Behavioral/Cognitive

Goal-Directed Modulation of Neural Memory Patterns: Implications for fMRI-Based Memory Detection

Melina R. Uncapher,[1]* J. Tyler Boyd-Meredith,[1]* Tiffany E. Chow,[3] Jesse Rissman,[3] and Anthony D. Wagner[1,2]

[1]Department of Psychology and [2]Neurosciences Program, Stanford University, Stanford, California 94305, and [3]Department of Psychology, University of California, Los Angeles, Los Angeles, California 90095

Remembering a past event elicits distributed neural patterns that can be distinguished from patterns elicited when encountering novel information. These differing patterns can be decoded with relatively high diagnostic accuracy for individual memories using multivoxel pattern analysis (MVPA) of fMRI data. Brain-based memory detection—if valid and reliable—would have clear utility beyond the domain of cognitive neuroscience, in the realm of law, marketing, and beyond. However, a significant boundary condition on memory decoding validity may be the deployment of "countermeasures": strategies used to mask memory signals. Here we tested the vulnerability of fMRI-based memory detection to countermeasures, using a paradigm that bears resemblance to eyewitness identification. Participants were scanned while performing two tasks on previously studied and novel faces: (1) a standard recognition memory task; and (2) a task wherein they attempted to conceal their true memory state. Univariate analyses revealed that participants were able to strategically modulate neural responses, averaged across trials, in regions implicated in memory retrieval, including the hippocampus and angular gyrus. Moreover, regions associated with goal-directed shifts of attention and thought substitution supported memory concealment, and those associated with memory generation supported novelty concealment. Critically, whereas MVPA enabled reliable classification of memory states when participants reported memory truthfully, the ability to decode memory on individual trials was compromised, even reversing, during attempts to conceal memory. Together, these findings demonstrate that strategic goal states can be deployed to mask memory-related neural patterns and foil memory decoding technology, placing a significant boundary condition on their real-world utility.

Key words: **countermeasures; episodic retrieval; functional MRI; neurolaw; pattern classification**

Introduction

Growing evidence indicates that it is possible to decode the presence or absence of memory for a stimulus or event from distributed patterns of human brain activity, as measured by functional MRI (fMRI) and multivoxel pattern analysis (MVPA; Johnson et al., 2009; McDuff et al., 2009; Chadwick et al., 2010; Quamme et al., 2010; Rissman et al., 2010; Polyn et al., 2012; Poppenk and Norman, 2012; Rissman and Wagner, 2012). The rapidly emerging literature on fMRI-based memory decoding not only informs neurocognitive theories of memory but also has implications for law, marketing, and beyond (Meegan, 2008). For example, a reliable and validated method to detect memory could advance the forensic ability of the criminal justice system to determine whether a suspect has guilty knowledge of crime-relevant information (Greely, 2011) or whether an eyewitness recognizes a critical event element. Given the high diagnostic accuracy observed in some fMRI-based memory decoding studies (up to 70–90%; Rissman et al., 2010), it may be tempting to conclude that these approaches have forensic utility for uncovering an individual's memory states and perhaps also their experiential history with event information.

However, fMRI-based memory detection techniques are still under development, with many significant challenges remaining before determining their appropriateness for field use (Brown and Murphy, 2010; Verschuere et al., 2011). One of the most significant open questions is whether memory decoding is vulnerable to "countermeasures": strategies deployed to mask memory signals and "beat" detection tests (Farah et al., 2014). Rissman et al. (2010) reported indirect evidence suggesting a vulnerability to strategic goal states, because the ability to detect previously encountered from novel faces was reduced to near chance when participants' memory was implicitly, rather than explicitly, probed. However, other data suggest that lack of attention to one's mnemonic state may not always thwart memory classification. For example, Kuhl et al. (2013) were able to decode memory details even when participants were not instructed to retrieve those details. Together, these findings reveal a need to identify

Received Dec. 18, 2014; revised April 23, 2015; accepted April 25, 2015.

Author contributions: M.R.U. and A.D.W. designed research; M.R.U. and T.E.C. performed research; M.R.U., J.T.B.-M., and J.R. contributed unpublished reagents/analytic tools; M.R.U., J.T.B.-M., and T.E.C. analyzed data; M.R.U., J.T.B.-M., and A.D.W. wrote the paper.

This study was supported by a grant from the John D. and Catherine T. MacArthur Foundation to Vanderbilt University, with a subcontract to Stanford University. Its contents do not necessarily represent official views of either the John D. and Catherine T. MacArthur Foundation or the MacArthur Foundation Research Network on Law and Neuroscience (www.lawneuro.org). We thank Hank Greely and Nita Farahany for helpful discussions.

*M.R.U. and J.T.B.-M. contributed equally to this work.

The authors declare no competing financial interests.

Correspondence should be addressed to Melina R. Uncapher, Stanford Memory Laboratory, Stanford University, Stanford, CA 94305-2130. E-mail: melina.u@stanford.edu.

DOI:10.1523/JNEUROSCI.5145-14.2015

 0270-6474/15/358531-15$15.00/0

conditions under which strategic goal states alter neural memory patterns: in particular, can participants willfully conceal their memory states through the use of countermeasures that appear cooperative? Addressing this question not only has implications for delineating the boundary conditions of fMRI methods to detect memory but also for understanding dynamics of goal-directed retrieval processes.

Here we investigated a situation that resembles eyewitness identification and required countermeasures that would appear cooperative on an eyewitness identification test. Participants viewed a series of faces, and their memory for these faces was then probed in one of two ways while undergoing fMRI. In the first test, participants made explicit recognition decisions about previously encountered and novel faces. In the second test, participants attempted to conceal their memory for the previously encountered faces and to feign memory for the novel faces. Using the explicit memory data, we trained classifiers to discriminate activity patterns associated with the subjective experiences of recognition and novelty. We then tested whether the classifiers could decode the participants' memory states when they engaged in countermeasures.

Materials and Methods

Participants. Twenty-four healthy, right-handed male participants were recruited from Stanford University and its surrounding communities. Participants were aged 18–31 years, with a mean ± SD age of 23 ± 4.29 years, were native English speakers with no history of neurological complications, and were either African American (AA; $n = 8$) or European American (EA; $n = 16$) according to self-report. Participants gave written informed consent, in accordance with Stanford University Institutional Review Board procedures, and were screened for fMRI compatibility.

Experiment. The experiment included two scan sessions conducted ~24 h apart and took approximately 5 h across both scan sessions. Each participant was compensated $20 for each hour of participation. Data from two additional participants were collected but omitted from subsequent analyses because of inadequate or incomplete performance: one was omitted because d' was at chance (−0.08) and the other because the participant withdrew from scanning before completing the experiment.

Stimuli. Face stimuli consisted of 400 color photographs of male faces, of which half were AA and half were EA (data examining the effects of race will be reported separately). Face stimuli were standardized for neutral facial expression and background illumination, and included head and neck only. Stimuli were presented against a gray background with a black central fixation crosshair. For each participant, face stimuli were divided into two samples using stratified random sampling by race to assign stimuli to be presented during the encoding phase (OLD items; 100 AA, 100 EA) or to serve as foil items at retrieval (NEW items; 100 AA, 100 EA).

Day 1: encoding. Participants were scanned while intentionally encoding 200 male faces (100 AA faces and 100 EA faces). Each face was presented for 2 s, with an 8 s interstimulus interval (ISI) for a total of 10 s per trial. Each face was shown twice during the course of the encoding phase: after the full set of 200 stimuli was presented, the same faces were presented again in a different order. Participants were given an elaborative encoding strategy to memorize the stimuli, whereby they were instructed to generate imaginative stories involving the individuals pictured in the stimuli. To confirm that participants were attending to stimuli and engaging in the task, they were instructed to press the right index finger button on a response box after each face appeared. Stimuli were presented across eight runs, with 50 faces per run. The first four runs consisted of first presentation of study stimuli, and the second four runs consisted of second presentation of study stimuli. At the end of each run (for both day 1 and day 2), participants were encouraged to take as long a break as they desired, self-terminating the break via button press. Breaks ranged between 10 and 55 s across participants.

Day 2: retrieval. Approximately 24 h later, participants returned for a scanned retrieval phase. In this phase, participants were presented with all 200 studied faces, interspersed pseudorandomly with 200 novel faces, and were asked to provide an "old/new" recognition response on each trial. Each face was presented for 2 s with an 8 s ISI. Thus, participants were given 10 s to respond to each face but were instructed that both speed and accuracy were important, with the latter emphasized to ensure adequate behavioral performance. Stimuli were presented across eight scanning runs. Runs were split into two tasks, with the first four comprising an "explicit memory" task and the second four of a "concealed memory" task. Each task contained 50 OLD AA faces, 50 OLD EA faces, 50 NEW AA faces, and 50 NEW EA faces. Stratified random sampling of the stimuli by condition (OLD and NEW) was used to assign faces to each of the two tasks. Stimuli were balanced for race within each run and presented pseudorandomly such that no more than three presentations of the same race or condition (OLD or NEW) appeared consecutively. All four runs of the explicit memory task were always performed before runs for the concealed memory task to maintain naivety to the concealed memory task (for data demonstrating that this run structure did not pose problems for the decoding findings, see Results). Instructions for the explicit memory task were given before participants were placed in the fMRI scanner. Instructions for the concealed memory task were only given after completion of the explicit memory task, while participants were in the scanner. Verbal instructions were supplemented by visual, onscreen instructions. Participants alternated responding with their left or right hand on each run.

Explicit memory task. In this task, participants made explicit memory judgments: they depressed a button with their (1) index finger if they judged a stimulus to be "old," i.e., previously encountered during the encoding session, or (2) middle finger if they judged the face to be unstudied, or "new."

Concealed memory task. In this task, participants were instructed to try to prevent computer algorithms from identifying whether or not they correctly recognized a stimulus. To help conceal the true memory status for each stimulus, participants were instructed to (1) use instructed cognitive strategies and (2) reverse the response mapping used to indicate their mnemonic states. We specifically selected countermeasure strategies that required attention to and classification of items based on the memory signal evoked on each trial to determine whether an ability to mask the neural correlates of memory is simply a consequence of lack of attention to memory (a possible interpretation of the preliminary implicit memory findings by Rissman et al., 2010). Requiring attention to memory not only allowed us to investigate the neurobiological consequences of goal-directed modulation of memory signals but also may more closely resemble real-world scenarios in which participants feign cooperativity.

Specifically, for faces thought to be studied previously, participants were instructed to indicate a "new" response and to rapidly switch to a cognitive strategy to conceal the neural correlates of their recognition: they were asked to focus on the technical and/or photographic aspects of the stimulus to which they had not attended previously, such as the exposure, lighting, line edges, etc. Here, it was stressed that the computer algorithm was sensitive enough to pick up memory signals and that, after determining that it was a studied face, they should—to the best of their ability—try to keep the memory for the face from coming to mind by attending to the perceptual aspects of the photographs. Conversely, for stimuli perceived to be unstudied, participants were instructed to indicate an "old" response and to use cognitive strategies to elicit the neural correlates of retrieval, i.e., bring to mind a known individual that resembled the face, and relive any memories associated with the individual. Again, it was stressed that the computer algorithm was sensitive enough to detect novelty signals, so they should rapidly switch to generating memories for the face they determined to be novel. For both old and new faces, participants were instructed to use the appropriate cognitive strategy for the entire duration of the trial (10 s). Experimenters confirmed that all participants understood the concealed memory task before proceeding with the experiment.

Functional localizer task. After retrieval, participants performed two runs of a block design functional localizer experiment (Weiner and Grill-Spector, 2010). In this task, photographs were presented of intact AA faces, intact EA faces, scrambled AA faces [wherein facial features (eyes,

nose, and mouth) were rearranged within the face], and scrambled EA faces, scenes, abstract objects, and body parts. Each image was presented for 0.8 s with a 0.2 s ISI in 12 s blocks. Each run consisted of two blocks of each condition presented in pseudorandom order, interspersed with four blank blocks. Participants were instructed to respond when two consecutive images were identical. The localizer was performed to permit analyses of the encoding data (not reported here) and thus is not relevant to the present data.

fMRI data acquisition. Whole-brain imaging was performed with a 3.0 T GE (Discovery MR750) MR scanner. A T2-weighted anatomical volume was collected immediately before the experimental runs, using a T2-weighted flow-compensated spin-echo pulse sequence. A T1-weighted whole-brain spoiled gradient recalled high-resolution anatomical image was collected at the end of day 1 (encoding). Each functional volume, collected with a T2*-weighted echo planar imaging pulse sequence, consisted of 36 slices acquired in an interleaved ascending progression, parallel to the anterior commissure–posterior commissure plane. Functional volumes were collected as a 64×64 matrix using a repetition time (TR) of 2 s, echo time of 30 ms, and a field of view of 21 cm. In-plane resolution was 3.28 mm^2, and slice thickness was 3.3 mm. A total of 260 volumes were collected for each of the eight runs, with the initial four volumes of each run discarded to allow for T1 equilibration.

Univariate fMRI analyses. Statistical Parametric Mapping (SPM8; Wellcome Department of Cognitive Neurology, London, UK; http://www.fil.ion.ucl.ac.uk/spm/software/spm8), run in MATLAB 7.7 (R2008b; MathWorks), was used for both data preprocessing and univariate analysis.

Standard preprocessing procedures were applied to the data. All functional volumes were slice-time corrected to account for acquisition time differences between slices, with the middle slice in time used as a reference. All functional volumes were motion corrected and spatially realigned to the first volume, followed by realignment to the mean volume of the session. The T2-weighted anatomical volume from the day 2 (retrieval) session was coregistered to the mean functional volume, the T1-weighted anatomical volume was then coregistered to this coregistered T2-weighted volume, and then the T1-weighted volume was segmented into gray matter, white matter, and CSF, with the resulting images normalized to templates in Montreal Neurological Institute (MNI) space. Functional volumes were normalized into standard space based on the transformation parameters obtained during segmentation, and resampled into 4 mm^3 voxels. All images were then spatially smoothed with an 8 mm full-width at half-maximum (FWHM) Gaussian kernel.

First-level general linear models (GLMs) were computed for each participant by modeling each retrieval event as a 2 s epoch and convolving each event with the canonical hemodynamic response function. Temporal and dispersion derivatives were additionally modeled to capture variance associated with onset and duration, respectively. A high-pass filter of [1/128] Hz was used to eliminate low-frequency noise. An AR(1) model was used to account for serial autocorrelations. Eight regressors of interest were included in the GLM: hits, correct rejections (CRs), misses, and false alarms (FAs), in both the explicit memory and concealed memory tasks. Other regressors modeled the 8 s lead-out time at the end of each run and any items for which no response was recorded, as well as six regressors modeling movement parameters estimated during realignment. GLM parameters were estimated with classical (restricted maximum likelihood) algorithms. Linear contrasts of the resulting parameter estimates were used to investigate and test effects of interest.

Second-level analyses, in which participants were treated as a random effect, were then conducted with the contrast images generated at the first level for each participant. Contrast effects were examined, thresholded at $p < 0.001$ with at least 16 contiguous voxels [to maintain a familywise error (FWE) rate of $p < 0.05$, as calculated using fMRIstat http://www.math.mcgill.ca/keith/fmristat].

Multivariate fMRI analyses. Classification was implemented in MATLAB using custom code and the framework provided by the Princeton Multi-Voxel Pattern Analysis toolbox (Detre et al., 2006). MVPA was conducted on spatially smoothed and normalized data, and voxels in motor and premotor cortex as well as cerebellum were masked out so that the classifier did not exploit activity differences linked to the motor response made in response to different memory decisions. This mask yielded 23,013 voxels to pass to the classifier. Before classification, additional preprocessing steps were performed. The time series of each voxel was high-pass filtered to remove frequencies <0.01 Hz, detrended to remove linear and quadratic trends, and z-scored so as to normalize the time series of each voxel to have a mean of zero and a variance of one. For the main analyses, fMRI time series data from each voxel were reduced to a single value for each of the 400 test trials by averaging over data acquired at TR3–TR5. For the TR-specific analyses, six separate classifiers were trained and tested on each of six 2-s poststimulus time points, again with one brain activity value per voxel per trial. For the cross-TR analyses, this latter process was repeated, but the six separate classifiers were trained on TR3 (for motivation, see Results) and tested on each of the six TRs.

The brain activity pattern associated with each trial was labeled according to its objective mnemonic status (OLD or NEW), its subjective mnemonic status ("old" or "new"), and the task (explicit or concealed memory), resulting in eight trial types. In each classification analysis, we assessed how accurately the classifier could discriminate between two distinct mnemonic conditions, each defined by a single trial type or a combination of trial types. When training and testing within task, classification performance was assessed separately on each participant's data using a fourfold, leave-one-run-out cross-validation, in which each of the four subsets consisted of trials from one of the four runs of interest. Trials from three subsets were used to train the classifier, and trials from the held-out run were used to test generalization performance. This process was repeated iteratively with each of the four subsets of trials held out, such that unbiased classifier outputs were measured for all of the selected trials. This within-task decoding procedure was modified in only one instance: cross-TR analyses did not use cross-validation to test generalization performance but rather used independent training and testing data. Data were trained within two runs and tested on two held-out runs in a non-iterative manner. For across-task analyses, training data were also independent of testing data (e.g., training on explicit trials and testing on concealed memory trials), and thus cross-validation was also not needed to test generalization performance.

For all classification schemes, trial counts were balanced across classes (via random subsampling) within the training and testing bins before classification to ensure a theoretical null hypothesis classification accuracy rate of 50% and an area under the curve (AUC; see below) of 0.50; analyses with shuffled class labels confirmed that chance classification performance converged around these levels ("null distribution"). After balancing, the data from each voxel were z scored again, such that the mean activity level of each voxel for Class A trials was the inverse of its mean activity level for Class B trials. For each analysis, the entire classification process was run 10 times to obtain stable estimates of performance (independent analyses confirmed that 10 iterations were sufficient to obtain stable performance estimates).

Regularized logistic regression (RLR) was used for all classification procedures. This was determined previously to be an advantageous choice in this classification paradigm by Rissman et al. (2010). This algorithm implemented a multiclass logistic regression function using a softmax transformation of linear combinations of the features (Bishop, 2006) with an additional ridge penalty term as a Gaussian prior on the feature weights. This penalty term provided L2 regularization, enforcing small weights. During classifier training, the RLR algorithm learned the set of feature weights that maximized the log likelihood of the data; feature weights were initialized to zero, and optimization was implemented with Carl Rasmussen's conjugate gradient minimization function (http://www.gatsby.ucl.ac.uk/~edward/code/minimize/) using the gradient of the log likelihood combined with the L2 penalty.

The L2 penalty was set to be half of the additive inverse of a user-specified parameter λ, multiplied by the square of the L2 norm of the weight vector for each class, added over classes. We elected to set this free λ parameter to a fixed value of 10 for all analyses reported in this study.

Assessing classifier performance. After fitting the RLR model parameters using the training set data, each brain activity pattern (i.e., trial) from the test set was then fed into the model and yielded an estimate of the probability of that example being from Class A or Class B (by construction, these two values always sum to one). These probability values were con-

catenated across all cross-validation testing folds and then ranked. The true positive (hit) rate and false positive (FA) rate of the classifier were calculated at 80 fixed cutoff thresholds along the probability continuum to generate receiver operating characteristic (ROC) curves. The AUC values associated with these curves were computed as described by Fawcett (2004) and can be interpreted formally as the probability that a randomly chosen member of one class has a smaller estimated probability of belonging to the other class than has a randomly chosen member of the other class. Stated another way, the AUC indexes the mean accuracy with which a randomly chosen pair of Class A and Class B trials could be assigned to their correct class (0.5 is random performance; 1.0 is perfect performance). If one's goal is high specificity in labeling examples of Class A and is unwilling to tolerate many false positives, one can interrogate the most confident guesses of the classifier. Here we arbitrarily set this threshold to be the top 10% of classification guesses. Note that we report accuracy rather than AUC values when reporting the most confident trials of the classifiers.

Importance maps. For each classification scheme, importance maps were constructed following the procedure described in previous MVPA studies (Johnson et al., 2009; McDuff et al., 2009). The importance value of a voxel provides an index of how much its signal increases or decreases influence the predictions of the classifier. After training, the logistic regression classification procedure yields a set of weight values reflecting the predictive value of each voxel (with positive values indicating that activity increases are generally associated with a Class A outcome and negative values indicating that activity increases are generally associated with a Class B outcome). These weights were then multiplied by the mean activity level of each voxel for Class A trials (which, because of our trial balancing and z-scoring procedure, is the additive inverse of its mean activity level for Class B trials). Voxels with positive values for both activity and weight were assigned positive importance values, voxels with negative activity and weight were assigned negative importance values, and voxels for which the activity and weight had opposite signs were assigned importance values of zero (Johnson et al., 2009; McDuff et al., 2009). Group-level summary maps were created by averaging the importance maps of the individual participants and are displayed in the figures at arbitrary thresholds: 3D-rendered maps thresholded between ± 0.02 and ± 0.5 and 2D-rendered maps between ± 0.05 and ± 0.5 (see Fig. 4) or between ± 0.15 and ± 0.5 (see Fig. 7). As a final note, although importance maps are a useful tool to evaluate which voxels were used by the classifier, these maps should not be interpreted as providing an exhaustive assessment of which voxels are individually informative about the distinction of interest.

Searchlight analyses. Importance maps reveal which voxels provide diagnostic information to the whole-brain classifiers. However, they do not reveal whether data from individual anatomical regions can be used on their own to discriminate hits from CRs. We conducted searchlight analyses to provide local decoding accuracies (Kriegeskorte et al., 2006) across the brain. Of particular interest was whether the regions in which mean blood oxygen level-dependent (BOLD; univariate) signal was modulated significantly by countermeasures (see Fig. 2*A*) also enabled trial-by-trial decoding accuracy that significantly departed from chance. We performed the critical classification (explicit → concealed hits vs CRs) on local spherical masks centered individually on every voxel in the whole-brain mask (excluding voxels in the motor cortex and cerebellum). Each spherical mask included any voxel that touched the edge of the center voxel; thus, the resulting spheres contained 19 voxels, except when the sphere extended beyond the whole-brain mask. To determine whether local decoding accuracies evolved across the trial (as would be expected if participants initially attended to memory signals and then attempted to conceal such signals), we conducted these searchlights separately for each of the six TRs.

We evaluated significance in each of our searchlight spheres as in the prior decoding analyses: AUCs were first computed for 10 classification iterations, and then a null distribution was simulated by computing 10 additional classification iterations using scrambled regressors. We generated group-level t maps showing spheres that reliably discriminated hits from CRs by performing a paired t test of each participant's mean scrambled versus unscrambled AUC value at each voxel, across all 10 iterations. These maps were thresholded at $p < 0.05$ (corrected) by applying a cluster-size threshold derived from Monte Carlo simulations (Xiong et al., 1995) as implemented in the AFNI (Automated Functional Neuro-Imaging) program 3dClustSim. The smoothness for the Monte Carlo simulation was estimated separately for each participant and each time point using the AFNI program 3dFWHMx from the average AUCs achieved across the iterations of scrambled classification. Smoothness was averaged across participants and time points to compute a single smoothness value for each dimension. A voxelwise height threshold of $p < 0.01$ resulted in a cluster size of 22 voxels to reach a cluster-level significance of $p < 0.05$ (FWE) within a given time point. To correct for multiple comparisons across our six time points, we applied a Bonferroni's correction, computing the extent threshold necessary to achieve a cluster-level significance of $p < 0.0083$ (or 0.05/6; FWE) at each time point or $p < 0.05$ (FWE) across space and time. Using this method, we determined that a cluster extent of 29 voxels was required to achieve a cluster-level significance of $p < 0.05$ (FWE) across space and the six time points.

Results

Behavioral performance

Explicit memory task

When participants truthfully reported their mnemonic experience elicited by each test face, they achieved a mean ± SD hit rate (rate at which OLD images accurately judged "old") of 0.73 ± 0.12 and an FA rate (rate at which NEW images inaccurately judged "old") of 0.27 ± 0.10, resulting in a mean d' of 1.27 ± 0.56. Mean response times (RTs) were faster for correct answers (hits, 1.74 ± 0.60 s; CRs, 2.10 ± 0.75 s) than for incorrect answers (FAs, 2.30 ± 1.01 s; misses: 2.33 ± 0.81 s; $t_{(23)} = 5.15$, $p = 3.22 \times 10^{-5}$). Hit responses were faster than CR responses ($t_{(23)} = 5.44$, $p = 1.56 \times 10^{-5}$).

Concealed memory task

When participants engaged in countermeasures, during which they responded contrary to their true mnemonic experience, they achieved a mean hit rate (rate at which OLD images were reported as "new," reflecting that the participant accurately believed the image to be OLD) of 0.65 ± 0.12. Participants exhibited a mean FA rate (rate at which NEW images were reported as "new" indicating that the participant inaccurately believed the image to be OLD) of 0.29 ± 0.12. Thus, d' in the concealed memory task was 0.99 ± 0.54. Mean RTs were again faster for correct answers (hits, 2.09 ± 1.08 s; CRs, 2.14 ± 1.09 s) than for incorrect answers (FAs, 2.17 ± 1.13 s; misses: 2.10 ± 1.15 s; $t_{(23)} = 5.85$, $p = 5.79 \times 10^{-6}$), but there was no difference between hits and CRs ($t_{(23)} = 0.16$, $p = 0.87$).

Comparing explicit and concealed memory tasks

Mean d' was significantly greater in the explicit memory than in the concealed memory task ($t_{(23)} = 2.99$, $p = 6.6 \times 10^{-3}$). There were no task differences in mean RT for any memory outcome (all p values > 0.05). However, there was a significant interaction between task and memory, in that the average difference in RT for hits and CRs was greater in the explicit memory than in the concealed memory task ($t_{(23)} = 6.25$, $p = 2.22 \times 10^{-6}$). This differential effect of memory on RT as a function of task may follow from the difference in d' between the explicit and concealed memory conditions and is likely a consequence of the dual-task nature of the concealed memory condition: participants were required to first determine whether faces were old or new and then rapidly switch to a memory/novelty concealment strategy while also reversing their motor responses.

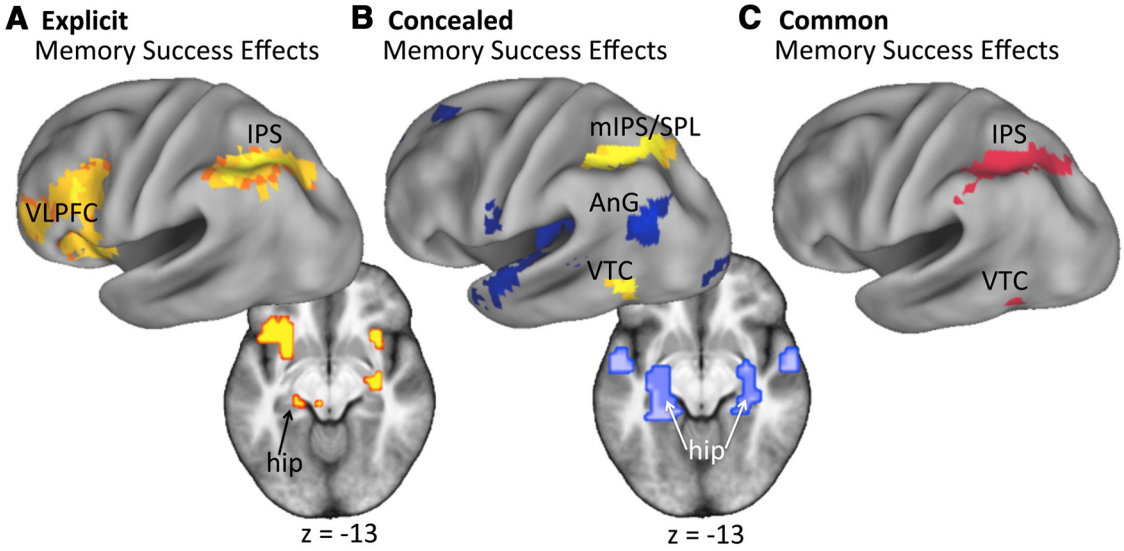

Figure 1. Memory success effects for each task. ***A***, In the explicit memory task, hits elicited greater activity than CRs (Memory Success Effects) in various regions, including those implicated previously in successful memory retrieval. ***B***, In the concealed memory task, memory success effects (warm colors) were found in a similar region in the IPS, whereas the reverse comparison (CRs > hits; cool colors) revealed activity in the left AnG and bilateral hippocampus. ***C***, Memory success effects overlapped in the two tasks in several regions, including the IPS and VTC. Activity is rendered on 3D Caret inflated brain and 2D mean across-subject brain, both in standardized MNI space; height and extent thresholds: $p < 0.001$, $k = 16$. hip, Hippocampus; mIPS/SPL, medial IPS/SPL.

Univariate fMRI analyses

We first investigated the question of whether participants could engage in strategic countermeasures to modulate memory-related BOLD signal across trials (i.e., univariate fMRI responses). To do so, we identified "memory success effects" (hits > CRs) for each task separately and then determined where memory success effects were common across tasks, as well as where they were modulated by task. We then investigated whether the ability to modulate memory success effects was influenced by memory strength.

Memory success effects by task

Greater activity for hits than CRs in the explicit memory task was observed in many regions identified previously in fMRI studies of recognition memory retrieval (for meta-analyses, see McDermott et al., 2009; Spaniol et al., 2009; Kim, 2010; Hutchinson et al., 2014), including the left intraparietal sulcus (IPS) and ventrolateral prefrontal cortex (VLPFC), as well as the left hippocampus (Fig. 1*A*). Similarly, memory success effects in the concealed memory task were identified in the left medial IPS/superior parietal lobule (SPL), with additional clusters in the right IPS and left ventral temporal cortex (VTC; Fig. 1*B*, warm colors). Although no regions showed the reverse effect (CRs > hits) in the explicit task, a number of regions exhibited a reversed effect in the concealed memory task, including the bilateral hippocampus and left angular gyrus (AnG; Fig. 1*B*, cool colors).

Activity in the hippocampus and AnG is often associated with episodic recollection (for meta-analyses, see Kim, 2010; Hutchinson et al., 2014); because these regions showed a reverse memory success effect in the concealed memory task, this pattern offers initial neural support that participants were able to successfully execute the instructed countermeasures in the concealed memory task. In other words, the finding of greater activity for CRs than hits in the concealed condition suggests that participants engaged successfully in greater recollection when cued to do so by faces they determined were novel (concealed memory CRs) relative to faces they determined were studied (concealed memory hits), for which they were to disengage from their memories of the face and attend to the novel aspects of the photograph.

Common memory success effects

Given that participants were required to determine whether a face was old or new in both tasks, we next sought to determine whether there were any regions that differentiated hits from CRs in both tasks. To do so, we inclusively masked the foregoing memory success contrasts (at $p < 0.01$ each, to result in a conjoint threshold of $p < 0.001$). The outcome of this masking procedure revealed effects in the left IPS and left VTC (Fig. 1*C*).

Memory success effects modulated by task

Although common memory success effects were observed in the left IPS and VTC, the apparent differences in regional activity for the two tasks reported in Figure 1, *A* and *B*, suggests that, at least in part, participants were able to modulate their memory-related neural responses between the explicit and concealed memory tasks. A direct test of whether memory success effects differed statistically between tasks revealed greater memory success effects in the explicit than the concealed memory task in many regions, including the bilateral hippocampus, left AnG, and VLPFC (Fig. 2*A*). In these regions, the memory success effects observed in the explicit task were reversed in the concealed task, such that activity elicited by concealed memory CRs was greater than that elicited by concealed memory hits. In contrast, the bilateral IPS exhibited greater memory success effects in the concealed relative to the explicit memory task (Fig. 2*B*).

Relationship between memory strength and memory success effects

We hypothesized that it would prove more difficult to modulate memory signals in the concealed memory task when memory for the faces was relatively strong. We tested this prediction by extracting parameter estimates from the left AnG and bilateral hippocampal clusters identified in the previous contrast (memory × task interaction; Fig. 2) and examining whether they tracked memory strength across participants. Specifically, for the concealed memory task, we regressed d' against the memory success effects (hits > CRs parameter estimates). As illustrated in Figure 3, the AnG showed a positive predictive relationship between memory strength and memory success effects, such that participants with stronger memories in the concealed memory task were

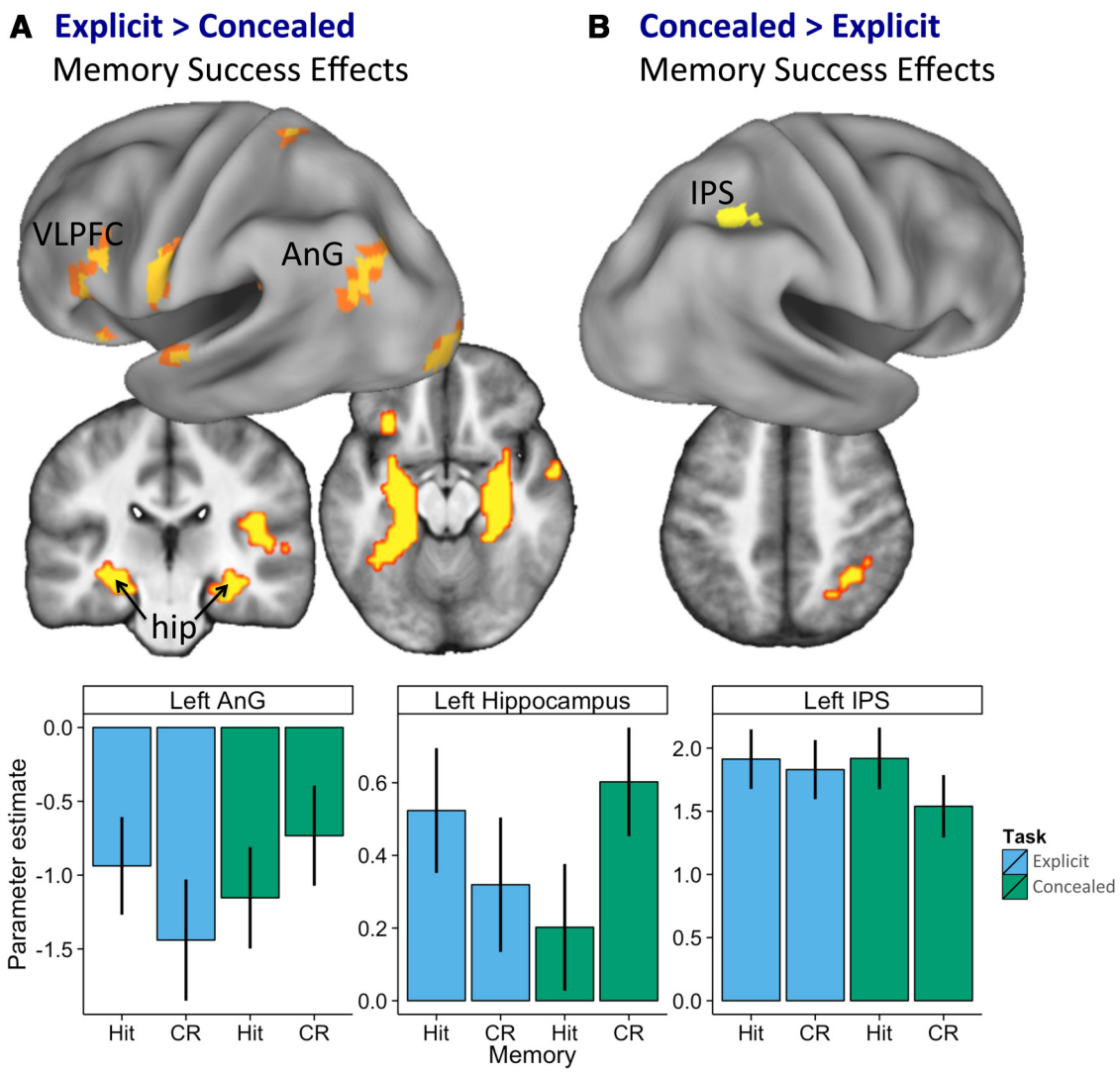

Figure 2. Memory success effects are modulated by attempts to conceal memory. ***A***, Memory success effects reverse between the explicit and concealed memory tasks in regions associated with memory retrieval, suggesting that participants were able to successfully deploy countermeasures. ***B***, Memory success effects were greater in the concealed versus explicit memory task in the right IPS. Activity rendered as described in Figure 1. Graphs depict mean univariate β weights across participants (errors indicate SEM) for clusters in the left hippocampus, AnG, and IPS (***B***).

less likely to show reversed memory success effects in the left AnG ($r = 0.36$, $p = 0.04$). In other words, their memory success effects persisted despite attempts to conceal their memory. This finding suggests that participants with stronger memories had greater difficulty modulating their AnG activity during the concealed task. Interestingly, neither hippocampal cluster showed a significant relationship between memory strength and activity (right hippocampus, $r = -0.05$, $p = 0.82$; left hippocampus, $r = -0.22$, $p = 0.29$), and the slopes of the correlations differed between the AnG and right hippocampus (Williams $t_{(21)} = -2.49$, $p = 0.021$) and marginally differed for the left hippocampus (Williams $t_{(21)} = -1.62$, $p = 0.12$). Together, these findings suggest that participants with stronger memories were less able to exert goal-directed control over memory-related activity in the left AnG, but this appeared not to be the case in the bilateral hippocampus.

Multivariate fMRI analyses

Our central question is whether use of cognitive (goal-directed) countermeasure strategies would enable participants to mask neural patterns related to memory, thus affecting the ability of multivariate techniques to read out their memory states for individual events. Accordingly, we next assessed the ability of MVPA classifiers to decode the memory status of individual retrieval trials by (1) first training and testing a classifier on data from the standard recognition memory task (explicit memory task) and (2) then assessing whether this classifier could also decode memory when participants were attempting to conceal their memory states (concealed memory task). Our process model suggested three alternative scenarios to test; we begin by explicating our process model and then describe each hypothesis in turn.

Process model and hypotheses

Given the standard recognition memory instructions in the explicit task, it is likely that, after encountering each face in this task, participants evaluated the strength of the memory evidence elicited by the face and endorsed it as "old" if it passed a certain threshold of "oldness" or "new" if it passed a different threshold of "novelty" or if it elicited only weak memory evidence [we remain agnostic as to whether participants adopted a two-high-threshold model (Ratcliff, 1978; Snodgrass and Corwin, 1988) or a signal detection model (Green and Swets, 1966) of memory

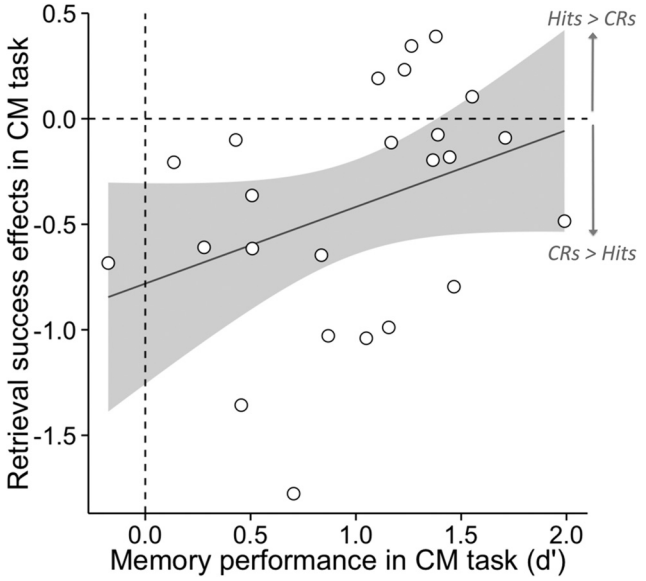

Figure 3. Memory strength in the concealed memory task predicts whether participants can modulate retrieval success effects in the left AnG. Clusterwise β values were extracted from a left AnG (L AnG) cluster identified in the univariate task × memory interaction (Fig. 2*A*), and *d'* in the concealed memory task was regressed against β values in the concealed memory task (hits > CRs). This regression reveals that participants with superior memory performance in the concealed memory task were less likely to show inverse retrieval success effects (CRs > hits) in the AnG, suggesting that participants with stronger memories were less able to exert goal-directed control over memory-related activity in the left AnG.

decisions, because it does not change our predictions or interpretations].

In the concealed task, participants were instructed to make a memory decision to determine which of the two countermeasure strategies to deploy ("feign memory" or "feign novelty"). Thus, participants likely initially followed a similar process as described for the explicit task, followed by deployment of one of the two countermeasure strategies. We intentionally required participants' decision tree to include attention to memory signals to maximize the possibility of detecting memory in the face of countermeasures.

This process model suggested three plausible scenarios regarding the ability of a classifier to decode memory during attempts to conceal memory state:

(1) A classifier trained to detect true memory (explicit hits) from true novelty (explicit CRs) is sensitive enough to detect the initial memory decision on concealed trials (again, this memory decision was made to determine which of the two countermeasure strategies to deploy). In this scenario, explicit hit patterns would consistently predict concealed hit trials, and explicit CR patterns would predict concealed CR trials. Therefore, classification performance would be above chance.

(2) The cognitive operations engaged to feign memory and novelty are qualitatively different than the operations engaged during true memory and novelty. In this scenario, patterns discriminating hits and CRs in the explicit condition would not help discriminate hits and CRs in the concealed condition. Here, explicit hit patterns would not consistently predict either concealed hit or CR patterns, and thus classification performance would be at chance.

(3) The cognitive operations engaged to feign memory elicit patterns that are sufficiently similar to the patterns elicited during true memory, such that feigned memory patterns (concealed CRs) are consistently classified as true memory patterns (explicit hits). Likewise, feigned novelty patterns (concealed hits) are consistently classified as true novelty patterns (explicit CRs). In this case, the classifier would consistently produce Class B guesses in response to Class A patterns (and vice versa) and would therefore exhibit significantly below-chance decoding performance.

To adjudicate between the foregoing scenarios, we trained classifiers to discriminate hits from CRs on trials when participants' responses were representative of their memory decisions (the explicit task). To first establish baseline decoding performance, we tested the ability of this classifier to discriminate hits from CRs on held-out trials from the explicit task. We next tested the ability of the classifier to discriminate hits from CRs on trials when participants were concealing their memory state (concealed task).

Memory decoding during explicit memory judgments

Based on previous findings using a similar explicit memory task (Rissman et al., 2010), we expected the "explicit → explicit" classifier to perform well. Consistent with this expectation, the mean AUC was 0.67 (Fig. 4*A*, blue; null distribution AUC of 0.50, as determined by permutation analyses; mean vs null AUCs, $t_{(23)} = 8.97$, $p = 5.74 \times 10^{-9}$). This performance corresponds to a mean accuracy of 63% when examining all choices made by the classifier, with mean accuracy reaching 73% on trials when the classifier was most confident (i.e., top 10% of trials with the strongest classifier-estimated evidence; Fig. 5*A*, blue).

The foregoing analyses interrogated memory decoding using data collapsed across multiple peristimulus time points (TR3—TR5; see Materials and Methods). Because a memory decision unfolds over time, a set of classifiers trained and tested on each time point during retrieval may provide additional diagnostic information (especially on the concealed memory trials; see below). Figure 4*B* (blue) illustrates that classification performance on each TR in the explicit → explicit scheme increased until TR4 (6–8 s after stimulus onset), when it reached a maximum AUC of 0.66. All but the first two TRs exhibited performance significantly above chance (TR3, $p = 3.64 \times 10^{-6}$; TR4, $p = 9.44 \times 10^{-10}$; TR5, $p = 2.52 \times 10^{-7}$; TR6, $p = 6.76 \times 10^{-5}$), surviving Bonferroni's correction for multiple comparisons ($p_{crit} = 0.0083$ for six TR comparisons).

Maps illustrating the diagnostic value of each voxel to the classifier ("importance maps"; Fig. 4*C*) reveal that several prominent foci biased the classifier toward a hit or a CR choice. Among the regions biasing hit predictions were the bilateral parahippocampal gyrus, bilateral fusiform gyrus, left VTC, left IPS, and bilateral VLPFC. Such a pattern is consistent with a rich retrieval literature and suggests that the classifier collectively relied on signals that represented face information in memory (parahippocampal and fusiform gyri, and VTC), as well as regions thought to subserve attention to and evaluation of accumulating mnemonic evidence (IPS and VLPFC). Conversely, regions that biased CR predictions included the left hippocampus and supramarginal gyrus (SMG) and bilateral frontal eye fields (FEFs). Collectively, the signals in these regions likely reflect the sampling of perceptual space in service of retrieval cue processing (FEF and SMG) and encoding of novel perceptual information (hippocampus).

Memory decoding during attempts to conceal

Although the univariate findings demonstrated that participants were able to successfully modulate their mean signal during the concealed task in regions traditionally associated with retrieval

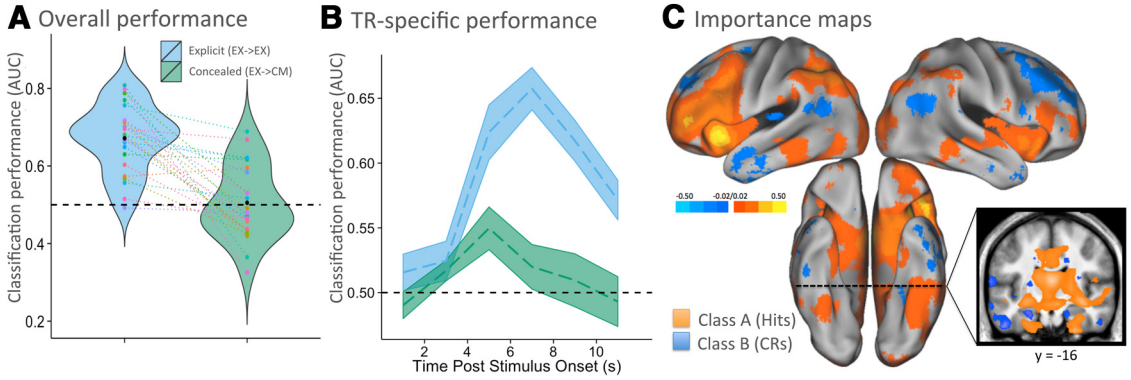

Figure 4. Memory decoding trained on the explicit recognition memory task. ***A***, Performance of memory classifiers (hits vs CRs) trained on standard recognition memory data and tested on held-out trials from the same condition (explicit memory trials, EX⟶EX; blue) or on trials when they were attempting to conceal their true memory state (concealed memory trials, EX⟶CM; green). Violin plots depict the mean AUC for ROC curves (for description of this metric, see text). Plots demonstrate that memory decoding was above chance during a standard recognition memory test (EX⟶EX, blue) but reduced to chance levels when participants attempted to conceal memory (EX⟶CM, green). AUC values are plotted for each participant's data using unique color identifiers, with lines connecting each participant's classification performance, and plot-width depicting participant density in that performance range. ***B***, Classifiers trained and tested on each time point (2 s volume acquisition, TR) revealed similar results as in ***A***, with above-chance decoding on standard memory trials (blue) and chance decoding on concealed memory trials (time point 4–6 s was slightly above chance). ***C***, Maps indicating the diagnostic value that each voxel provides the classifier (importance maps) in biasing a hit choice (warm colors) or a CR choice (cool colors). Error bars in ***B*** indicate SEM. EX, Explicit memory task; CM, concealed memory task; EX⟶EX, classifier trained on explicit trials, tested on held-out explicit trials (4-fold cross-validation); EX⟶CM, classifier trained on explicit trials, tested on concealed trials.

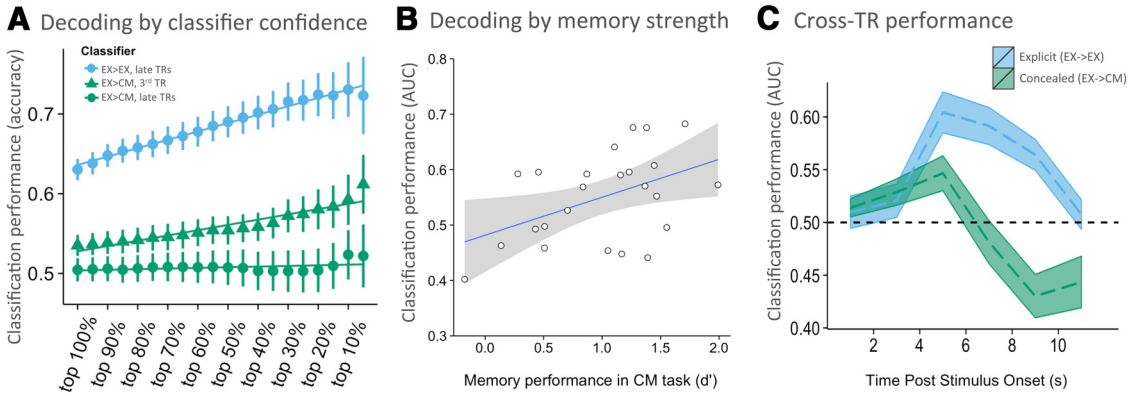

Figure 5. Classification by confidence, memory strength, and time point. ***A***, Classifier accuracy at different levels of confidence, ranging from all trials (top 100%) to the top 10% most confident classifier responses. Values in blue represent classifier performance from the whole-brain explicit ⟶ explicit classifier (EX⟶EX) for late TRs (TR3–TR5), which reveal better performance than the explicit ⟶ concealed classifier (EX⟶CM) for late TRs (green circles), as well as the explicit ⟶ concealed classifier for TR3 (green triangles). ***B***, Individual differences in classification performance on TR3 for concealed trials correlated with memory strength (*d'*) in the concealed trials suggests that the classifier may have detected transient memory signals that emerged before participants' attempts to conceal the signals. ***C***, Classifier trained on memory patterns from TR3 in the explicit task detect novelty patterns on TR5 and TR6 in the concealed condition (green), suggesting that feigned memory can indeed fool a classifier into reliably predicting the opposite response, at later TRs (i.e., when countermeasures have been deployed, after memory state has been determined). EX⟶EX, Classifier trained on explicit trials, tested on held-out explicit trials (4-fold cross-validation); EX⟶CM, classifier trained on explicit trials, tested on concealed trials.

(Fig. 2), this does not necessitate that trial-by-trial decoding accuracy will be affected by attempts to conceal memory and novelty. We next determined whether a classifier trained on data from the explicit memory test could be used to discriminate hits from CRs when participants attempted to conceal memory status (i.e., explicit ⟶ concealed). Again, whether classification performance is revealed to be above, at, or below chance may provide insight into the relationship between true and feigned memory/novelty patterns (see hypotheses above). We found that the explicit-trained classifier performed no differently from chance in the discrimination of hits versus CRs in the concealed memory task (mean AUC of 0.51 vs null distribution AUC of 0.50, as determined by permutation analyses; $t_{(23)} = 0.33$, $p = 0.75$; Fig. 4*A*, green). Given this poor performance, it was particularly important to determine whether the most confident guesses of the classifier could reliably distinguish the two classes. To identify the upper bound of performance, we considered the 10% of trials with the strongest classifier evidence. In this case, accuracy was again no different from chance (mean of 52%; Fig. 5*A*, green circles). These findings suggest that feigned memory and novelty mostly elicit patterns qualitatively different from true memory and novelty.

However, time course analysis (TR-specific classifiers) demonstrated a slight elevation in performance above chance at one time point. Classification based on TR3 (4–6 s after stimulus onset) was modest but significantly above chance (AUC of 0.55 vs null AUC of 0.50; $t_{(23)} = 3.35$, $p = 0.003$, surviving Bonferroni's correction; Fig. 4*B*, green) and the top 10% of classifier evidence trials, based on TR3, resulted in a mean accuracy of 59% (Fig. 5*A*, green triangles). This finding indicates that some memory-related signal may have been present even when countermeasures are in use but that this signal is transient.

Classification performance on TR3 trials is preliminarily suggestive that the classifier may be able to detect transient memory patterns in the concealed task (i.e., offering preliminary support for hypothesis 1 above). Such an interpretation would further

predict that classification performance should scale as participants' memory strength increases. To test this prediction, we examined whether individual differences in classification performance on TR3 for concealed trials correlated with memory strength (d') in the concealed trials. Indeed, classification tended to be better for participants with the strongest memory ($r = 0.46$, $p = 0.025$; Fig. 5*B*), suggesting that the slight and transient above-chance decoding at TR3 may reflect the detection of transient memory signals that emerged before participants' attempts to conceal these signals (bringing the classifier back down to chance levels).

Our process model posits that concealed trials contain true memory/novelty patterns followed by feigned memory/novelty patterns, because participants first attended to their mnemonic state before determining which countermeasure to deploy. To the extent that these feigned memory/novelty patterns emerge only after true memory/novelty patterns, we predict that a classifier trained to detect the emergence of true memory patterns but tested at the emergence of feigned memory patterns would display below-chance performance (i.e., hypothesis 3 above; note: analogous predictions hold for true novelty patterns). To test this prediction, we trained classifiers on TR3 of explicit trials and tested them separately at each of TR4–TR6 on concealed trials. This enabled us to determine whether early memory signals would be detectable during later feigned memory attempts. Note this is different from the TR-specific classifier reported in Figure 4*B*, because those data reflect classifiers trained and tested on the same TR. Also, although our predictions are relevant to TR4–TR6, we report all six TRs in Figure 5*C* to provide a full temporal profile of classification performance.

Strikingly, these explicit → concealed cross-TR classifiers performed slightly but significantly below chance on TR5 (Fig. 5*C*, green; AUC of 0.43 vs null AUC of 0.50; $t_{(23)} = -3.74$, $p = 0.001$) and performed slightly below chance at TR6 (AUC of 0.44 vs null AUC of 0.50; $t_{(23)} = -2.19$, $p = 0.038$). Performance on TR5 but not TR6 survived Bonferroni's correction. The finding of significantly below-chance decoding on TR5 suggests that (1) early emerging true memory patterns in the concealed condition can be subsequently (i.e., at later TRs) masked by attending to novelty, "fooling" a classifier into guessing that the pattern elicited by a recognized item is from the novelty class, and (2) early emerging true novelty patterns can be subsequently masked by feigning memory, "fooling" a classifier into guessing that the pattern elicited by a new item is from the recognized class. The pattern of findings informs our mechanistic predictions by indicating that the classifier exposed some (transient) shared neural patterns between feigned and true memory/novelty. However, it is worth noting that they place a significant boundary condition on the field use of fMRI-based memory decoders, because they suggest that participants are indeed able to willfully and consistently confuse a memory decoder.

Potential issues affecting memory decoding

We next sought to determine whether the overall pattern of findings—that hits and CRs from concealed memory trials could not be differentiated robustly by a classifier trained on explicitly reported hits and CRs—could be explained by (1) differences in memory performance between tasks, (2) poor memory on the explicit task, (3) poor classification of hits versus CRs even in the training data (i.e., explicit trials), or (4) temporal or motor confounds.

First, it is possible the poor performance on the explicit → concealed classifier was attributable to the subset of participants whose memory performance differed markedly between the explicit and concealed memory tasks. To test this possibility, we repeated all analyses on the median split of participants showing the best matched d' between explicit and concealed memory tasks (mean d' difference of best matched participants = 0.17, ranging from 0.02 to 0.29). These analyses revealed that the overall pattern of findings held when examining these 12 participants (mean AUCs: explicit → explicit, 0.66; explicit → concealed, 0.52).

Likewise, it is important to rule out that the low mean accuracy on the explicit → concealed classifier was attributable to poor classification for those participants with poor memory. Accordingly, we recomputed all analyses on a median split of participants with the best memory performance on the explicit task (mean d' of the top 12 participants = 1.74, ranging from 1.26 to 2.31). Again, the overall pattern of findings held on this subset of participants (mean AUCs: explicit → explicit, 0.69; explicit → concealed, 0.48).

It is also conceivable that the explicit → concealed classifier performed poorly because the distinction between hits and CRs in the training data (i.e., explicit trials) was not well learned by the classifier. We tested this possibility by rerunning analyses on a median split of participants, using the 12 participants in whom the explicit → explicit classifier performed the best (mean AUC of top 12 participants, 0.74; ranging from 0.69 to 0.82). The overall pattern of findings held on this subset of participants, in that explicit → concealed performance still did not differ from chance in these participants (mean AUC of 0.48).

An additional concern about the poor performance on the explicit → concealed classifier is the possibility of temporal confounds. Temporal confounds may have arisen because the concealed memory task runs always followed the explicit memory task runs, and thus the classifier may have suffered as a result of factors that drifted or differed simply as a function of time. For instance, participants may have experienced cognitive fatigue or interference from previous trials that served to reduce classification performance on later (concealed memory) trials. Cognitive fatigue or interference could serve to (1) introduce stochastic noise to the later trials or (2) reduce participants' ability to differentiate old and new items on later trials. Both possibilities would reduce the ability of classifiers to identify consistent patterns associated with hits and CRs on these later trials. However, as reported below, we instead found that classifiers trained and tested on the later (concealed) trials yielded very high performance, in fact, better than that of classifiers trained and tested on the earlier (explicit) trials. Moreover, participants reported that the concealed memory task was more cognitively engaging than the earlier explicit memory task, because participants felt they were trying to "beat" a computer algorithm. These observations partially mitigate the concern that these later trials comprised cognitively fatigued—and therefore noisy or more variable—trials that reduce the overall ability to classify trials.

To further address this concern, we next directly tested whether the ability of classifiers to discriminate hits from CRs varied according to explicit task run. To do so, we trained a classifier to discriminate hits from CRs on the first run of the explicit task and tested on each successive run (runs 2–4). Importantly, we observed no effect of run on classification performance ($p = 0.48$, with all individual AUCs greater than null distributions, p values < 0.05). To rule out a temporal distance effect, we then implemented the reverse scheme, training on the last explicit run, and testing on the previous runs (runs 1–3). Here again, there was no effect of run on classification performance ($p = 0.80$, with all

individual AUCs greater than null distributions, p values < 0.05). These analyses suggest that fatigue or interference is not likely to be driving the low explicit → concealed classification performance. Thus, we have objective metrics (high concealed → concealed classification performance and no effect of run on explicit → explicit classification performance) and subjective report data that together suggest that cognitive fatigue or interference is unlikely to be driving the poor explicit → concealed classification performance.

As a final comment on this concern, it is worth noting that, in our previous study (Rissman et al., 2010, their Experiment 2), participants encountered an overall design that paralleled the present, but with the first four scan runs consisting of an implicit memory test and the latter four scan runs consisting of an explicit old/new test analogous to that used in the present study. Thus, the potential concerns of fatigue and interference in the later (concealed) task here would have been present in the later (explicit) task in the study by Rissman et al. (2010). Importantly, Rissman et al. observed that classification of hits versus CRs (explicit → explicit) yielded a mean AUC of 0.71 (later trials), which is comparable with the AUC of 0.67 observed in the present explicit → explicit classifications (earlier trials). Together with the objective and subjective metrics reported above, multiple lines of evidence strongly argue that fatigue and interference effects are unlikely to account for the poor explicit → concealed performance.

As an additional control analysis, it is important to ensure that the classifiers were not simply decoding participants' motor responses. To test this possibility, we trained a set of classifiers to detect left hand versus right hand responses within motor-related voxels (i.e., motor and premotor cortices and cerebellum), across different runs. A motor classifier trained on the first two explicit runs and tested on the last two explicit runs (runs 3 and 4) performed no differently than if tested on the last two concealed runs (runs 7 and 8; mean AUCs of 0.64 and 0.64, respectively, for late TRs). This suggests that fatigue was not driving the reduced performance on the memory-related explicit → concealed classifier. Furthermore, given that index versus middle finger responses mapped to hits versus CRs responses, we trained a set of classifiers to detect left index finger versus middle finger responses in these same motor-related voxels. Interestingly, we found that index versus middle finger responses were difficult to decode, whether within-task (trained on first explicit left hand run, tested on second explicit left hand run: mean AUC of 0.54) or across task (trained on all explicit, tested on all concealed: mean AUC of 0.51). This poor classification performance allows us to be confident that hits versus CRs classifications are not simply decoding motor responses. Together, these tests, which rule out temporal and motor confounds, additionally help mitigate concerns that noncognitive factors (such as scanner drift or scanner noise) may have caused the poor explicit → concealed performance.

It is additionally worth noting that performance of the foregoing classification schemes (explicit → explicit and explicit → concealed) was uncorrelated ($r = 0.30$, $p = 0.15$), further lending weight to the conclusion that decodable memory signals in a traditional retrieval task were essentially undetectable when cognitive countermeasures were deployed.

Finally, it is important to note that poor explicit → concealed decoding performance was not attributable to weak univariate memory signals in the concealed task, particularly in regions associated with memory retrieval, such as hippocampus and AnG (Fig. 1*B*, cool colors).

Searchlight analyses

Whole-brain decoding accuracy can be complemented by searchlight approaches that provide local decoding accuracies (Kriegeskorte et al., 2006). Of particular interest was whether the regions in which mean BOLD (univariate) signal was modulated significantly by countermeasures (Fig. 2*A*) also showed trial-by-trial decoding accuracy that significantly departed from chance. Therefore, we performed the critical classification (explicit → concealed hits vs CRs) on local spherical masks centered individually on every voxel in the whole-brain mask (excluding voxels in motor and premotor cortex and cerebellum). Each spherical mask included any voxel that touched the edge of the center voxel; thus, the resulting spheres contained 19 voxels, except when the sphere extended beyond the whole-brain mask. To determine whether local decoding accuracies evolved across the trial (as would be expected if participants initially attended to memory signals and then attempted to conceal such signals), we conducted these searchlights separately for each of the six TRs.

These searchlight analyses revealed several regions in the medial frontal and parietal cortices that demonstrated slight but significantly above-chance decoding at early time points (0–4 s; Fig. 6), when participants were presumably making their initial memory judgment. The medial parietal cortex is implicated strongly in episodic memory tasks, whether considering univariate data (for meta-analyses, see McDermott et al., 2009; Spaniol et al., 2009; Kim, 2010; Hutchinson et al., 2014), multivariate data (Fig. 4*C*; Rissman et al., 2010), or intracranial EEG data (Foster et al., 2015). Thus, this above-chance decoding demonstrates that some regional activity patterns may be transiently detectable and likely reflects the requirement for participants to first retrieve mnemonic information before deploying countermeasures.

Several clusters yielded below-chance decoding at later time points, when participants were presumably attempting to deploy countermeasures, including clusters centered on the left AnG (Fig. 6) and right FEF. These clusters exhibited below-chance decoding at 6–8 s after stimulus that was robust to multiple comparisons correction across space but did not survive Bonferroni's correction across time points. The AnG finding, while falling one voxel under the 29-voxel extent threshold required to surpass a stringent spatiotemporal correction, aligns well with the univariate data showing that mean BOLD signal in the left AnG was significantly modulated by countermeasures (reversing the typical retrieval success effects; Fig. 2*A*). The finding is also consistent with the importance maps in Figure 4*C*, indicating that the left AnG provided high diagnostic value in influencing the predictions of the whole-brain classifier. By taking a searchlight approach, we determined that local patterns of activity in the left AnG may confuse the classifier at later time points, when participants are attempting to conceal their memory state.

These findings may help adjudicate between potentially conflicting previous data on the decodability of memory under various goal states. On the one hand, Kuhl et al. (2013) showed that mnemonic information could be decoded when participants were not instructed to attend to their mnemonic state but may have been doing so incidentally. On the other hand, when participants were instructed explicitly to perform a task orthogonal to a memory task (make an attractiveness rating on the old and new faces instead of an explicit memory judgment), memory decoding was reduced to at- or near-chance levels (Rissman et al., 2010). Furthermore, recent data suggest that participants can be taught to adopt a global strategy wherein they avoid attending to memory signals elicited by salient, mock crime-related retrieval cues, therefore obscuring event-related potential indices of guilty

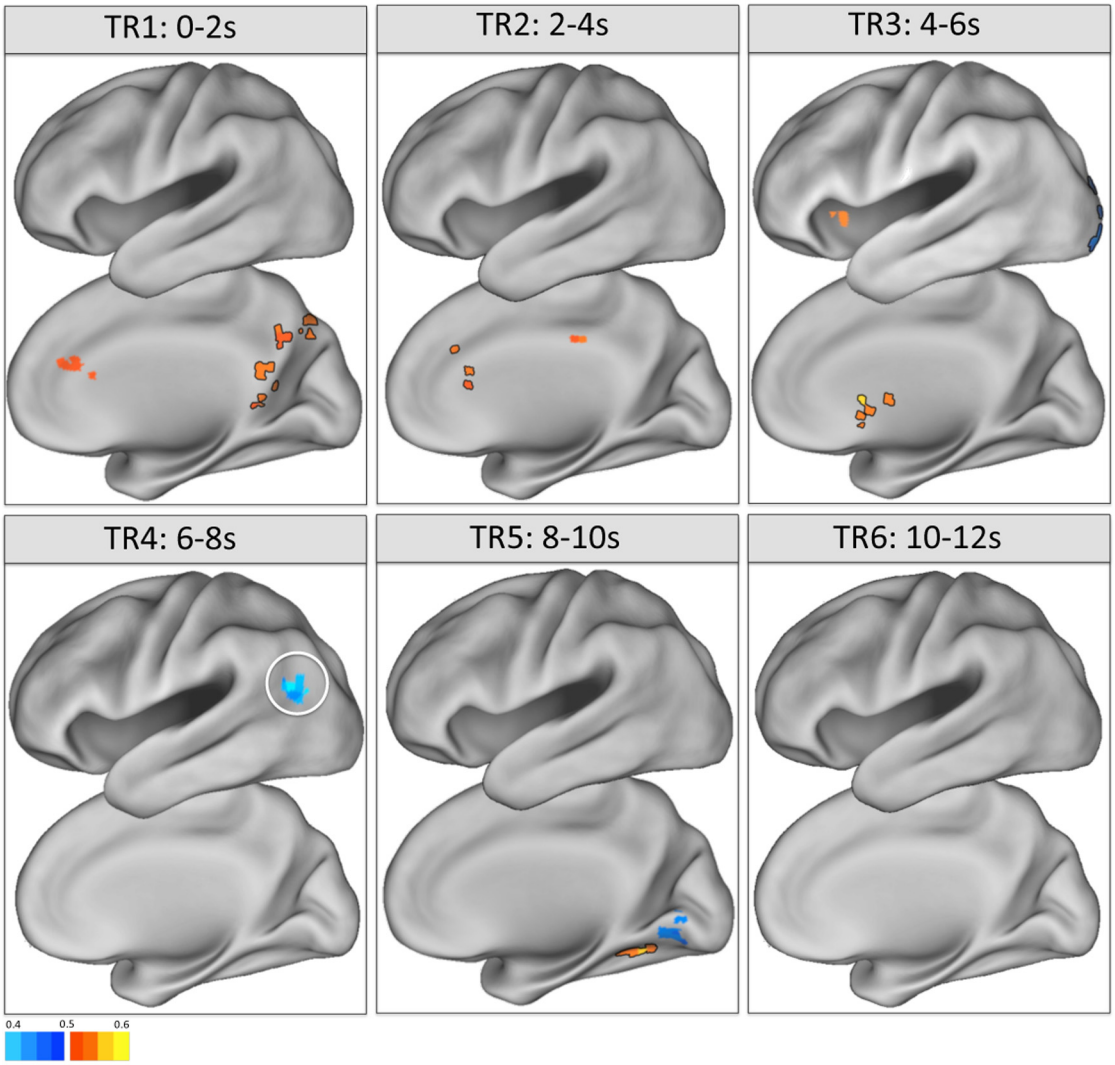

Figure 6. Searchlight analysis of memory decoding during use of countermeasures. Maps of group-average searchlight accuracy indicated voxels that exhibited above- or below-chance accuracy (warm and cool colors, respectively) on the critical explicit→concealed classification of hits versus CRs. Searchlights were performed separately on each time point, revealing local information that demonstrated above-chance decoding in medial frontoparietal regions at early time points (0–4 s) when participants were presumably making memory decisions and below-chance decoding at later time points (6–8 s), including the left AnG (circled), when participants were presumably attempting to conceal memory state. Maps are thresholded at $p < 0.05$ (corrected for multiple comparisons, FWE). Clusters outlined in black additionally survive Bonferroni's correction for the six temporal comparisons. Color map represents AUC values (with 0.5 representing chance).

knowledge (Bergström et al., 2013). The aim of the present study was to determine whether memory signals can be rendered undetectable even after initially attending to them, which may be inevitable in real-world scenarios that use highly salient retrieval cues, such as depictions of a crime scene. The searchlights showing below-chance performance extend the findings of Bergström et al. by showing that, when memory is attended, memory detection can still be obscured at later time points through goal-directed modulation of regions related to memory retrieval and perception (see AnG and occipital clusters in Fig. 6).

The full pattern of results observed in the searchlight analysis provides information at a fine-grained spatial scale that can aid the interpretation of the observed performance in our previous (whole-brain) attempt to decode memory during countermeasures. The above-chance searchlight performance reveals that it is possible to decode memory across tasks given an appropriate choice of voxels and time points. It is important to note that our whole-brain classification analyses that averaged across late time points included voxels that contributed to above-, below-, and at-chance performance on searchlight analyses. Thus, the whole-brain, late-TR classifier was trained on a combination of regions that, individually, exhibited reliably similar, reliably reversed, or unreliable patterns across tasks. Together, this pattern of findings may explain why the whole-brain, late-TR classifier performed no differently from chance.

An alternative interpretation of the chance performance on the explicit → concealed whole-brain, late-TR classifier is that the model was overfit to cognitive processes or stochastic noise that varied over time or that differed according to task. For instance, a cognitive process that likely differed by task was the retrieval orientation adopted by participants during explicit versus concealed tasks. To the extent that retrieval orientation was driving performance on the whole-brain, late-TR classifier, differing retrieval orientations would give rise to very low across-task performance. However, by using more fine-grained searchlight analyses, we were able to show that regions thought to represent, bind, or subjectively process recollected information (particularly the left AnG) were able to be modulated dynamically by attempts to conceal memory on a trial-by-trial basis. Thus, it is unlikely that differing retrieval orientations drove searchlight

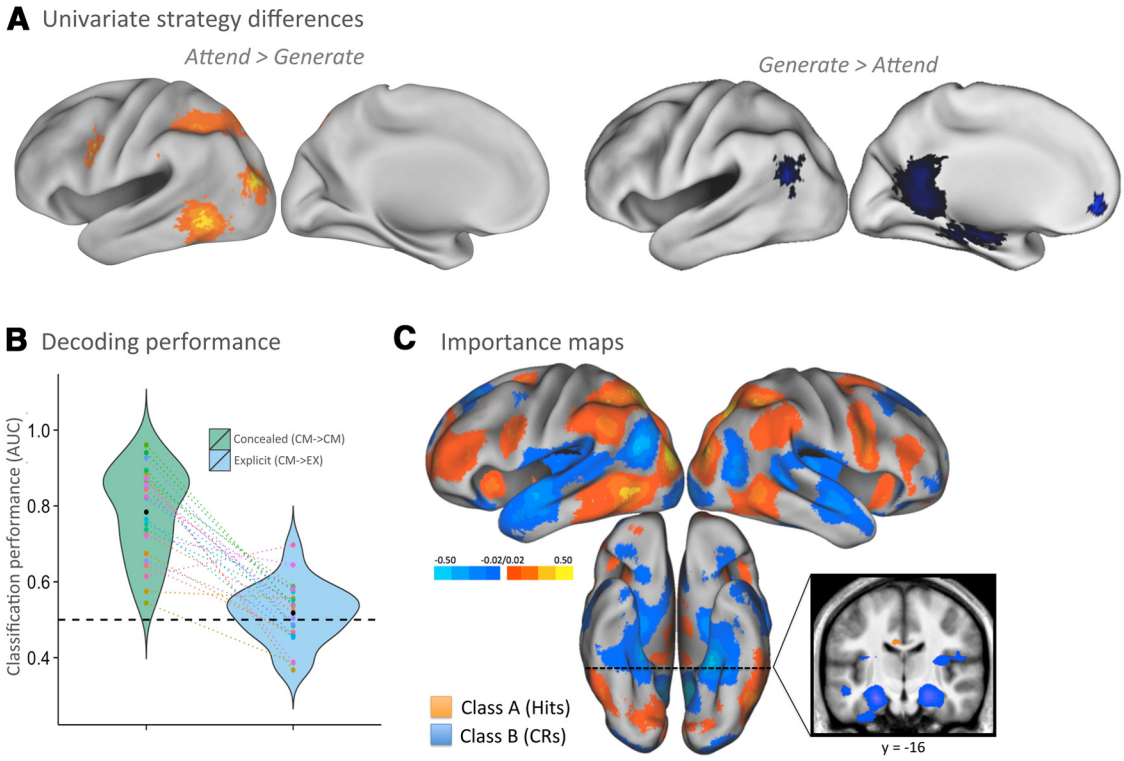

Figure 7. Strategic goal states associated with countermeasures. ***A***, Left, Univariate contrast reveals brain regions engaged during attempts to conceal memory by shifting attention to novel perceptual aspects of the face stimuli. Right, Complementary contrast reveals brain regions involved in concealing novelty, by generating memories in response to novel faces. ***B***, Plots of memory decoding for classifiers trained on patterns from concealed memory trials. Green plots reveal above-chance decoding when tested on concealed trials (CM⟶CM, green) but reduced to chance levels when tested on explicit recognition data (EX⟶CM, blue). AUC values are plotted for each participant's data using unique color identifiers, with lines connecting each participant's classification performance and plot-width depicting participant density in that performance range. ***C***, Maps indicating the diagnostic value that each voxel provides the concealed memory-trained classifier (importance maps) in biasing a hit choice (warm colors) or a CR choice (cool colors). CM⟶CM, Classifier trained on concealed trials, tested on held-out concealed trials (4-fold cross-validation); CM⟶EX, classifier trained on concealed trials, tested on explicit trials.

performance, although they may have played a role in the at-chance whole-brain analyses.

Several other lines of research further support the conclusion that the observed chance performance resulted from the successful deployment of countermeasures and not simply a failure of the classifier to generalize to slightly different task sets. First, we have observed previously that comparable classifiers—also trained to discriminate hits from CRs on laboratory-based face stimuli—can reliably decode hits from CRs on photo sequences from cameras worn by participants (and vice versa), achieving well above chance classification performance (Rissman et al., 2011). Notably, the training and testing data differed along multiple domains, including participants (independent samples), retrieval cues (photo sequences vs single faces), memory content (memory for rich, real-world autobiographical episodes vs laboratory-based memory for single faces), retrieval responses (eight-option responses made with two hands vs five-option responses made with one hand), and study-test interval (1–4 weeks vs ~1 h). Thus, very different training and testing data can give rise to robust memory decoding using our approach. Second, in our previous study, we found robust generalization across participants, in that training on 15 participants' data produced reliable classification of the left-out participant's patterns, with accuracies similar to within-participant decoding (Rissman et al., 2010; note that, in the present dataset, we replicated this robust across-participant classification; data are available on request). Together with the above-chance searchlight classifications, these observations suggest that there is sufficient consistency in hits versus CRs memory patterns across tasks and participants to allow for generalization, again suggesting that the poor explicit → concealed classification should be at least partially attributed to the successful deployment of countermeasures, not stochastic or cognitive sources of noise.

Strategic goal states associated with countermeasures

To gain insights about the strategic goals states that modulated memory-related patterns, we next sought to elucidate the neurocognitive mechanisms associated with the two countermeasure tasks (one for old items, another for new items). To do so, we first investigated the mean (univariate) BOLD signal engaged by each countermeasure task and then explored how well a classifier could discriminate between the tasks.

When participants identified a face as studied on concealed runs, they were to attempt to conceal their memory state by shifting attention to novel aspects of the photograph. Therefore, we predicted that a univariate comparison of trials on which they performed this "attend" task (hits and FAs) would elicit greater mean signal in regions implicated in top-down visual attention relative to trials in which they performed the "generate" task (misses and CRs). Because this comparison amounts to a contrast of subjective oldness versus newness (as all conditions were equally weighted), an additional question is whether this comparison would also reveal neural correlates of the subjective experience of memory or whether engagement of strategic countermeasures would override the subjective memory signal, rendering it undetectable with univariate measures. Interestingly,

this contrast (illustrated in Fig. 7*A*, left) revealed robust activity in the left medial IPS/SPL, an area reliably engaged during goal-directed shifts of visual attention (Corbetta et al., 2008), and the lateral occipital complex, an area implicated in object-based processing (Grill-Spector et al., 2001), but no medial temporal activity. Interestingly, this contrast also revealed a cluster in the left inferior frontal junction (IFJ), a region implicated in task-switching related cognitive control demands (Derrfuss et al., 2005), as well as retrieval suppression strategies that require "thought substitution" (i.e., attention to thoughts other than the cued memory; Benoit and Anderson, 2012). Together, these patterns lend support to the idea that participants were able to strategically engage cognitive control mechanisms to disengage from their subjective experience of memory to attend to novel features of the stimulus.

When participants identified a face as novel in the concealed task, they were to conceal their cognitive state by generating memories in response to the novel face. Therefore, we predicted that trials for which participants attempted to generate memories would engage regions implicated in mnemonic retrieval. As shown in Figure 7*A* (right), this contrast (generate > attend) revealed regions associated consistently with episodic remembering: hippocampus and parahippocampal gyrus, AnG, and retrosplenial gyrus (McDermott et al., 2009; Spaniol et al., 2009; Kim, 2010; Hutchinson et al., 2014), lending additional support for the idea that participants were able to generate other memories in response to novel faces. Again, it is notable that neural correlates of novelty were not observed, suggesting that strategic control measures were able to override initial cognitive states, at least at the level of mean signal. This finding further highlights the importance of using multivariate analyses to attempt to decode transient, trial-by-trial signals that may be undetectable in univariate contrasts.

Given that the subjective experience of oldness and novelty cued highly distinct strategic goal states (as evidenced by the engagement of distinct univariate patterns), we predicted that MVPA-based classification of hits versus CRs in the concealed task would be more robust than that observed in the explicit task. That is, the foregoing explicit → explicit multivariate analyses demonstrated that hits versus CRs in the explicit task were discriminable because of their separable mnemonic experiences, and we reasoned that these conditions would be even more discriminable in the concealed task given that each trial type additionally cued a qualitatively different goal-directed task (attend vs generate, respectively). In line with this prediction, a classifier trained to discriminate hits from CRs in the concealed task (i.e., concealed → concealed) performed well above chance (AUC of 0.78 vs null AUC of 0.49; $t_{(23)} = 12.74$, $p = 6.5 \times 10^{-12}$; Fig. 7*B*, green) and performed better than the analogous classifier from the explicit task (i.e., explicit → explicit, mean AUC of 0.67; $t_{(23)} = 4.42$, $p = 0.0002$). As can be seen in Figure 7*C*, the regions providing the most diagnostic information in favor of hit predictions were those associated with the attend task (i.e., univariate attend > generate; Fig. 7*A*, left); likewise, the regions providing information in favor of CR predictions were primarily those associated with the generate task (i.e., univariate generate > attend; Fig. 7*A*, right). Thus, the classifier appears to be using task-related, and perhaps also memory-related, signals to discriminate hits from CRs in the concealed task.

Finally, we sought to determine whether the cognitive operations elicited during feigned memory and novelty were similar enough to those during veridical memory and novelty as to detect across-task memory and novelty patterns. Although the overall poor performance on the explicit → concealed classifier suggests that veridical memory and novelty patterns could not be used to identify feigned memory and novelty operations, it does not necessitate the reverse to be true. However, the concealed → explicit classifier performance was at chance (AUC of 0.52 vs null AUC of 0.50; $t_{(23)} = 1.17$, $p = 0.25$; Fig. 7*B*, blue), additionally bolstering the idea that the operations engaged to feign memory and novelty are primarily distinct from those elicited during veridical memory and novelty.

In summary, it is clear that strategic goal states were engaged to allow memory status to be concealed at the level of mean activity, as well as on a trial-by-trial basis. To conceal memory, participants were able to use frontoparietal cognitive control mechanisms to reorient attention from a mnemonic representation and toward novel perceptual features represented in the VTC. To conceal novelty, participants were able to use face cues to generate medial temporal lobe (MTL)-mediated mnemonic representations and hold those representations online for the duration of the trial (likely supported by AnG mechanisms).

Discussion

Recent advances in neuroimaging methods show promise in aiding the detection of individual memories. The present study tested whether an fMRI-based memory detection technique would be able to decode memory even when participants attempt to conceal their memory states, while appearing cooperative. We report several key findings. First, univariate analyses demonstrated that countermeasures modulated neural activity, such that memory success effects in a standard retrieval task reversed when participants deployed countermeasures. Notably, participants exhibiting stronger memories had more difficulty reversing the univariate memory success effects. Second, MVPA classifiers reliably decoded individual memories when participants truthfully reported their memory. Third, this ability to decode memory mostly failed, and even slightly reversed, when participants used simple cognitive strategies to thwart classifiers trained on truthful memory responses. Finally, strategic goal states were engaged to allow memory status to be concealed in the mean signal, as well as on a trial-by-trial basis. Frontoparietal control mechanisms enabled memory to be concealed by reorienting attention away from cued memories and toward novel perceptual features, whereas MTL–AnG mechanisms allowed novelty to be concealed by generating and holding online previous memories. Together, these findings demonstrate the power that strategic goal states exert over brain patterns, because they can mostly mask individual memories from detection.

The first clue revealing the profound influence that countermeasures exert over mnemonic operations was revealed in the univariate findings. Here we observed memory success effects (hits > CRs) when participants truthfully reported their memory states, which reversed (CRs > hits) when participants generated memories in response to novel faces or focused on novel aspects of studied faces. Notably, these reversals of activity occurred in regions implicated consistently in memory retrieval, specifically the bilateral hippocampus and left AnG (for meta-analyses, see McDermott et al., 2009; Spaniol et al., 2009; Kim, 2010; Hutchinson et al., 2014). To the extent that hippocampal and AnG BOLD signal reflects retrieved mnemonic content, the finding that the effects reversed between tasks suggests that participants were indeed able to willfully engage the instructed countermeasure strategies across trials. Indeed, univariate comparisons of the two countermeasure tasks revealed the regions that likely mediated this strategic modulation of memory-related activity (Fig. 7*A*):

(1) for memory concealment—requiring reorienting attention from mnemonic information and toward novel perceptual aspects of the stimuli—regions associated with goal-directed allocation of attention and perceptual representations were engaged; and (2) for novelty concealment—requiring generation of memories in response to novel faces—retrieval-related regions were engaged.

It is notable that the memory concealment strategy used here is similar to the "thought substitution" strategy of Benoit and Anderson (2012), whereby retrieval suppression is induced when participants attend to event features other than cued memories. At the neural level, our IFJ cluster overlapped their mid-VLPFC region [surviving small-volume correction, $p_{FWE} = 0.011$, 6 mm sphere centered on their mid-VLPFC cluster; Benoit and Anderson, 2012, their Table S3 (−51, 8, 31)]. In their paradigm, the thought substitution strategy did not modulate hippocampal activity, but here it appeared to (in analogous conditions: explicit vs concealed hits). This discrepancy may be attributable to differences in the substitution strategies between studies: here, the substitution consisted of perceptual analysis of the stimuli, which may drive hippocampal activity less than the substitution strategy of Benoit and Anderson, which required explicit retrieval of an alternate learned association. This highlights an opportunity for additional investigations to determine when and how strategic goal states modulate hippocampal activity.

Participants' ability to reverse univariate retrieval success effects appeared contingent on overall memory strength, suggesting that participants with stronger memories found it more difficult (1) to direct attention away from (strong) memories or (2) to generate memories cued by novel faces. Interestingly, this relationship between memory strength and memory signal in the concealed memory condition was only found in the AnG and not in the hippocampus. This regional dissociation bolsters the proposal that AnG activity is not simply a reflection of hippocampal output, but rather the two regions subserve different functional roles (Vilberg and Rugg, 2012). In such proposals, the hippocampus reinstates neural activity elicited during the initial experience, whereas the AnG may contribute to the online representation of recollected information in service of a memory decision (see also Shimamura, 2011). Beyond providing support for differing roles of the AnG and hippocampus, this regional dissociation may address preliminarily why participants with stronger memories were less likely to show reversed univariate retrieval success effects. To the extent that AnG activity reflects the active maintenance of a memory representation, the finding that AnG, and not hippocampal, activity varies with memory strength provides preliminary support for the idea that participants whose memories were maintained in the AnG with greater fidelity found it more difficult to direct attention away from strong memories. Memory strength also influenced classification performance, in that participants' ability to mask their memory signal from the classifier was contingent on memory strength (albeit only transiently, on TR3 only). As such, our data suggest that cognitive countermeasures may be more difficult to deploy or may be less effective when stronger memories are at play, which has clear implications for memory decoding applications and neurobiological theories of retrieval.

When participants truthfully reported memory, MVPA classifiers reliably decoded the memory state each participant possessed for each test face. Regions that provided diagnostic signal to this classifier were similar to those reported in our previous study (Rissman et al., 2010), with PFC and posterior cingulate cortex biasing hit classifications and the anterior hippocampus biasing CR classifications. Unlike the present study, which required simple "old" or "new" responses (to match responses on the concealed memory task), Rissman et al. required explicit decisions about whether any detail of the previous encounter with the face was recollected or, in the absence of recollected detail, whether the face was retrieved with high or low confidence. Thus, the similarity in importance maps across experiments preliminarily suggests that classifiers rely on similar neural signals to discriminate correctly identified old from new faces when memory judgments are made, regardless of the precise nature of the memory judgment. However, it is important to note that decoding accuracy dramatically diminishes to chance when memory judgments are not made (Rissman et al., 2010, their Implicit task in Experiment 2) or, as revealed by the present findings, when participants attend to their memory states but then deploy countermeasures to conceal them.

This latter finding has implications for neurobiological theories of memory and recommendations regarding readiness for field use. Specifically, the finding that mnemonic information could be decoded even when participants are not instructed to attend to that information (Kuhl et al., 2013) raises the question as to whether memory signals can be rendered undetectable even after initially attending to them, as may be the case in real-world scenarios with highly salient retrieval cues, such as a crime scene. We found this to mostly be the case, thus imposing a significant boundary condition on the validity of fMRI-based memory decoders and suggesting that the method is not yet ready for field use.

The finding that cognitive countermeasures can thwart or even reverse the ability of a classifier to accurately detect individual memories even when memory is attended raises the question of whether transient memory signals may be detectable in the face of countermeasures. Here, in the concealed memory data, we were able to weakly decode hits from CRs at only one time point (Figs. 4*B*, 5*A*), and hits were reliably confused with CRs at a later time point (Fig. 5*C*). In particular, the finding that classification performance at this time point varied according to memory strength warrants additional investigation. This finding preliminarily suggests that, for example, individuals with strong memories of relevant information may have difficulty concealing their memory. The whole-brain classifications were complemented by searchlights showing that early above-chance classification was possible using data from medial frontal and parietal cortical regions and that left AnG patterns may reliably confuse memory classifiers (resulting in significantly below-chance decoding) at later time points when participants were presumably deploying countermeasures. Future studies using more temporally resolved data (electroencephalography, EEG, or simultaneous fMRI–EEG) may further inform whether transient memory signals can be detected in the presence of cognitive countermeasures. Such investigations may reveal that attention to memory representations necessarily gives rise to detectable memory signals, although we note that extant EEG-based memory detection methods also appear vulnerable to countermeasures (Rosenfeld et al., 2004, 2013).

Together, our data demonstrate that cognitive strategies can dramatically compromise the ability of fMRI-based classifiers to detect memory. These findings extend those of a previous fMRI investigation using physical countermeasures to compromise fMRI-based "lie detection" (Ganis et al., 2011) to the domain of memory and further demonstrate that covert measures, which may not be physically detectable, can nevertheless foil fMRI-based memory detection. Furthermore, given previous findings

that memory was not decodable when probed implicitly (Rissman et al., 2010), our findings further reveal that, even when memories are attended to, neural signatures of memory states can still be rendered difficult to detect. It is also important to note that concealed information, mnemonic or otherwise, can be detected using a variety of different methods (outlined in Introduction), including those that have been designed to be relatively countermeasure-resistant (such as the complex trial protocol of the Concealed Information Test; for review, see Rosenfeld et al., 2013). Future studies are needed to understand the effects of ecologically valid factors, such as the passage of time, rehearsal, and the richer mnemonic information held in real-world (rather than laboratory-presented, list-based) memories, as well as the influence of false memory on memory detection techniques. Although there may exist great promise for these techniques, it is clear that there are important boundary conditions that may render their real-world potential uncertain.

References

Benoit RG, Anderson MC (2012) Opposing mechanisms support the voluntary forgetting of unwanted memories. Neuron 76:450–460. CrossRef Medline

Bergström ZM, Anderson MC, Buda M, Simons JS, Richardson-Klavehn A (2013) Intentional retrieval suppression can conceal guilty knowledge in ERP memory detection tests. Biol Psychol 94:1–11. CrossRef Medline

Bishop CM (2006) Pattern recognition and machine learning. New York: Springer.

Brown T, Murphy E (2010) Through a scanner darkly: functional neuroimaging as evidence of a criminal defendant's past mental states. Stanford Law Rev 62:1119–1208. Medline

Chadwick MJ, Hassabis D, Weiskopf N, Maguire EA (2010) Decoding individual episodic memory traces in the human hippocampus. Curr Biol 20:544–547. CrossRef Medline

Corbetta M, Patel G, Shulman GL (2008) The reorienting system of the human brain: from environment to theory of mind. Neuron 58:306–324. CrossRef Medline

Derrfuss J, Brass M, Neumann J, von Cramon DY (2005) Involvement of the inferior frontal junction in cognitive control: meta-analyses of switching and Stroop studies. Hum Brain Mapp 25:22–34. CrossRef Medline

Detre GJ, Polyn SM, Moore CD, Natu VS, Singer BD, Cohen JD, Haxby JV, Norman KA (2006) The multi-voxel pattern analysis (MVPA) toolbox. 12th Annual Meeting of the Organization of Human Brain Mapping, Florence, Italy, January.

Farah MJ, Hutchinson JB, Phelps EA, Wagner AD (2014) Functional MRI-based lie detection: scientific and societal challenges. Nat Rev Neurosci 15:123–131. CrossRef Medline

Fawcett T (2004) ROC graphs: notes and practical considerations for researchers. Machine Learn 31:1–38.

Foster BL, Rangarajan V, Shirer WR, Parvizi J (2015) Intrinsic and task-dependent coupling of neuronal population activity in human parietal cortex. Neuron 86:578–590. CrossRef Medline

Ganis G, Rosenfeld JP, Meixner J, Kievit RA, Schendan HE (2011) Lying in the scanner: covert countermeasures disrupt deception detection by functional magnetic resonance imaging. Neuroimage 55:312–319. CrossRef Medline

Greely HT (2011) Reading minds with neuroscience–possibilities for the law. Cortex 47:1254–1255. CrossRef Medline

Green DM, Swets JA (1966) Signal detection theory and psychophysics. New York: Wiley.

Grill-Spector K, Kourtzi Z, Kanwisher N (2001) The lateral occipital complex and its role in object recognition. Vision Res 41:1409–1422. CrossRef Medline

Hutchinson JB, Uncapher MR, Weiner KS, Bressler DW, Silver MA, Preston AR, Wagner AD (2014) Functional heterogeneity in posterior parietal cortex across attention and episodic memory retrieval. Cereb Cortex 24:49–66. CrossRef Medline

Johnson JD, McDuff SGR, Rugg MD, Norman KA (2009) Recollection, familiarity, and cortical reinstatement: a multivoxel pattern analysis. Neuron 63:697–708. CrossRef Medline

Kim H (2010) Dissociating the roles of the default-mode, dorsal, and ventral networks in episodic memory retrieval. Neuroimage 50:1648–1657. CrossRef Medline

Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863–3868. CrossRef Medline

Kuhl BA, Chun MM (2014) Successful remembering elicits event-specific activity patterns in lateral parietal cortex. J Neurosci 34:8051–8060. CrossRef Medline

Kuhl BA, Johnson MK, Chun MM (2013) Dissociable neural mechanisms for goal-directed versus incidental memory reactivation. J Neurosci 33:16099–16109. CrossRef Medline

McDermott KB, Szpunar KK, Christ SE (2009) Laboratory-based and autobiographical retrieval tasks differ substantially in their neural substrates. Neuropsychologia 47:2290–2298. CrossRef Medline

McDuff SGR, Frankel HC, Norman KA (2009) Multivoxel pattern analysis reveals increased memory targeting and reduced use of retrieved details during single-agenda source monitoring. J Neurosci 29:508–516. CrossRef Medline

Meegan DV (2008) Neuroimaging techniques for memory detection: scientific, ethical, and legal issues. Am J Bioeth 8:9–20. CrossRef Medline

Polyn SM, Kragel JE, Morton NW, McCluey JD, Cohen ZD (2012) The neural dynamics of task context in free recall. Neuropsychologia 50:447–457. CrossRef Medline

Poppenk J, Norman KA (2012) Mechanisms supporting superior source memory for familiar items: a multi-voxel pattern analysis study. Neuropsychologia 50:3015–3026. CrossRef Medline

Quamme JR, Weiss DJ, Norman KA (2010) Listening for recollection: a multi-voxel pattern analysis of recognition memory retrieval strategies. Front Hum Neurosci 4. CrossRef

Ratcliff R (1978) Theory of memory retrieval. Psychol Rev 85:59–108. CrossRef

Rissman J, Wagner AD (2012) Distributed representations in memory: insights from functional brain imaging. Annu Rev Psychol 63:101–128. CrossRef Medline

Rissman J, Greely HT, Wagner AD (2010) Detecting individual memories through the neural decoding of memory states and past experience. Proc Natl Acad Sci U S A 107:9849–9854. CrossRef Medline

Rissman J, Chow T, Hardekopf K, Greely HT, Wagner AD (2011) Decoding real-world autobiographical retrieval experiences with fMRI multi-voxel pattern analysis. Soc Neurosci Abstr 37:327.06.

Rosenfeld JP, Soskins M, Bosh G, Ryan A (2004) Simple, effective countermeasures to P300-based tests of detection of concealed information. Psychophysiology 41:205–219. CrossRef Medline

Rosenfeld JP, Hu X, Labkovsky E, Meixner J, Winograd MR (2013) Review of recent studies and issues regarding the P300-based complex trial protocol for detection of concealed information. Int J Psychophysiol 90:118–134. CrossRef Medline

Shimamura AP (2011) Episodic retrieval and the cortical binding of relational activity. Cogn Affect Behav Neurosci 11:277–291. CrossRef Medline

Snodgrass JG, Corwin J (1988) Pragmatics of measuring recognition memory: applications to dementia and amnesia. J Exp Psychol Gen 117:34–50. CrossRef Medline

Spaniol J, Davidson PSR, Kim ASN, Han H, Moscovitch M, Grady CL (2009) Event-related fMRI studies of episodic encoding and retrieval: meta-analyses using activation likelihood estimation. Neuropsychologia 47:1765–1779. CrossRef Medline

Verschuere B, Ben-Shakhar G, Meijer E (2011) Memory detection. Cambridge, UK: Cambridge UP.

Vilberg KL, Rugg MD (2012) The neural correlates of recollection: transient versus sustained fMRI effects. J Neurosci 32:15679–15687. CrossRef Medline

Weiner KS, Grill-Spector K (2010) Sparsely-distributed organization of face and limb activations in human VTC. Neuroimage 52:1559–1573. CrossRef Medline

Xiong J, Gao JH, Lancaster JL, Fox PT (1995) Clustered pixel analysis for functional MRI activation studies of the human brain. Hum Brain Mapp 3:287–301.