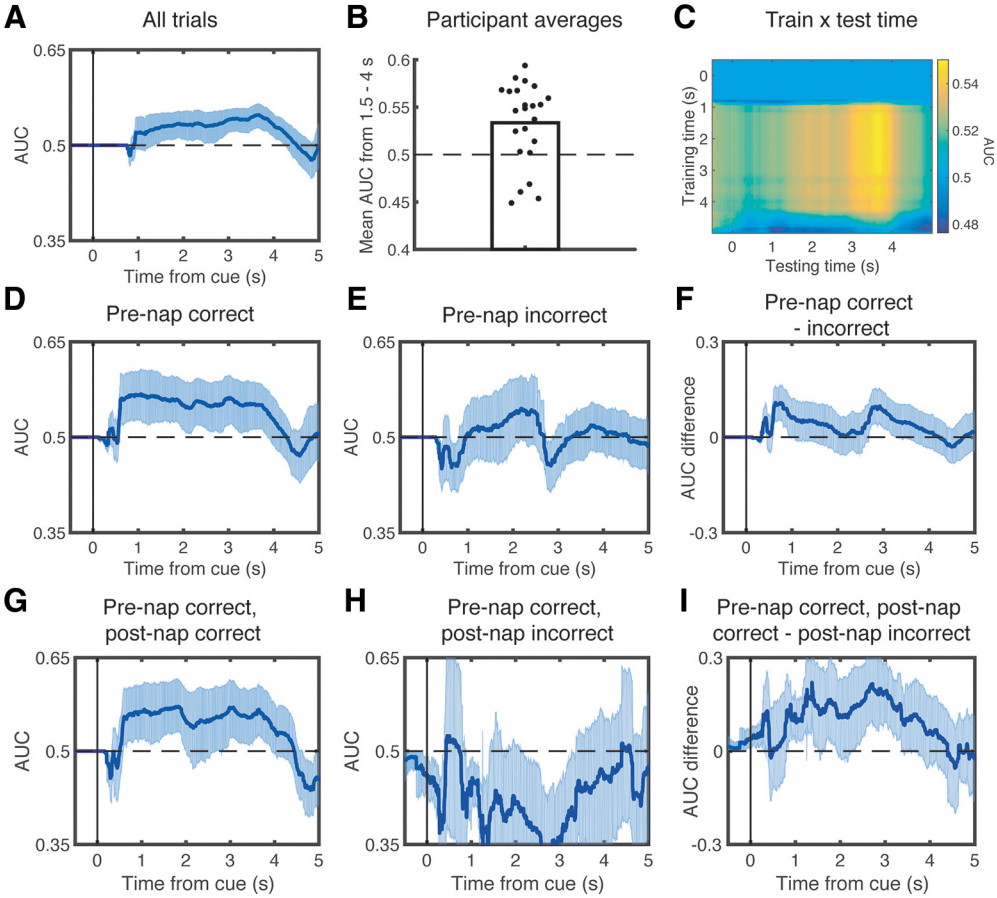

Figure 3. Lateralized retrieval of learning content during sleep related to prenap and postnap memory. Using the lateralized components that discriminated left versus right motor imagery during wake as features, we trained a leave-one-subject-out classifier on sleep data and computed AUC-based classification performance time-locked to TMR cues. We found significant classification performance when using all trials (***A***). ***B***, Significant across-subject decoding from 1.5 to 4 s. ***C***, Performance when crossing training and testing time. We also found significant classification performance when focusing just on trials that were remembered prenap (***D***), but not for trials that were not remembered prenap (***E***). However, there were no significant differences in classification performance between these conditions (***F***). Using only trials that were remembered before the nap, we then found significant classification performance for trials that were subsequently remembered postnap (***G***), but not for those forgotten postnap (***H***), and there were significant differences in classification performance between these conditions (***I***). In all graphs, classification was performed on time-averaged signals within 1000 ms windows centered on each time point. Error bars indicate the edges of the middle 95th percentile of classification accuracies.

gory, most prominently between 1376 and 4270 ms postcue ($p = 0.01$; Fig. 3*A*). This analysis remains significant if we remove independent component 1 ($p = 0.005$), which topographically suggests that it may include some horizontal eye activity during wake. To depict the size of this effect across subjects, we plotted the mean AUC values from 1.5 to 4 s after the cue, which significantly differed from chance (mean AUC: 53.4 ± 0.89, $t_{(22)} = 3.76$, $d_z = 0.78$, $p = 0.001$; Fig. 3*B*). To gain a better understanding of how the features detected by the classifier might change over time, we also plotted AUC values for every combination of training and testing time points (Fig. 3*C*). It appears from the "block-like" pattern (i.e., similar results across multiple training times) that the features that discriminate between classes are relatively stable over time.

Second, we reasoned that lateralized activity should be greater for targets whose position was correctly remembered before the nap. As there was no feedback given after the prenap test, these were the only cues that we could verify were correctly encoded. For items recalled prenap, we again found significant classification, in this case, most prominently between 592 and 3924 ms postcue ($p = 0.01$; Fig. 3*D*). Conversely, for items not remembered prenap, classification was not significantly predictive (no significant time points, $p > 0.99$; Fig. 3*E*). We also directly contrasted decoding ability for items initially remembered versus those not initially remembered by subtracting the classification accuracy for each condition on each resampled subset, and then assessed whether this differed from what is expected by chance by repeating this analysis using scrambled labels. This difference was not significant ($p = 0.195$; Fig. 3*F*).

Third, limiting ourselves to items that were correctly remembered prenap, we further asked whether there were differences in classification based on postnap memory. For items that were remembered on both the prenap and postnap tests, we found significant above-chance classification, most prominently from 592 to 2016 ms and 2804 to 3932 ms ($p = 0.015$; Fig. 3*G*). For items remembered only on the prenap test but not the postnap test, we found no meaningful above-chance segments ($p = 0.605$; Fig. 3*H*). Finally, we directly contrasted these signals and found that classification was significantly higher for items remembered prenap and postnap than for items only remembered prenap ($p < 0.01$; Fig. 3*I*).

Postcue and precue σ power positively and negatively predicts subsequent memory

Recent evidence has suggested that post-TMR cue σ (spindle-band) power positively predicts memory, an effect that is maximal over the centroparietal midline electrode, CPz (Antony et al., 2018a,b). We therefore asked whether these signals predict sub-

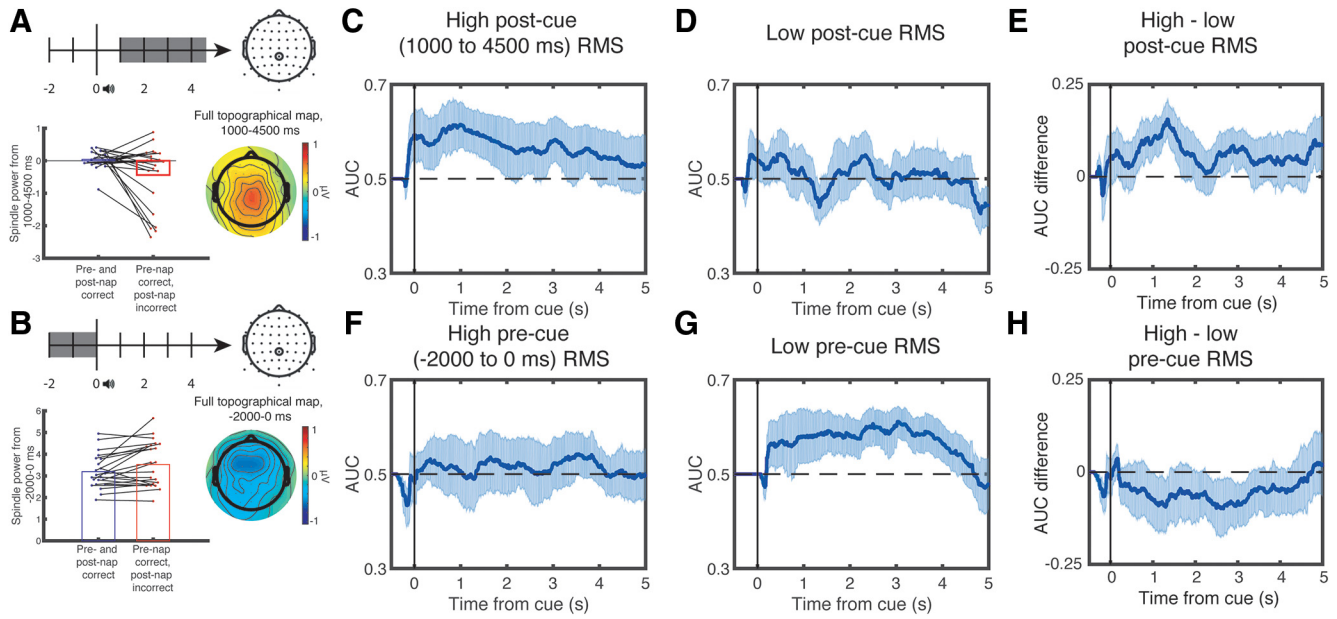

Figure 4. Retrieval of learning content increased with higher postcue σ power and decreased with higher precue σ power. Subsequent memory differed as a function of postcue (1000 to 4500 ms, baseline-corrected) (***A***, left) and precue σ power (−2000 to 0 ms, no baseline correction) (***B***, left) over electrode CPz. We also plotted the full topography across the scalp (***A***, ***B***, right). We then plotted AUC-based classification performance as a function of postcue and precue σ power. Trials were split by the median RMS power in the σ band of each subject. Classifier accuracy was significantly above chance on trials with above-median postcue σ power (***C***), but not below-median σ power (***D***). Moreover, there were significant differences in classifier accuracy between these conditions (***E***). Conversely, classifier accuracy was not significantly above chance on trials with above-median precue σ power (***F***), but it was for trials with below-median σ power (***G***), and there were marginally significant differences in classifier accuracy between these conditions (***H***). In all graphs, classification was performed using 1000 ms windows centered on each time point. Error bars indicate the edges of the middle 95th percentile of classification accuracies.

sequent memory for items that were remembered prenap. We found that postcue σ power over CPz was significantly higher from 1000 to 4500 ms for later-remembered than later-forgotten targets (later-remembered: 0.016 ± 0.06; later-forgotten: -0.74 ± 0.36, $t_{(22)} = 2.15$, $d_z = 0.45$, $p = 0.043$; Fig. 4*A*). This analysis did not reach significance using only the early interval (1000 to 1500 ms) that we chose *a priori* based on previous studies, as described in Materials and Methods (later-remembered: 0.69 ± 0.17; later-forgotten: -0.05 ± 0.48, $t_{(22)} = 1.49$, $d_z = 0.31$, $p = 0.15$).

Intriguingly, while postcue σ power may indicate that successful retrieval has occurred, precue σ may prevent retrieval because cues fall within the spindle refractory period (Antony et al., 2018b). Therefore, we next asked whether precue σ power over CPz negatively predicted subsequent memory for items that were remembered prenap. For this analysis, we did not perform standard baseline correction, as σ power within the baseline interval is our variable of interest. Indeed, precue σ power over CPz was significantly lower from −2000 to 0 ms for later-remembered than later-forgotten items (later-remembered: 3.24 ± 0.16; later-forgotten: 3.73 ± 0.29, $t_{(22)} = 2.42$, $d_z = 0.50$, $p = 0.024$; Fig. 4*B*). In sum, both postcue and precue effects replicate previous findings and allow us to use σ power over CPz as an independent predictor of lateralized retrieval evidence.

Postcue and precue σ power positively and negatively predicts lateralized memory retrieval

We next asked whether postcue σ power was associated with improved lateralized memory retrieval and precue σ power associated with impaired lateralized memory retrieval. First, we took all trials where prenap memory was correct and split these trials by median postcue σ power values. We found significant lateralized memory retrieval for trials with above-median ($p < 0.005$), but not below-median ($p = 0.42$), σ power (Fig. 4*C*,*D*). Moreover, when we subtracted the AUC values between these conditions on each resampled subset (as above), the strength of lateralized memory retrieval differed significantly ($p < 0.025$; Fig. 4*E*). Next, we took all trials where prenap memory was correct and split these trials by median precue σ power values. As predicted, we found significant lateralized memory retrieval for trials with below-median ($p < 0.005$), but not above-median ($p = 0.53$), σ power (Fig. 4*F*,*G*). When contrasting high and low precue σ power directly, there was marginally higher lateralized memory retrieval in the low precue σ power condition ($p = 0.07$; Fig. 4*H*).

While the above analyses used a median split on σ power to ensure equal numbers of trials in each group, we also performed similar analyses based on whether a postcue (starting 0–4 s relative to cue) or precue spindle (starting −3 to 0 s relative to cue) was present. Postcue spindles were found on 26.3% of trials, and precue spindles were found on 14.4% of trials. As expected from the above analysis using median split on σ power, we found significant classification for trials with postcue spindles ($p = 0.005$). Unlike the above analysis, we also found significant classification for trials without postcue spindles ($p = 0.005$). This divergence from the previous analysis likely stems from the fact that there is a higher proportion of trials (73.7%) without postcue spindles than for below median σ power (50% by convention). Nevertheless, directly contrasting the two conditions shows that there is significantly higher classification for trials with than without spindles ($p = 0.03$), underscoring the importance of postcue spindles. Similar to the above median split analysis, the opposite pattern emerges for precue spindles. On trials with a precue spindle, we see no significant classification ($p = 0.20$), whereas on trials with one, we see significant classification ($p < 0.005$). However, there was no significant difference between these conditions when compared directly ($p = 0.29$).

No reinstatement of specific, learning-related pattern from wake imagery

The above results involved extracting learning-related features and then training and testing a classification algorithm using those features via cross-validation during sleep. However, it is worth asking whether the specific learning-related pattern involved in wakeful motor imagery is reinstated during sleep. First, we established that wakeful motor imagery patterns could be successfully decoded via cross-validation ($p < 0.005$). Next, using the exact classifier weights that were obtained by training on wakeful motor imagery, we tested whether these patterns were reinstated during sleep. We found no evidence of this ($p = 0.43$). We also attempted to do this without independent component 1; while we were able to successfully cross-validate during wake ($p < 0.005$), we again found no evidence of these patterns during sleep ($p = 0.22$). Therefore, there seem to be differences in the feature weights that are most important for discriminating between motor imagery during wake versus retrieval during sleep.

Lateralized spindle incidence and power analyses

One report found that, after training subjects to learn visual stimuli in a spatial array in either left or right hemisphere-heavy blocks, TMR cues elicited differences in the number and amplitude of lateralized spindles (Cox et al., 2014). Their analysis focused specifically on fast (>13.5 Hz) spindles. Therefore, we tested whether these differences existed using (1) both all lateralized channels pooled together and (2) specifically using significant channels from the above study. For these analyses, we considered spindles that started between 0 and 4 s after the cue and were >13.5 Hz in frequency. For spindle incidence, we performed double subtractions on the average number of spindles occurring over each electrode based on cue type and channel hemisphere [e.g. (number of spindles after left cues in P7–P8) − (number of spindles after right cues in P7–P8)]. For spindle amplitude, we averaged the power of each spindle (calculated as the sum of the absolute value of its data points) for all spindles within a particular condition [e.g. (mean spindle power after left cues in P7–P8) − (mean spindle power after right cues in P7–P8)]. Pooling all 27 lateralized channels together, we found no significant differences in the spindle incidence [mean subtracted difference: 0.012 ± 0.013 spindles, $t_{(22)} = 0.87$, $d_z = 0.18$, $p = 0.40$] or amplitude [15 ± 43 μV, $t_{(22)} = 0.34$, $d_z = 0.07$, $p = 0.74$]. For the specific channel analyses, we tried two channel pairs (TP7–TP8 and P7–P8). Neither produced significant differences in incidence [TP7–TP8 difference: 0.008 ± 0.04 spindles, $t_{(22)} = 0.22$, $d_z = 0.05$, $p = 0.83$; P7–P8 diff: 0.05 ± 0.03 spindles, $t_{(22)} = 1.43$, $d_z = 0.30$, $p = 0.17$] or amplitude [TP7-TP8 difference: 169 ± 85 μV, $t_{(22)} = 1.38$, $d_z = 0.42$, $p = 0.20$; P7-P8 diff: 9 ± 53 μV, $t_{(22)} = 0.13$, $d_z = 0.04$, $p = 0.90$]. Finally, these analyses were similarly not significant when considering all slow and fast spindles together (all $p > 0.29$).

Discussion

We demonstrated that learning-related signals from individual episodic memories emerge when cued during sleep by decoding lateralized activity associated with specific sound cues. Evidence for retrieval was only significant if prenap memory was correct, suggesting that this effect only occurs for learned information. Furthermore, lateralized retrieval evidence also predicted postnap memory, specifically linking retrieval with memory stabilization. We also found that postcue and precue spindle power positively and negatively predicted subsequent memory, respectively. Moreover, lateralized retrieval evidence was stronger for trials with high postcue spindle power and low precue spindle power (the former contrast was significant; the latter contrast was trending). Collectively, these analyses support the idea that sleep benefits memory retention via retrieval of individual learning episodes.

In our study, content-related differences in neural activity were generally detectable from ~1 to 4 s after the cue. This result fits with those from other studies showing bouts of replay that extend for multiple seconds. For example, two studies using TMR in rodents found differences in learning-related neural activity for a number of seconds, lasting until the next cue arrived (Bendor and Wilson, 2012; Rothschild et al., 2017). While some human TMR studies have only reported neural differences predicting subsequent memory within the first ~1.5 s (Schreiner et al., 2015; Lehmann et al., 2016; Farthouat et al., 2017; Antony et al., 2018a,b), others have found differences up to and later than 2 s postcue (Schreiner et al., 2015, 2018; Cairney et al., 2018). Importantly, this long time window may reflect brief bouts of retrieval occurring at variable times within the window, rather than continuous retrieval throughout the window.

To the best of our knowledge, our results are the first to show that, across trials, retrieval evidence on any particular trial increases with higher postcue spindle power. These results nicely dovetail with findings showing that retrieval of neural content temporally coincides with spindles (Cairney et al., 2018). We also found a trend toward lateralized retrieval evidence decreasing with higher precue spindle power; this converges with results showing that TMR cues are effective when followed by spindles and less effective when presented in the spindle refractory period, when ensuing spindles are unlikely (Antony et al., 2018b). Our results strongly support a recent mechanistic framework proposing that memories are retrieved during spindle events and spindle refractoriness prevents further retrieval of other memories (Antony et al., 2019). However, we did not find evidence for the related idea that localized spindles support retrieval (Bergmann et al., 2012; Johnson et al., 2012; Cox et al., 2014). We hope future studies, perhaps using techniques with finer spatial resolution, such as electrocorticography or magnetoencephalography, or tasks engaging different areas of the neocortex, will clarify this issue.

One limitation of our study is that, although our cueing manipulation resulted in numerical differences between cued and uncued targets in the expected direction, this difference was not significant (when we limited the analysis to items that were remembered correctly prenap, the difference was marginal, $p = 0.07$). One possible explanation for this nonsignificant effect is that the prenap and postnap tests differed in a way that was likely to engage different sorts of retrieval processing. The retrieval cue in each trial of the prenap test was the sound associated with the object (picture not shown), whereas the retrieval cue in each trial of the postnap test was the item picture (the sound was not played). We omitted pictures in the prenap test to create fewer distractions for subjects to perform motor imagery, and we omitted sounds in the postnap test to ensure that subjects correctly associated visual items with locations (and were not relying on sound-location memory), as we have done in previous studies (Antony et al., 2018a,b). These differences in retrieval demands may have added noise to prepost comparisons, obscuring TMR effects.

Another possible limitation stems from the inclusion of nontarget stimuli; these stimuli were included to ensure that subjects encoded the semantic nature of the target stimuli and not just the spatial location. While these could feasibly become a source of

interference during TMR, such interference was not reflected in the behavior (memory for cued and uncued nontarget stimuli did not differ), it would have been equal for both left and right target cues (because nontarget stimuli were present for every trial), and any lateralized neural activity it may have caused was clearly overridden by lateralized evidence for the target stimuli.

We present evidence for the retrieval of learning content by showing that learning-related (lateralized) neural features can be successfully decoded when the classifier is trained and tested on sleep data (in a cross-validated fashion). This approach resembles the approach taken by recent papers that also trained and tested using only sleep data (Schönauer et al., 2017; Cairney et al., 2018), and extends these approaches by selecting only features that discriminated between learning conditions during wake. However, it is important to note that a classifier trained on lateralized differences from wake did not show above-chance classification during sleep, suggesting that the specific neural pattern of lateralized differences from wake was not reinstated in response to TMR cues during sleep.

In conclusion, our findings support the idea that retrieval of learning-related neural activity during sleep benefits memory. Furthermore, retrieval following an auditory cue was linked with spindle occurrence, in keeping with our prior description of spindle refractoriness (Antony et al., 2019). Memory retrieval was facilitated when the cue was followed by a spindle, as evidenced both by later memory performance and by EEG analyses. Moreover, this paradigm, using lateralized differences as a means of tracking retrieval of specific episodic associations, could be useful for uncovering nuanced relationships between retrieval and memory (Lewis-Peacock and Norman, 2014) and investigating how retrieval differs across physiological states (Andrillon et al., 2016; Jiang et al., 2017; Schönauer et al., 2017) and human populations (Mander et al., 2013; Westerberg et al., 2015).

References

Andrillon T, Poulsen AT, Hansen LK, Léger D, Kouider S (2016) Human markers of responsiveness to the environment in human sleep. J Neurosci 36:6583–6596.

Antony JW, Paller KA (2017) Hippocampal contributions to declarative memory consolidation during sleep. In: The hippocampus from cells to systems (Hannula DE, Duff MC, eds), pp 245–280. New York: Springer.

Antony JW, Cheng LY, Brooks PP, Paller KA, Norman KA (2018a) Competitive learning modulates memory consolidation during sleep. Neurobiol Learn Mem 155:216–230.

Antony JW, Piloto L, Wang M, Pacheco P, Norman KA, Paller KA (2018b) Sleep spindle refractoriness segregates periods of memory reactivation. Curr Biol 28:1736–1743.e4.

Antony JW, Schönauer M, Staresina BP, Cairney SA (2019) Sleep spindles and memory reprocessing. Trends Neurosci 42:1–3.

Belal S, Cousins J, El-Deredy W, Parkes L, Schneider J, Tsujimura H, Zoumpoulaki A, Perapoch M, Santamaria L, Lewis P (2018) Identification of memory reactivation during sleep by EEG classification. Neuroimage 176:203–214.

Bendor D, Wilson MA (2012) Biasing the content of hippocampal replay during sleep. Nat Neurosci 15:1439–1444.

Bergmann TO, Mölle M, Diedrichs J, Born J, Siebner HR (2012) Sleep spindle-related reactivation of category-specific cortical regions after learning face-scene associations. Neuroimage 59:2733–2742.

Cairney SA, Guttesen AA, El Marj N, Staresina BP (2018) Memory consolidation is linked to spindle-mediated information processing during sleep. Curr Biol 28:948–954.e4.

Cox R, Hofman WF, de Boer M, Talamini LM (2014) Local sleep spindle modulations in relation to specific memory cues. Neuroimage 99:103–110.

Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 134:9–21.

Deuker L, Olligs J, Fell J, Kranz TA, Mormann F, Montag C, Reuter M, Elger CE, Axmacher N (2013) Memory consolidation by replay of stimulus-specific neural activity. J Neurosci 33:19373–19383.

Farthouat J, Gilson M, Peigneux P (2017) New evidence for the necessity of a silent plastic period during sleep for a memory benefit of targeted memory reactivation. Sleep Spindl Cortical Up States 1:14–26.

Jamalabadi H, Alizadeh S, Schönauer M, Leibold C, Gais S (2016) Classification based hypothesis testing in neuroscience: below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. Hum Brain Mapp 37:1842–1855.

Ji D, Wilson MA (2007) Coordinated memory replay in the visual cortex and hippocampus during sleep. Nat Neurosci 10:100–107.

Jiang X, Shamie I, Doyle WK, Friedman D, Dugan P, Devinsky O, Eskandar E, Cash SS, Thesen T, Halgren E (2017) Replay of large-scale spatiotemporal patterns from waking during subsequent NREM sleep in human cortex. Sci Rep 7:17380.

Johnson LA, Blakely T, Hermes D, Hakimian S, Ramsey NF, Ojemann JG (2012) Sleep spindles are locally modulated by training on a brain-computer interface. Proc Natl Acad Sci U S A 109:18583–18588.

Kouider S, Andrillon T, Barbosa LS, Goupil L, Bekinschtein TA (2014) Inducing task-relevant responses to speech in the sleeping brain. Curr Biol 24:2208–2214.

Lehmann M, Schreiner T, Seifritz E, Rasch B (2016) Emotional arousal modulates oscillatory correlates of targeted memory reactivation during NREM, but not REM sleep. Sci Rep 6:39229.

Lewis-Peacock JA, Norman KA (2014) Competition between items in working memory leads to forgetting. Nat Commun 5:5768.

Mander BA, Rao V, Lu B, Saletin JM, Lindquist JR, Ancoli-Israel S, Jagust W, Walker MP (2013) Prefrontal atrophy, disrupted NREM slow waves and impaired hippocampal-dependent memory in aging. Nat Neurosci 16:357–364.

Oudiette D, Paller KA (2013) Upgrading the sleeping brain with targeted memory reactivation. Trends Cogn Sci 17:142–149.

Oudiette D, Antony JW, Creery JD, Paller KA (2013) The role of memory reactivation during wakefulness and sleep in determining which memories endure. J Neurosci 33:6672–6678.

Rasch B, Büchel C, Gais S, Born J (2007) Odor cues during slow-wave sleep prompt declarative memory consolidation. Science 315:1426–1429.

Rechtschaffen A, Kales A (1968) A manual of standardized terminology, techniques and scoring system of sleep stages in human subjects. Washington, DC: Public Health Service, U.S. Government Printing Office.

Rothschild G, Eban E, Frank LM (2017) A cortical-hippocampal-cortical loop of information processing during memory consolidation. Nat Neurosci 20:251–259.

Rudoy J, Voss J, Westerberg C, Paller K (2009) Strengthening individual memories by reactivating them during sleep. Science 326:2009.

Schönauer M, Alizadeh S, Jamalabadi H, Abraham A, Pawlizki A, Gais S (2017) Decoding material-specific memory reprocessing during sleep in humans. Nat Commun 8:15404.

Schreiner T, Lehmann M, Rasch B (2015) Auditory feedback blocks memory benefits of cueing during sleep. Nat Commun 6:8729.

Schreiner T, Doeller CF, Jensen O, Rasch B, Staudigl T (2018) Theta phase-coordinated memory reactivation reoccurs in a slow-oscillatory rhythm during NREM sleep. Cell Rep 25:296–301.

Shanahan LK, Gjorgieva E, Paller KA, Kahnt T, Gottfried JA (2018) Odor-evoked category reactivation in human ventromedial prefrontal cortex during sleep promotes memory consolidation. Elife 7:e39681.

Westerberg CE, Florczak SM, Weintraub S, Mesulam MM, Marshall L, Zee PC, Paller KA (2015) Memory improvement via slow-oscillatory stimulation during sleep in older adults. Neurobiol Aging 36:2577–2586.

Wilson MA, McNaughton BL (1994) Reactivation of hippocampal ensemble memories during sleep. Science 265:676–679.