Detecting individual memories through the neural decoding of memory states and past experience

Jesse Rissman[a,1], Henry T. Greely[b], and Anthony D. Wagner[a,c,1]

[a]Department of Psychology, [b]Law School, and [c]Neurosciences Program, Stanford University, Stanford, CA 94305

Edited by Edward E. Smith, Columbia University, New York, NY, and approved April 12, 2010 (received for review January 26, 2010)

A wealth of neuroscientific evidence indicates that our brains respond differently to previously encountered than to novel stimuli. There has been an upswell of interest in the prospect that functional MRI (fMRI), when coupled with multivariate data analysis techniques, might allow the presence or absence of individual memories to be detected from brain activity patterns. This could have profound implications for forensic investigations and legal proceedings, and thus the merits and limitations of such an approach are in critical need of empirical evaluation. We conducted two experiments to investigate whether neural signatures of recognition memory can be reliably decoded from fMRI data. In Exp. 1, participants were scanned while making explicit recognition judgments for studied and novel faces. Multivoxel pattern analysis (MVPA) revealed a robust ability to classify whether a given face was subjectively experienced as old or new, as well as whether recognition was accompanied by recollection, strong familiarity, or weak familiarity. Moreover, a participant's subjective mnemonic experiences could be reliably decoded even when the classifier was trained on the brain data from other individuals. In contrast, the ability to classify a face's objective old/new status, when holding subjective status constant, was severely limited. This important boundary condition was further evidenced in Exp. 2, which demonstrated that mnemonic decoding is poor when memory is indirectly (implicitly) probed. Thus, although subjective memory states can be decoded quite accurately under controlled experimental conditions, fMRI has uncertain utility for objectively detecting an individual's past experiences.

declarative memory | episodic retrieval | experiential knowledge | memory detection | pattern classification | functional MRI

Our brains are remarkable in their ability to encode and store an ongoing record of our experiences. The prospect of using advanced brain imaging technologies to identify a neural marker that reliably indicates whether or not an individual has previously encountered a particular person, place, or thing has generated much interest in both neuroscientific and legal communities (1, 2). A memory detection technique could conceivably be used to interrogate the brains of suspected criminals or witnesses for neural evidence that they recognize certain individuals or entities, such as those from a crime scene. Indeed, data from one electroencephalographic (EEG) procedure [Brain Electrical Oscillation Signature (BEOS) Profiling] was recently admitted in a murder trial in India to establish evidence that the suspect's brain contained knowledge that only the true perpetrator could possess (3). Results from another EEG-based technique, which relies on the P300 response to infer that an individual "recognizes" a probe stimulus, were admitted into evidence in a U.S. court case in 2001 (4). Given these precedents, coupled with the rapid strides being made in cognitive neuroscience research, other parties will almost certainly eventually seek to exploit brain-recording data as evidence of a person's past experiences, in judicial proceedings or in civil, criminal, military, or intelligence investigations. The scientific validity of such methods must be rigorously and critically evaluated (5–12).

Although there are no peer-reviewed empirical papers describing the BEOS Profiling method (to our knowledge), this approach appears to follow in the tradition of prior EEG methods for detecting the presence or absence of memory traces (13–15). Because EEG-based techniques have been argued to suffer several major limitations (*SI Discussion*), recent interest has focused on applying fMRI as a means to probe experiential knowledge (1). The greater spatial resolution of fMRI data may allow researchers to better detect and more precisely characterize the distributed pattern of brain activity evoked by a particular stimulus or cognitive state. Using multivoxel pattern analysis (MVPA) methods (16, 17) that can be applied to index memory-related neural responses (18–22), we capitalized on the rich information contained within distributed fMRI activity patterns to attempt to decode the mnemonic status of individual stimuli.

A substantial body of neuroscientific evidence demonstrates that an individual's brain responds differently when it experiences a novel stimulus as compared with a stimulus that has been previously encountered (23–26). For example, prior fMRI data submitted to univariate analysis have documented regions of prefrontal cortex (PFC), posterior parietal cortex (PPC), and medial temporal lobe (MTL) wherein activation tracks the degree to which a stimulus gives rise to the subjective mnemonic perception that it was previously experienced (i.e., perceived oldness), independent of the stimulus's true mnemonic history (27–30). Other fMRI studies have identified regions of the MTL and posterior sensory cortices wherein activity appears to track the objective mnemonic history of stimuli, independent of an individual's subjective mnemonic experience (30–35). Neural correlates of past stimulus experience have also been revealed in fMRI and EEG studies of priming, a form of nondeclarative memory in which a previously encountered stimulus is processed more fluently upon subsequent presentation in an indirect (implicit) memory test (23, 36–38). Although these rich literatures suggest that fMRI memory detection may be possible, it is presently unknown whether the subjective and objective neural signatures of old/new recognition can be reliably detected on individual test trials. Moreover, to the extent that memory detection is possible, the across-subject consistency of the neural evidence affording such classification is unknown.

In two experiments, we assessed whether distinct mnemonic categories—subjective memory states and objective old/new status—can be classified from single-trial fMRI data using an MVPA approach. In both, participants were exposed to a large set of faces and then were scanned ≈1 h later while viewing the studied faces as well as novel faces. Exp. 1 examined classification of subjective and objective memory while individuals were engaged in a task that required explicit recognition decisions regarding the test stimuli. Exp. 2 was virtually identical, with the key differences being that (*i*) mnemonic encoding was incidental, rather than intentional, and (*ii*) during the first half of the scanning session,

Author contributions: J.R., H.T.G., and A.D.W. designed research; J.R. performed research; J.R. contributed new reagents/analytic tools; J.R. analyzed data; and J.R., H.T.G., and A.D.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. E-mail: jesse.rissman@stanford.edu or awagner@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1001028107/-/DCSupplemental.

participants made male/female judgments about old and new faces (rather than explicit memory judgments), whereas during the second half of scanning, participants made explicit recognition decisions. Thus, Exp. 2 assessed classification under circumstances in which old/new recognition was indirectly probed and examined whether the neural signatures that characterize explicit recognition are also diagnostic of indirect (implicit) recognition.

Results

Exp. 1: Explicit Recognition Task. ***Behavioral performance.*** Sixteen participants were scanned while making explicit memory judgments on 400 probe faces. For each, participants indicated their mnemonic experience using one of five responses: recollected as studied ("R old"), high confidence studied ("HC old"), low confidence studied ("LC old"), low confidence unstudied ("LC new"), or high confidence unstudied ("HC new") (39). Mean recognition accuracy was 0.71 [(hit rate (0.70) + correct rejection (CR) rate (0.71))/2]; mean d' (1.15) differed from chance [$t_{(15)} = 7.42$, $p < 10^{-5}$]. The distribution of responses to objectively old (OLD) and objectively new (NEW) faces confirmed that participants used the response options appropriately, rarely responding "R old" or "HC old" to NEW faces or "HC new" to OLD faces (Table S1, Exp. 1). Reaction times (RTs) followed an inverted U-shaped function, with the fastest RTs occurring for responses at the endpoints of the recognition scale (i.e., "R old" and "HC new") and the slowest RTs for LC responses. Despite increased study-test lag, mnemonic interference, and potential fatigue, performance was relatively stable (mean d' in the first (1.23) and second (1.09) half of the session did not significantly differ [$t_{(15)} = 1.65$, $P = 0.11$].

fMRI analyses. Assessing classifier performance. We used regularized logistic regression to classify the mnemonic status of individual trials based on distributed fMRI activation patterns. Classification performance was indexed by receiver operating characteristic (ROC) curves, which rank the classification outputs according to their probability estimates (from strongly favoring Class A to strongly favoring Class B) and chart the relationship between the classifier's *true positive rate* (probability of correctly labeling examples of Class A as Class A) and *false positive rate* (probability of incorrectly labeling examples of Class B as Class A) across a range of decision boundaries. The area under the curve (AUC) indexes the mean accuracy with which a randomly chosen pair of Class A and Class B trials could be assigned to their correct classes (0.5 = random performance; 1.0 = perfect performance).

Classifying faces as OLD vs. NEW. As a first assessment of the MVPA classifier's ability to decode whether a face was OLD or NEW, we analyzed trials in which the participant correctly labeled the face's objective mnemonic status, training the classifier to discriminate OLD faces that participants called "old" (Hits) from NEW faces called "new" (CRs). In this classification scheme, the objective and subjective old/new status of the faces in each class were identical, and thus the classifier could capitalize on neural

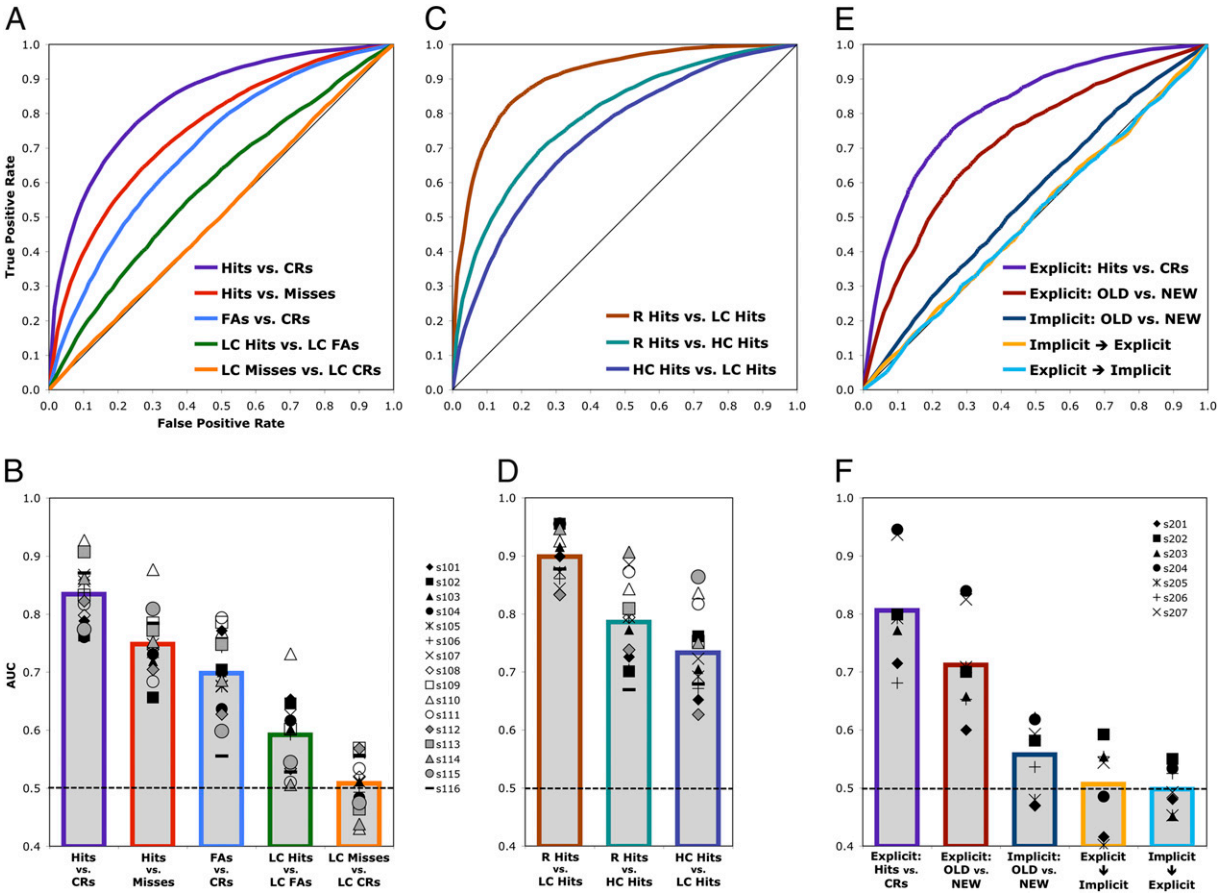

Fig. 1. Mnemonic decoding results. Mean ROC curves (*A*, *C*, and *E*) and their corresponding AUC values (*B*, *D*, and *F*) summarize classifier performance for various classification schemes in Exp. 1 (*A*–*D*) and Exp. 2 (*E* and *F*). AUC values are plotted for each participant's data using unique identifiers, with the group means indicated by the vertical bars. Chance performance (AUC = 0.5) is indicated by the dashed line. For each classification scheme, participants with fewer than 18 trials in each class were excluded from analysis (Table S2). In *E* and *F*, "Implicit → Explicit" refers to a classifier trained to discriminate OLD vs. NEW on the Implicit Recognition Task data and tested on the Explicit Recognition Task data, and "Explicit → Implicit" refers to the converse classification scheme.

signals pertaining to either or both. The results (Fig. 1 *A* and *B*) revealed that the classifier successfully discriminated Hits from CRs, with a mean AUC of 0.83 [*t* test vs. null hypothesis (AUC = 0.50): $t_{(15)} = 27.01$, $P < 10^{-13}$]. Notably, robust classification performance was obtained for all 16 participants, with AUC levels ranging from 0.76 to 0.93. Across-participant variance in Hit/CR classification performance was partially driven by individual differences in recognition memory performance, as evidenced by a significant correlation between classification AUC and behavioral recognition accuracy [$r = 0.55$, $P < 0.05$].

Although AUC provides a more sensitive single metric of classification performance than does overall accuracy (40), mean classification accuracy levels were also computed (Fig. S1*A*). Hits could be discriminated from CRs with a mean accuracy of 76% when the classifier was forced to make a guess on every trial. However, when the classifier's guesses were restricted to only those trials for which it had the strongest "confidence" in its predictions, mean classification accuracy rose to as high as 95% (Fig. S1*A*). Thus, the classification procedure can be calibrated to produce few classification errors when the classifier is made to refrain from guessing on all but those trials where the neural evidence for a particular mnemonic state is most robust.

Classifying subjective mnemonic experience. A variety of classification schemes were used to assess the ability to decode the subjective mnemonic experience associated with individual faces. To isolate the purely subjective components of retrieval, objective mnemonic status was held constant for any given classification. In this manner, a classifier trained to discriminate between studied faces correctly recognized as "old" (Hits) and studied faces incorrectly perceived as "new" (Misses) indicates how well the subjective "old"/"new" status of faces can be decoded when the objective status is always OLD. Likewise, a classifier trained to discriminate between FAs and CRs indicates the ability to decode the subjective "old"/"new" status when the objective status is always NEW. The results (Fig. 1 *A* and *B*) revealed well above chance classification of the subjective mnemonic experience associated with both OLD faces (mean AUC = 0.75) [$t_{(14)} = 18.08$, $P < 10^{-10}$] and NEW faces (AUC = 0.70) [$t_{(15)} = 11.43$, $P < 10^{-8}$]. Mean classification accuracy across classifier "confidence" further revealed that the Hit/Miss classification approached 90% and the FA/CR classification approached 80% at the highest "confidence" level (Fig. S1*A*). These effects remained robust when only LC responses were considered, indicating that the classifier can decode neural signatures of subjective oldness even when the participant's decision confidence is held constant (*SI Results*).

Classifying distinct manifestations of subjective recognition. We next assessed the accuracy with which classifiers could decode the specific type of subjective recognition experienced by participants. First, we trained a classifier to discriminate Hits on which participants reported the experience of contextual recollection (R Hits) from Hits on which participants reported low confidence in their recognition judgments (LC Hits). Differentiating between these two subjective memory states proved to be an easy task for the classifier (Fig. 1 *C* and *D*), with a mean AUC of 0.90 [$t_{(10)} = 29.75$, $P < 10^{-10}$] and a mean accuracy for the upper classifier "confidence" decile of 97% (Fig. S1*B*). Second, and strikingly, separate classifiers were able to robustly discriminate HC Hits from both R Hits (AUC = 0.79) [$t_{(12)} = 13.56$, $P < 10^{-7}$] and LC Hits (AUC = 0.73) [$t_{(12)} = 11.57$, $P < 10^{-7}$], with the former classification scheme significantly outperforming the later [$t_{(10)} = 3.07$, $P < 0.05$] (Fig. 1 *C* and *D*); mean accuracy at the highest classifier "confidence" level was ≈90% and 84%, respectively (Fig. S1*B*). Thus, classifications of different subjective recognition states from distributed patterns of fMRI data were well above chance when the memory test was explicit, with discrimination between recollection (R Hits) and strong familiarity (HC Hits) being superior to that between strong familiarity and weak familiarity (LC Hits).

Classifying objective mnemonic status. Next, we assessed whether the objective OLD/NEW status of faces can be decoded, holding subjective mnemonic status constant. Because most participants made few "R old" or "HC old" responses for NEW faces and few "HC new" responses to OLD faces (average number of trials: R FAs = 2.3; HC FAs = 11.2; HC Misses = 11.4; see also Table S2), analyses were restricted to trials on which participants made low confidence responses. Importantly, when participants responded "LC old," the classifier demonstrated above-chance discrimination of OLD faces (LC Hits) from NEW faces (LC FAs), with a mean AUC of 0.59 [$t_{(12)} = 5.04$, $P < 10^{-3}$]. However, classification accuracy was markedly, and significantly, lower (Fig. 1 *A* and *B* and Fig. S1*A*) than in the above subjective memory classifications (all $P < 0.05$). Moreover, when participants responded "LC new" (i.e., LC Misses vs. LC CRs), the classifier was at chance in discriminating OLD from NEW faces [mean AUC = 0.51; $t_{(13)} = 0.66$, n.s.]. Thus, while classification of subjective mnemonic states was robust, classification of the objective mnemonic status of a face, holding subjective status constant, was relatively poor.

Neural signals that drive classifier performance. Although the goal of the present investigation was to quantify the discriminability of distinct mnemonic states based on their underlying fMRI activity patterns, it is valuable to examine which brain regions provided diagnostic signals to each classifier. Importance maps for the classifications of subjective mnemonic states are displayed in Fig. 2 {see *SI Methods* for details and Fig. S2 for expanded data reporting; see *SI Results* for additional analyses exploring decoding performance when classification was restricted to individual anatomical regions of interests (Table S3) or focal voxel clusters [i.e., spherical searchlights (Fig. S3 and Fig. S4*A*)]}. The importance maps for the "old"/"new" classifications (Hit/CR, Hit/Miss, and FA/CR) revealed a common set of regions wherein activity increases were associated with the classifier's prediction of an "old" response. Prominent foci included the left lateral PFC (inferior frontal gyrus; white arrows) and bilateral PPC falling along the intraparietal sulcus (IPS) [yellow arrows; for the FA/CR classification, bilateral IPS can be visualized in a more ventral slice (Fig. S2)]. Although few regions exhibited negative importance values, a region of anterior hippocampus, extending into the amygdala, emerged in the Hit/CR and FA/CR maps as showing activity increases that predicted a "new" response.

The importance maps for the two classifications that isolated distinct experiences of subjective recognition revealed several notable findings. In the R Hits vs. HC Hits classification, bilateral hippocampal regions (orange arrows) and left angular gyrus (blue arrow) were associated with the prediction of an R Hit (Fig. 2). The hippocampal regions had a more dorsal and posterior extent than the hippocampal areas described above, and overlapped with a region of left posterior hippocampus that was predictive of an "old" response in the Hit/CR and Hit/Miss classifications (Fig. S2). Critically, these robust hippocampal and angular gyrus effects were substantially diminished in the HC Hits vs. LC Hits importance map. Rather, this classification of item recognition strength appeared to depend more strongly on the dorsal PPC and left lateral PFC regions that were also observed for the subjective "old"/"new" classifications.

Across-participant classification. The above analyses were conducted on classifiers trained and tested on within-participant fMRI data. It is also of interest to know whether memory-related neural signatures are sufficiently consistent across individuals to allow one individual's memory states to be decoded from a classifier trained exclusively on fMRI data from other individuals' brains. Accordingly, we reran the classification analyses, but this time we used a leave-one-participant-out cross-validation approach. Across-participant classification performance levels were similar to those of the corresponding within-participant analyses (Fig. S5; compare with Fig. 1 *A–D*), suggesting high across-

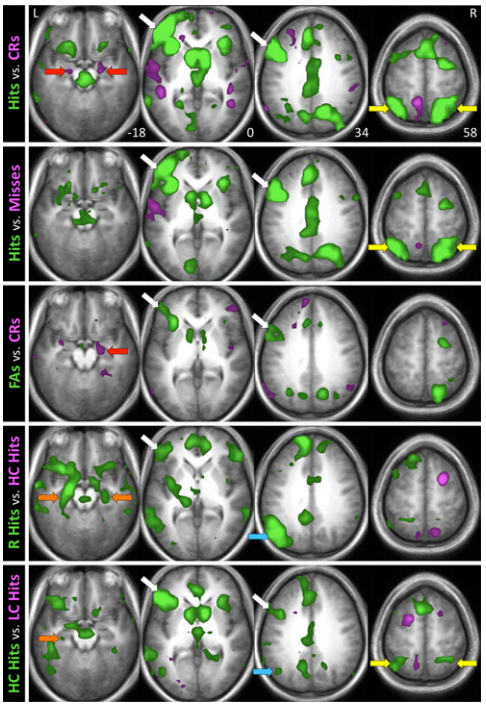

Fig. 2. Classification importance maps. For each classification scheme, group mean importance maps highlight voxels wherein activity increases drive the classifier toward a Class A prediction (green) or Class B prediction (violet). Importance values were arbitrarily thresholded at ±0.0002 and overlaid on selected axial slices of the mean normalized anatomical image (coordinates indicate *z* axis position in Montreal Neurological Institute space). See text for references to colored arrows.

participant consistency in memory-related activation patterns. Indeed, the corresponding within- and across-participant AUCs did not significantly differ (all $P > 0.01$; $p_{crit} = 0.0063$ with Bonferroni correction for 8 comparisons), although performance for the across-participant LC Hit/FA classification no longer exceed chance ($P = 0.1$).

Exp. 2: Implicit vs. Explicit Recognition. A new group of seven participants performed a modified version of Exp. 1, in which prescan mnemonic encoding was incidental and the first five scanning runs required male/female judgments, rather than explicit memory judgments. Because old/new recognition during these runs was not relevant to the male/female decision, memory in these runs was indirectly (implicitly) probed; we refer to this task as the "Implicit Recognition Task". For the remaining five scanning runs, participants performed the same "Explicit Recognition Task" used in Exp. 1.

Behavioral performance. On the Implicit Recognition Task, participants were over 99% accurate at judging the male/female status of the faces. On the Explicit Recognition Task, the distribution of responses to OLD and NEW faces (Table S1, Exp. 2) was comparable to those obtained in Exp. 1. When directly contrasted with the performance levels obtained in the last five runs of Exp. 1 (mean $d' = 1.09$), participants in Exp. 2 exhibited superior recognition performance (mean $d' = 1.71$) [$t_{(21)} = 2.46$, $P < 0.05$], which may be attributable to the deep encoding afforded by the Exp. 2 study task (attractiveness ratings).

fMRI analyses. We first assessed whether MVPA classification performance during Explicit Recognition was comparable across Exps. 1 and 2. A classifier trained to discriminate Hits vs. CRs during the Explicit Recognition Task runs in Exp. 2 achieved a mean AUC of 0.81 (Fig. 1 *E* and *F*). To compare classification rates across experiments, we reran the Hits vs. CRs classification from Exp. 1 using only the last 5 scanning runs; when doing so, the mean AUC in Exp. 1 was 0.77, which was not significantly different from that in Exp. 2 [$t_{(21)} = 0.46$, n.s.].

Having confirmed that mnemonic classification during the Explicit Recognition Task was roughly equivalent across the two experiments, we ran a series of analyses to compare classification performance between the Explicit and Implicit Recognition Tasks of Exp. 2 (Fig. 1 *E* and *F*). Because participants did not make memory judgments during the Implicit Recognition Task, the faces encountered during this task could only be labeled by their objective OLD/NEW status. Thus, we assessed how accurately we could classify the OLD/NEW status of faces during the Implicit Recognition Task, where any effects of memory are indirect, and during the Explicit Recognition Task (for the latter, this entailed classifying OLD vs. NEW faces without taking participants' subjective recognition responses into account; note that subjective and objective mnemonic status are correlated). Importantly, whereas OLD/NEW classification was well above chance using the Explicit Recognition Task data from Exp. 2 (mean AUC = 0.71) [$t_{(6)} = 6.27$, $P < 10^{-3}$], classification performance did not markedly differ from chance using the Implicit Recognition Task data (mean AUC = 0.56) [$t_{(6)} = 2.39$, $P = 0.054$; $p_{crit} = 0.025$ with Bonferroni correction for 2 comparisons] (Fig. 1 *E* and *F*). The task-dependent decline in OLD/NEW classification performance across the explicit and implicit tests was significant [$t_{(6)} = 5.46$, $P < 0.01$]. Classification remained at chance levels when the classifier was trained on trials from the Explicit Recognition Task and tested on trials from the Implicit Recognition Task (mean AUC = 0.50) [$t_{(6)} = 0.13$, n.s.] (Fig. 1 *E* and *F*). The converse classification scheme (i.e., trained on Implicit and tested on Explicit) also yielded chance performance (mean AUC = 0.51) [$t_{(6)} = 0.24$, n.s.]. Taken together, these analyses suggest that our classification methods are not capable of robustly decoding the OLD/NEW status of faces encountered during the Implicit Recognition Task.

Discussion

The present experiments evaluated whether individuals' subjective memory experiences, as well as their veridical experiential history, can be decoded from distributed fMRI activity patterns evoked in response to individual stimuli. MVPA yielded several notable findings that have implications both for our understanding of neural correlates of recognition memory and for possible use of these methods for forensic investigations. First, MVPA classifiers readily differentiated activity patterns associated with faces participants' correctly recognized as old from those associated with faces correctly identified as novel. Second, it was possible to reliably decode which faces participants subjectively perceived to be "old" and which they perceived to be "new," even when holding the objective mnemonic status of the faces constant. Third, MVPA classifiers accurately determined whether participants' recognition experiences were associated with subjective reports of recollection, a strong sense of familiarity, or only weak familiarity, with the discrimination between recollection and strong familiarity being superior to that between strong vs. weak familiarity. Fourth, neural signatures associated with subjective memory states were sufficiently consistent across individuals to allow one participant's mnemonic experiences to be decoded using a classifier trained exclusively on brain data from other participants. Fifth, in contrast to the successful decoding of subjective memory states, the veridical experiential history associated with a face could not be easily classified when subjective recognition was held constant. For faces that participants claimed to recognize, the classifier achieved only limited success at determining which were actually old vs. novel; for faces that participants claimed to be novel, the classifier was unable to determine which had been previously seen. Finally, a neural signature of past experience could not be reliably decoded

during implicit recognition, during which participants viewed previously seen and novel faces outside the context of an explicit recognition task. Taken together, these findings demonstrate the potential power of fMRI to detect neural correlates of subjective remembering of individual events, while underscoring the potential limitations of fMRI for uncovering the veridical experiential record and for detecting individual memories under implicit retrieval conditions.

The robust classification of participants' subjective recognition states indicates that the perceptions of oldness and novelty are associated with highly distinctive neural signatures. Assessment of the importance maps for the Hit/Miss and FA/CR classifications (Fig. 2) revealed a common set of lateral PFC and PPC regions for which increased activity favored an "old" response; a qualitatively similar pattern was apparent in univariate statistical maps (Fig. 3; see also Fig. S3). These frontoparietal regions have been previously shown to track perceived oldness (27, 28, 41, 42) and are likely involved in cognitive and attentional control processes that guide the recovery of information from memory, as well as the evaluative processes that monitor retrieval outcomes and guide mnemonic decisions. Beyond successful classification of items perceived to be "old" or "new," MVPA classifiers could also reveal the specific type of "oldness" experienced by participants. In particular, Hits associated with subjectively reported contextual recollection were reliably discriminated from Hits associated with high confidence recognition without recollection, which were in turn discriminated from Hits associated with low confidence recognition. These classification analyses likely capitalized on neural signals related to recollection and item familiarity, respectively. Indeed, the importance maps revealed that regions of the hippocampus and angular gyrus, commonly associated with recollective retrieval (28, 43, 44), signaled diagnostic information for the classifier trained to differentiate R Hits from HC Hits, and yet provided limited information for the classifier trained to differentiate HC Hits from LC Hits. By contrast, this later classifier appeared to rely more heavily on regions of ventrolateral PFC and dorsal PPC, whose activity levels have previously been shown to track one's level of familiarity or mnemonic decision confidence (27, 39).

In sharp contrast to the robust classification of subjective recognition states, classifying an item's objective OLD/NEW status was far more challenging. When we assessed the decoding of objective recognition independent from subjective recognition—items were matched on their level of perceived oldness or perceived novelty—above-chance OLD/NEW classification was restricted to items participants assigned a "LC old" response (LC Hit/FA). Although the predictive value of this classification was relatively poor (mean AUC = 0.59), the modest success of this classifier suggests that the neural signatures of true and false recognition are at least sometimes distinguishable. This finding is consistent with previous fMRI studies using univariate statistical analyses, which have reported activation differences in the MTL (31–34, 45, 46) and sensory neocortex (30, 35, 45, 46) during true and false recognition. However, our inability to classify the objective OLD/NEW status of items that received a "LC new" response (LC Miss/CR) raises the possibility that our limited success on the LC Hit/FA classification exploited small subjective differences rather than neural signatures that tracked the veridical experiential history of stimuli per se.

To further assess whether stimulus experiential history can be decoded, we examined whether an MVPA approach could differentiate brain responses associated with OLD and NEW faces when participants performed an indirect (implicit) memory task. Numerous neuroimaging studies have documented activity reductions ("repetition suppression" or "fMRI adaptation") associated with the facilitated processing of previously encountered, relative to novel, stimuli (23, 37, 38); such "neural priming" effects are thought to be a hallmark of neocortical learning that supports nondeclarative memory. Although univariate analyses of the Implicit Recognition Task data from Exp. 2 revealed repetition suppression in regions of visual association cortex and anterior MTL (Fig. S4*B*), when these data were submitted to MVPA, the classifier exhibited an extremely poor ability to detect the OLD/NEW status of faces. Thus, these neural priming signals were likely too weak and variable across trials to effectively drive classifier performance. Furthermore, there was a low degree of overlap between the brain patterns associated with explicit and implicit recognition, as evidenced by the failure of a classifier trained on OLD vs. NEW discrimination using explicit retrieval data to perform above chance when tested on implicit retrieval data. These findings highlight the profound influence that goal states exert on the neural processes triggered by sensory inputs (47).

Taken together, our data raise critical questions about the utility of an fMRI-based approach for the detection of experiential knowledge. If one's goal is to detect neural correlates of subjective remembering, the data provide novel evidence that, at least under the constrained experimental conditions assessed here, this could be achieved with high accuracy, especially if only the classifier's most "confident" predictions are considered. Moreover, it appears that a participant's subjective recognition experiences can be decoded even when the classifier is trained on brain data from other participants, indicating that macroscopic (1) neural signatures of subjective recognition are highly consistent across individuals. Thus, from an applied perspective, this method might be able to indicate whether an individual subjectively remembers a stimulus, even when data from that individual are not available to train the classifier. On the other hand, an ideal memory detection technology would also be able to reveal whether a person had actually experienced a particular entity, without regard to his or her subjective report. Our data indicate that neural signatures of objective memory, at least for the simple events assessed here, are extremely challenging to detect reliably with current fMRI methods. This finding reveals a potentially significant boundary condition that may limit the ultimate utility of fMRI-based memory detection approaches for real-world application (see *SI Discussion* for consideration of additional boundary conditions). The neuroscientific and legal communities must maintain an ongoing dialogue (5) so that any future real-world applications will be based on, and limited by, controlled scientific evaluations that are well understood by the legal system before their use. Although false positives and false negatives can have important implications for memory theory, their consequences can be much more serious within a legal context.

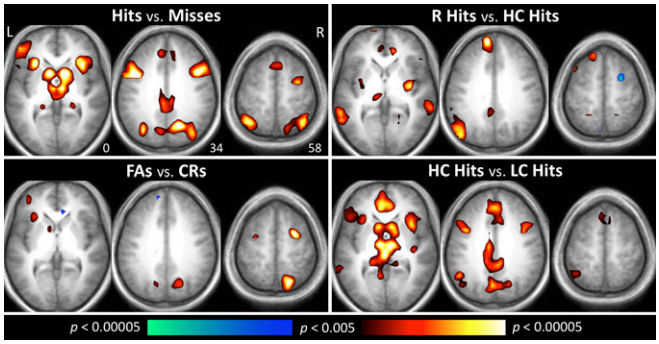

Fig. 3. Univariate contrast maps. Group *t* tests on activity parameter estimates (derived from a standard voxel-level general linear model-based analysis) illustrate regions with greater activity for trials from Class A (warm colors) or Class B (cool colors). The general correspondence between these univariate maps and the importance maps (Fig. 2) suggests that the classification analyses at least partially capitalized on large-scale (macroscopic) signal differences between conditions (see Figs. S3 and S4 for expanded univariate data reporting).

Methods

Exp. 1 Procedure. Before scanning, participants intentionally studied 210 faces, viewing each on a laptop computer for 4 s. Approximately 1 h later, participants were scanned while performing 400 trials of the Explicit Recognition Task (40 trials during each of 10 scanning runs). On each trial, a face was presented for 2 s, and participants indicated (with a 5-button response box in their right hand) whether they (*i*) recollected having studied the face (i.e., remembered contextual details associated with the initial encounter), (*ii*) were highly confident they studied it, (*iii*) thought they studied it, but had low confidence in this assessment, (*iv*) thought it was novel, but had low confidence in this assessment, or (*v*) were highly confident it was novel (see *SI Methods* for additional details). Stimulus presentation was followed by an 8-s fixation interval. One half of the test faces were novel (NEW) and one half were studied (OLD), with assignment counterbalanced across participants.

Exp. 2 Procedure. Exp. 2 was identical to Exp. 1, except for the following critical changes. Rather than being instructed to memorize the faces during the "study phase," participants were instructed to rate the attractiveness of each face on a 4-point scale. This task promoted attentive viewing and incidental encoding of the faces. Then, during the first five scanning runs, participants were instructed to make a button press response indicating whether each face was male or female. Half of the faces in each scan were OLD and half were NEW, but OLD/NEW status was not relevant to the male/female decision (Implicit Recognition Task). Immediately before the sixth scanning run, participants unexpectedly received a new set of task instructions—the same explicit recognition memory test instructions given to participants in Exp. 1—and they performed this Explicit Recognition Task for the remaining five scanning runs.

fMRI Data Analysis. Whole-brain imaging was conducted on a 3.0-T GE Signa MRI system, and standard data preprocessing procedures, including spatial normalization, were implemented. To reduce the fMRI time series data (TR = 2 s) to a single brain activity pattern for each of the 400 trials, the time points corresponding to the peak event-related hemodynamic response—namely, those occurring 4–8 s poststimulus, which translates to the third and fourth poststimulus TRs—were extracted and averaged. MVPA classification analyses were conducted using a regularized logistic regression algorithm, and performance was assessed using a cross-validation procedure (*SI Methods*).

ACKNOWLEDGMENTS. We thank Nina Poe, Felicity Grisham, Anna Parievsky, and Vincent Bell for helpful assistance with stimulus development, scanning, and data processing. Francisco Pereira contributed code for the RLR classification algorithm. This work was supported by grants from the John D. and Catherine T. MacArthur Foundation's Law and Neuroscience Project, and by National Institutes of Heath Grants R01-MH080309, R01-MH076932, and F32-NS059195.

1. Bles M, Haynes JD (2008) Detecting concealed information using brain-imaging technology. *Neurocase* 14:82–92.
2. Meegan DV (2008) Neuroimaging techniques for memory detection: scientific, ethical, and legal issues. *Am J Bioeth* 8:9–20.
3. Giridharadas A (Sept. 15, 2008) India's Novel Use of Brain Scans in Courts Is Debated. *The New York Times*, Section A, p 10.
4. Harrington v. Iowa No. PCCV 073247 (Dist. Ct. Iowa, March 5, 2001), discussed in Harrington v. Iowa, 659 NW 2d 509 (Iowa 2003).
5. Gazzaniga MS (2008) The law and neuroscience. *Neuron* 60:412–415.
6. Garland B, Glimcher PW (2006) Cognitive neuroscience and the law. *Curr Opin Neurobiol* 16:130–134.
7. Langleben DD, Dattilio FM (2008) Commentary: the future of forensic functional brain imaging. *J Am Acad Psychiatry Law* 36:502–504.
8. Feigenson N (2006) Brain imaging and courtroom evidence: on the admissibility and persuasiveness of fMRI. *Int J Law in Context* 2:233–255.
9. Greely HT, Illes J (2007) Neuroscience-based lie detection: the urgent need for regulation. *Am J Law Med* 33:377–431.
10. Rosenfeld JP (2005) 'Brain fingerprinting': A critical analysis. *Sci Rev Ment Health Pract* 4:20–37.
11. Editorial (2008) Deceiving the law. *Nat Neurosci* 11:1231.
12. Brown T, Murphy E (2010) Through a scanner darkly: Functional neuroimaging as evidence of a criminal defendant's past mental states. *Stanford Law Review* 62: 1119–1208.
13. Farwell LA, Donchin E (1991) The truth will out: interrogative polygraphy ("lie detection") with event-related brain potentials. *Psychophysiology* 28:531–547.
14. van Hooff JC, Brunia CH, Allen JJ (1996) Event-related potentials as indirect measures of recognition memory. *Int J Psychophysiol* 21:15–31.
15. Allen JJ, Iacono WG, Danielson KD (1992) The identification of concealed memories using the event-related potential and implicit behavioral measures: a methodology for prediction in the face of individual differences. *Psychophysiology* 29:504–522.
16. Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* 10:424–430.
17. Haynes JD, Rees G (2006) Decoding mental states from brain activity in humans. *Nat Rev Neurosci* 7:523–534.
18. Hassabis D, et al. (2009) Decoding neuronal ensembles in the human hippocampus. *Curr Biol* 19:546–554.
19. Polyn SM, Natu VS, Cohen JD, Norman KA (2005) Category-specific cortical activity precedes retrieval during memory search. *Science* 310:1963–1966.
20. Johnson JD, McDuff SG, Rugg MD, Norman KA (2009) Recollection, familiarity, and cortical reinstatement: a multivoxel pattern analysis. *Neuron* 63:697–708.
21. 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.
22. Chadwick MJ, Hassabis D, Weiskopf N, Maguire EA (2010) Decoding individual episodic memory traces in the human hippocampus. *Curr Biol* 20:544–547.
23. Grill-Spector K, Henson R, Martin A (2006) Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn Sci* 10:14–23.
24. Ranganath C, Rainer G (2003) Neural mechanisms for detecting and remembering novel events. *Nat Rev Neurosci* 4:193–202.
25. Kumaran D, Maguire EA (2009) Novelty signals: a window into hippocampal information processing. *Trends Cogn Sci* 13:47–54.
26. Desimone R (1996) Neural mechanisms for visual memory and their role in attention. *Proc Natl Acad Sci USA* 93:13494–13499.
27. Montaldi D, Spencer TJ, Roberts N, Mayes AR (2006) The neural system that mediates familiarity memory. *Hippocampus* 16:504–520.
28. Wagner AD, Shannon BJ, Kahn I, Buckner RL (2005) Parietal lobe contributions to episodic memory retrieval. *Trends Cogn Sci* 9:445–453.
29. Gonsalves BD, Kahn I, Curran T, Norman KA, Wagner AD (2005) Memory strength and repetition suppression: multimodal imaging of medial temporal cortical contributions to recognition. *Neuron* 47:751–761.
30. Danckert SL, Gati JS, Menon RS, Köhler S (2007) Perirhinal and hippocampal contributions to visual recognition memory can be distinguished from those of occipito-temporal structures based on conscious awareness of prior occurrence. *Hippocampus* 17:1081–1092.
31. Cabeza R, Rao SM, Wagner AD, Mayer AR, Schacter DL (2001) Can medial temporal lobe regions distinguish true from false? An event-related functional MRI study of veridical and illusory recognition memory. *Proc Natl Acad Sci USA* 98:4805–4810.
32. Daselaar SM, Fleck MS, Prince SE, Cabeza R (2006) The medial temporal lobe distinguishes old from new independently of consciousness. *J Neurosci* 26:5835–5839.
33. Hannula DE, Ranganath C (2009) The eyes have it: hippocampal activity predicts expression of memory in eye movements. *Neuron* 63:592–599.
34. Kirwan CB, Shrager Y, Squire LR (2009) Medial temporal lobe activity can distinguish between old and new stimuli independently of overt behavioral choice. *Proc Natl Acad Sci USA* 106:14617–14621.
35. Slotnick SD, Schacter DL (2004) A sensory signature that distinguishes true from false memories. *Nat Neurosci* 7:664–672.
36. Boehm SG, Paller KA (2006) Do I know you? Insights into memory for faces from brain potentials. *Clin EEG Neurosci* 37:322–329.
37. Schacter DL, Wig GS, Stevens WD (2007) Reductions in cortical activity during priming. *Curr Opin Neurobiol* 17:171–176.
38. Race EA, Shanker S, Wagner AD (2009) Neural priming in human frontal cortex: multiple forms of learning reduce demands on the prefrontal executive system. *J Cogn Neurosci* 21:1766–1781.
39. Yonelinas AP, Otten LJ, Shaw KN, Rugg MD (2005) Separating the brain regions involved in recollection and familiarity in recognition memory. *J Neurosci* 25: 3002–3008.
40. Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit* 30:1145–1159.
41. Wheeler ME, Buckner RL (2003) Functional dissociation among components of remembering: control, perceived oldness, and content. *J Neurosci* 23:3869–3880.
42. Kahn I, Davachi L, Wagner AD (2004) Functional-neuroanatomic correlates of recollection: implications for models of recognition memory. *J Neurosci* 24:4172–4180.
43. Diana RA, Yonelinas AP, Ranganath C (2007) Imaging recollection and familiarity in the medial temporal lobe: a three-component model. *Trends Cogn Sci* 11:379–386.
44. Vilberg KL, Rugg MD (2008) Memory retrieval and the parietal cortex: a review of evidence from a dual-process perspective. *Neuropsychologia* 46:1787–1799.
45. Garoff-Eaton RJ, Slotnick SD, Schacter DL (2006) Not all false memories are created equal: the neural basis of false recognition. *Cereb Cortex* 16:1645–1652.
46. Okado Y, Stark C (2003) Neural processing associated with true and false memory retrieval. *Cogn Affect Behav Neurosci* 3:323–334.
47. Dudukovic NM, Wagner AD (2007) Goal-dependent modulation of declarative memory: neural correlates of temporal recency decisions and novelty detection. *Neuropsychologia* 45:2608–2620.

Supporting Information

Rissman et al. 10.1073/pnas.1001028107

SI Results

Assessment of Subjective Old vs. New Classification Performance When Participants' Decision Confidence Is Held Constant. When trained/tested on the Hit/Miss or FA/CR classification scheme, it is possible that the classifier does not learn to identify neural signatures of subjective oldness vs. novelty per se but rather learns correlates of participants' decision confidence. This issue is especially pertinent to the Hit/Miss classification, where most of the Misses were low confidence responses ("LC new") whereas the majority of Hits were high confidence responses ("R old" or "HC old"). Accordingly, we reran these classifications while holding participant decision confidence constant (because there were relatively few high confidence Misses or FAs, we used only low confidence responses). For the LC Hit/Miss classification, AUC was 0.67 [$t_{(11)} = 8.22$, $P < 10^{-5}$]; for the LC FA/CR classification, AUC was 0.66 [$t_{(14)} = 8.42$, $P < 10^{-6}$]. Although modest, both performance declines were significant ($P < 10^{-3}$), possibly revealing that the more inclusive Hit/Miss and FA/CR classifiers partially exploited neural signals related to participant decision confidence (in addition to signals related to subjective mnemonic experience). Alternatively, the declines could result from the classifier being trained on fewer trials in the LC-only classification schemes, or from a diminution in the average strength of the subjective oldness or novelty experienced by participants (including the fact that the percentage of guesses is inherently higher for LC, relative to HC, responses). In either case, the fact that classification performance levels were still robust after controlling for both objective oldness and subjective decision confidence—with mean accuracy for the LC Hit/Miss and LC FA/CR classifications reaching 77% and 74%, respectively, at the strongest classifier "confidence" level—suggests that the subjective mnemonic experience triggered by a test face can be reliably decoded from single-trial fMRI activity patterns.

Evaluating Mnemonic Decoding Performance Based on Individual Anatomical Regions. Although the classifier-derived importance maps presented in Fig. 2 and Fig. S2 reveal the voxels whose activity levels most strongly influenced classifier performance when all 23,000 voxels within our anatomical mask were used as features for classification, it is also of interest to examine whether the brain activity patterns within individual anatomical regions contain sufficient information to allow decoding of mnemonic states. To this end, we evaluated classification performance within 80 distinct anatomical regions-of-interest (ROIs), defined by intersecting individual ROI masks from the Automated Anatomical Labeling (AAL) library (1) with our 23,000-voxel anatomical mask.

For the various classification schemes assessing subjective recognition, a number of ROIs (particularly those in PFC and PPC) supported classification performance (AUC) levels that were ≈4–7% lower than that obtained with whole-brain classification (Table S3). Moreover, for the LC Hits vs. LC FAs classification of objective OLD/NEW status, several ROIs supported performance levels on par or just slightly below the whole-brain level (Table S3, rightmost column). Here, the top performing ROIs included a few of the same PFC and PPC regions that emerged as top performers in the subjective mnemonic classification analyses. Importantly, however, this modest classification of objective OLD/NEW status was also possible from data within visual areas commonly associated with face processing, including regions of the fusiform gyrus, middle occipital cortex, and middle temporal gyrus. These later effects raise the possibility that objective recognition is associated with changes in local brain activity patterns linked to the perceptual analysis of the stimulus. Although ROI-based classification did not reveal medial temporal lobe ROIs as being among the top 10 highest performing regions for any of the six classification schemes (Table S3), classification based on hippocampal and parahippocampal ROIs reached or exceeded an AUC of 0.60 for the Hits vs. CRs and R Hits vs. HC Hits analyses.

Evaluating Mnemonic Decoding Performance Based on Local Distributed Patterns. A complementary method of evaluating the informational content represented within local brain activity patterns is the spherical searchlight mapping approach (2). This method involves running a large series of classification analyses, each using only a small spherical clique of voxels, and recording the classification performance level for each sphere. The center of the sphere is systematically shifted (like a searchlight) until classification performance has been recorded for spheres centered at every voxel location with the brain. For our purposes, this involved running 23,000 classification analyses, each using only a 123-voxel cluster of unsmoothed fMRI data (i.e., those voxels within a 3-voxel radius of the central voxel; note that sphere size diminishes for regions near the edge of the brain mask), and recording the classification AUC value at the center voxel of each sphere.

Group-averaged searchlight maps for classifications of subjective mnemonic status are displayed in Fig. S3, along with the corresponding set of univariate contrasts. These searchlight maps, which reveal brain regions whose local voxel activity patterns are capable of differentiating the two mnemonic states of interest, highlight similar brain networks as those seen in the importance maps that were derived from the whole-brain classification analyses (Fig. S2). As one might expect from the individual ROI classification analyses reported above, no single 123-voxel spherical cluster was capable of achieving classification performance levels as high as those obtained for the whole-brain classification analysis. Nonetheless, spheres centered in many brain regions showed fairly robust classification abilities. Importantly, when the searchlight maps are viewed alongside the corresponding univariate contrasts (i.e., group-level t tests on the voxel-by-voxel activation parameter estimates generated from a standard general linear model analysis; Fig. S3), it is readily apparent that the regions where the searchlight analysis produced the highest classification performance levels were typically the same regions that showed robust univariate activation differences between the two conditions. This makes the point that differences in the mean signal level within a region across examples of Class A and Class B may strongly drive classification performance [i.e., the signals driving classifier performance were macroscopic (3)].

By contrast, an examination of the searchlight maps for the classifications of objective recognition revealed a somewhat different picture. The LC Hits vs. LC FAs classification analysis, which produced above-chance performance using a whole-brain classifier (mean AUC = 0.59), yielded a few regions with modest levels of classification success when analyzed with the searchlight mapping approach (Fig. S4*A*). In particular, a region of right fusiform/inferotemporal cortex and a region of left medial superior frontal gyrus (mSFG) showed the most robust ability to discriminate LC Hits from LC FAs (peak AUC values for both regions reached 0.57, which is still quite poor in comparison with the classification performance levels observed for the subjective recognition classification schemes). In contrast to the general correspondence between the searchlight maps and univariate contrasts that was noted for the analyses of subjective recognition, the univariate contrast of LC Hits vs. LC FAs did not reveal effects in the fusi-

form/inferotemporal cortex or mSFG at an uncorrected threshold of $P < 0.005$ (Fig. S4*A*), nor at an even more liberal uncorrected threshold of $P < 0.05$. Thus, it is possible that the neural signatures of objective recognition in these regions are more readily detectable when the fine-grained local activation patterns are exploited using an MVPA searchlight approach. The finding that LC Hits can be distinguished from LC FAs (albeit weakly) based on brain activity within the right fusiform/inferotemporal cortex, an area associated with face processing, is consistent with the notion that the perceptual qualities of true memories may differ from those associated with false memories (4). Finally, no brain regions showed even modest classification abilities in the searchlight analysis of LC Misses vs. LC CRs (all AUCs < 0.53), which is unsurprising, given that performance for this classification scheme was also at chance using the whole-brain data.

Classifying Mnemonic States Using Large-Scale Regional Activity Profiles. All of the classification analyses described thus far have used individual voxel activity values as features. Such an approach allows the classification algorithm to capitalize on the information that might be represented within fine-grained activation patterns. However, as discussed above, the marked correspondence between our spherical searchlight classification maps and univariate contrast maps suggests that relatively large clusters of adjacent voxels may exhibit activation levels that favor one condition over another, and thus are diagnostic of mnemonic status. To the extent that large-scale regional activity profiles are able to distinguish trials from two distinct conditions, then a classifier that is trained with a spatially coarse representation of the data should still achieve reasonable performance.

We assessed the degree to which macroscopic activation patterns could be exploited to decode mnemonic states by rerunning our classification analyses after averaging the fMRI activity levels across all voxels within each of 80 distinct anatomical ROIs (defined, as described above, using the AAL library). Thus, rather than feeding our classification algorithm the activity values of 23,000 voxels as features, here we performed the same analysis with only 80 features. Despite the fact that this data reduction procedure eliminated the fine-grained information contained within individual brain regions, classification performance remained surprisingly robust [mean AUCs: Hits vs. CRs = 0.77; Hits vs. Misses = 0.70; FAs vs. CRs = 0.64; R Hits vs. HC Hits = 0.71; HC Hits vs. LC Hits = 0.68; LC Hits vs. LC FAs = 0.57; performance for the LC Misses vs. LC CRs classification (AUC = 0.51) remained at chance]. These results illustrate that a variety of mnemonic states can be differentiated based on their macroscopic activity profile, which likely tracks the generalized engagement of cognitive processes associated with distinct memory states rather than retrieval of specific fine-grained mnemonic representations (3). The diagnostic value of such macroscopic activation effects may also explain the ability of our standard whole-brain classification analyses to succeed at across-participant memory decoding (Fig. S1).

SI Discussion

Additional Factors That Will Likely Influence fMRI-Based Memory Detection. The results of our study suggest that the forensic value of an fMRI-based memory detection technique may be limited by the fact that objective mnemonic classification performance is very poor when (*i*) subjective memory judgments are held constant or (*ii*) memory is indirectly (implicitly) probed. These findings highlight the possibility of additional boundary conditions. For example, it has been shown that participants can adopt simple countermanding strategies to conceal the presence of "guilty knowledge" in studies that use EEG (5), skin conductance (6, 7), or reaction time (ref. 8; cf. ref. 9) measures to probe participants' memories. The present fMRI data similarly indicate that a change in participants' goal states (e.g., making male/female judgments instead of recognition memory judgments) can dramatically reduce the ability to decode neural correlates of experiential knowledge. As such, it seems likely that the use of countermanding strategies will also decrease fMRI-based classifier accuracy for discriminating both subjective and objective memory states. Future studies should systematically address the effects of countermanding strategies to determine whether they place further constraints on the forensic utility of fMRI for memory detection. It will also be critical to assess mnemonic classification accuracy for more ecologically valid experiences, because life's events are often considerably richer than the simplified events assessed here; rich autobiographical memories may have different neural signatures than those emerging in highly controlled, list-learning experiments (10, 11). It is theoretically possible that fMRI-based decoding of objective experiential history may be superior for complex real-world events, relative to laboratory-induced experiences with individual stimuli.

Before accepting the validity of potential forensic applications, it will also be important to evaluate memory detection in more realistic "forensic" contexts, such as scenarios in which participants commit a mock-crime and subsequently attempt to conceal their guilty knowledge while their memory for particular events is probed (12, 13). However, such experimental paradigms may still fail to induce the feelings of anxiety and sense of jeopardy that characterize real-world interrogations, and thus their ecological validity remains in question. It will also be imperative to enroll a more diverse sample of participants to assess whether our results can be generalized to the broader population (e.g., older vs. younger adults). Finally, the error rates (false positives and false negatives) of any viable memory detection approach will have to be quite well established, and the legal system will ultimately have to determine whether those error rates are acceptable for any particular use that might be made of the technique (14).

Whereas the present investigation focused on fMRI-based classification of recognition memory states, other recent fMRI studies have achieved some success at applying MVPA techniques to probe the nature of the representations that are retrieved from memory. Within the context of circumscribed task conditions, it is possible to achieve above-chance classification of the category of information about to be recalled from memory (15), the particular contextual associations brought back to mind during a retrieval attempt (16, 17), some details about one's recent navigational history in an environment (18), and which of several discrete episodes an individual is currently recollecting (19). Such findings highlight the potential of fMRI to read out categorical aspects of the content of what an individual is currently retrieving from memory. Although future work will bear on the forensic potential of such demonstrations of mnemonic decoding, it is possible that the pervasive phenomenon of false remembering (20) will limit conceivable practical applications.

Limitations of EEG-Based Approaches to Mnemonic Classification. Due to the inherently noisy nature of scalp recordings, extant EEG-based techniques for probing experiential knowledge (21–23) are generally unable to classify the mnemonic status of individual stimuli, but rather must average across a large number of distinct memory probes to achieve their results. Thus, this approach principally assesses whether a certain set of stimuli are recognized by the participant, whereas an ideal memory detection technique should be capable of classifying the mnemonic status of each individual probe stimulus. Furthermore, because EEG-based techniques often rely on detecting the neural signature of an attentional orienting response to "guilty knowledge" stimuli, they are susceptible to a variety of countermeasures (5), in which participants willfully manipulate their attentional state in such a manner as to substantially diminish the classification accuracy of the procedure. Thus, current EEG-based memory detection techniques may fall short of the methodological rigor, reliability, and scientific acceptance necessary to meet the standards (24) (Federal Rules of Evidence 702) for legal admissibility of scientific evidence (refs.

25 and 26, but see ref. 27 for an alternative perspective). That said, it remains an open question whether EEG-based methods might be capable of achieving more reliable mnemonic decoding performance if the EEG data were collected using a similar experimental paradigm and submitted to an analogous trial-by-trial multivariate pattern classification approach as that used in the present fMRI study.

SI Methods

Participants. Two independent samples were enrolled. Sixteen healthy right-handed adults (10 females; ages 18–27 yr, mean age = 21.4 yr) participated in Exp. 1 and seven participated in Exp. 2 (3 females; ages 19–30 yr, mean age = 22.7 yr). Participants were recruited from the Stanford University community and surrounding areas. Written informed consent was obtained in accordance with procedures approved by the institutional review board at Stanford University. Participants received $20/h for their participation, and the experimental sessions lasted ≈3.5 h. For Exp. 1, data from three additional individuals were excluded from analysis due to inadequate behavioral performance (one participant reported falling asleep for brief intervals throughout the experiment, one exhibited poor recognition memory: $d' = 0.39$, and one had atypically slow reaction times, with over 25% of responses taking > 5 s to execute). For Exp. 2, data from two additional individuals were excluded due to excessive head motion and poor recognition memory performance ($d' = 0.43$), respectively.

Stimulus Materials. A set of 420 color photographs of faces was used in Exps. 1 and 2. The collection of 210 male and 210 female faces was comprised of individuals of varied ages and races/ethnicities, and was compiled from an in-house stimulus collection as well as from the following online databases (with permission, where applicable): AR Face Database (http://cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html), CalTech Database (www.vision.caltech.edu/html-files/archive.html), CVL Face Database (www.lrv.fri.uni-lj.si/facedb.html), FERET Database (www.nist.gov/humanid/feret/feret_master.html), GTAV Face Database (http://gps-tsc.upc.es/GTAV/ResearchAreas/UPCFaceDatabase/GTAVFaceDatabase.htm), NimStim Face Stimulus Set (www.macbrain.org/resources.htm), and the Productive Aging Laboratory Face Database (https://pal.utdallas.edu/facedb). Using Adobe Photoshop, the faces were extracted from their backgrounds (with the hair included), manually retouched to ensure proper brightness and contrast, cropped in a consistent manner (from the base of the neck to the top of the head), and presented against a solid white background. For presentation outside the scanner (laptop computer) and inside the scanner (projection screen), stimuli were displayed at 240 × 300 pixels on an 800 × 600-resolution screen.

Supplemental Procedural Details, Exp. 1. For the study phase of the experiment, participants were explicitly instructed to attentively view the faces and try their best to encode them into memory. To ensure attention throughout the study session, participants were instructed to press the space bar during the 1.5-s interstimulus interval that followed each face. Participants were given an opportunity to take a brief break after every 30 stimuli; the entire study session lasted ≈25 min. After completing the study session, participants received detailed instructions describing the five response options for the Explicit Recognition Task, with emphasis on the qualitative distinction between recollection (recognition accompanied by reinstatement of contextual details) and high confidence recognition (putatively strong familiarity in the absence of recollection, but see ref. 28). Participants were instructed to respond to every face and were encouraged to do so quickly, but accurately.

Proper understanding of the recognition task instructions was confirmed during a practice testing session. Participants performed 20 practice trials on a laptop computer (10 OLD and 10 NEW faces, intermixed; the OLD faces consisted of the first five and last five faces encountered in the study session). During the first 10 practice trials, each face remained on the screen until the participant made a response, and gave a verbal description to the experimenter as to the basis for the particular recognition rating assigned to the face. After each response, participants received feedback as to whether the face was actually studied or novel. The next 10 trials had the same stimulus presentation parameters as in the actual fMRI experiment.

Supplemental Procedural Details, Exp. 2. In contrast to Exp. 1, during the "study phase" of Exp. 2 (i.e., attractiveness ratings task), participants were not informed that their memory for the faces would eventually be tested. Moreover, during the first five scanning runs, participants were instructed to make a male/female judgment about each face using the index and middle fingers of the right hand, with button assignment counterbalanced across participants. Only after completion of these five runs of the Implicit Recognition Task were participants informed of the impending Explicit Recognition memory test. At this point, participants were given instructions for the Explicit Recognition Task; before scanning recommenced, they practiced this task using the same stimuli and practice protocol as in Exp. 1 (although participants did not give the experimenter verbal descriptors justifying each response).

fMRI Data Acquisition. Functional images were collected using a T2*-weighted 2D gradient echo spiral-in/out pulse sequence (TR = 2.0 s; TE = 30 ms; flip angle = 75; FoV = 22 cm, in-plane resolution = 3.4375 mm × 3.4375 mm) (29). Each functional volume consisted of 30 contiguous slices acquired parallel to the AC-PC plane. Slice thickness was 4.0 mm in Exp. 1 and 3.8 mm in Exp. 2. Anatomical images coplanar with the functional data were collected at the start of the experiment using a T2-weighted flow-compensated spin-echo pulse sequence. A T1-weighted whole-brain spoiled gradient recalled (SPGR) 3D anatomical image was acquired at the end of the experimental session. Owing to technical difficulties, in Exp. 1, one participant's fMRI dataset is missing two functional runs (s102) and two additional participants (s103 and s115) are each missing one run; in Exp. 2, one participant's dataset (s205) is missing one functional run from the Explicit Recognition Task, and for another participant (s201) one run of the Implicit Recognition Task was discarded due to excessive nonresponses.

fMRI Data Analysis. The six initial volumes of each run were discarded to allow for T1 equilibration. Following reconstruction, a series of fMRI data preprocessing routines were implemented using SPM5 (www.fil.ion.ucl.ac.uk/spm). Functional images were corrected to account for differences in slice acquisition times using sinc interpolation, with the center slice used as a reference point. These data were then motion corrected using a two-pass, six-parameter, rigid-body realignment procedure. If, during the course of any trial, the participant moved at a rate of >0.5 mm/TR or the global signal (averaged across all brain voxels) deviated by more than 3.5 SDs from the run's mean, then that trial's data were excluded from analysis. Each participant's T1-weighted whole-brain anatomical image was coregistered to the T2-weighted coplanar anatomical image, and these in turn were coregistered to the mean functional image. The coregistered T1 image was then segmented into gray matter, white matter, and cerebrospinal fluid, and the gray matter image was spatially normalized to a gray matter template image in Montreal Neurological Institute (MNI) stereotactic space. The resulting transformation parameters were used to warp all structural and functional images into MNI space, and the functional images were resampled into 4-mm isotropic voxels and spatially smoothed with an 8-mm FWHM Gaussian kernel. Although not always used in MVPA analyses, spatial smoothing can increase the signal-to-noise ratio, making large-scale spatial patterns easier to detect. We found that smoothing

generally improved our classification accuracy by several percentage points.

Additional preprocessing steps were performed separately for each functional run using MATLAB routines provided in the Princeton MVPA Toolbox (www.csbmb.princeton.edu/mvpa). Each voxel's time series was high-pass filtered to remove frequencies below 0.01 Hz, detrended to remove linear and quadratic trends, and *z*-scored, so as to normalize each voxel's time series to have a mean of zero and a variance of one. To reduce the fMRI time series data to a single brain activity measure for each of the 400 test trials, the time points corresponding to the peak event-related hemodynamic response—namely, those occurring 4–8 s post stimulus, which translates to the third and fourth post-stimulus TRs—were extracted and averaged. A common 23,000-voxel inclusive mask was applied to the spatially normalized data of all participants to exclude the cerebellum and motor, premotor, and somatosensory cortices, which prevented the classifier from exploiting brain activity differences that might be linked to the motor responses associated with the distinct mnemonic states.

Pattern classification analyses were implemented in MATLAB using routines from the Princeton MVPA Toolbox and custom code. The brain activity pattern associated with each trial was labeled according to its objective mnemonic status (OLD or NEW) and its subjective mnemonic status ("R old," "HC old," "LC old," "LC new," "HC new"), resulting in 10 trial types. Trials in which no behavioral response was recorded were excluded from analysis. In each classification analysis, we assessed how accurately the classifier could discriminate between trials from two distinct mnemonic conditions (abstractly referred to as Class A vs. Class B), each of which was defined by a single trial type or a combination of trial types. Except where otherwise indicated, classification performance was assessed separately on each participant's data using a 10-fold cross-validation procedure. Trials from Class A and Class B were randomly divided into 10 balanced subsets, with each subset containing an equal number of trials from each class (note that the division of trials into these 10 subsets was not constrained by scanning run boundaries, and thus 10-fold cross-validation was used even for participants with missing runs and for the Exp. 2 data, which had only five runs for each task). The trials from 9 of these subsets were used for classifier training, and the held-out trials were used as a test set for assessing generalization performance. This process was iteratively repeated with each of the 10 subsets of trials held-out, such that unbiased classifier outputs were measured for all of the selected trials. Balancing the number of trials from each class prevented the classifier from developing a bias to identify trials as belonging to the more plentiful class, and ensured a theoretical null hypothesis classification accuracy rate of 50% and AUC of 0.5 (analyses with shuffled class labels confirmed that chance classification performance converged tightly around these levels for all classification schemes, as well as across all levels of classifier "confidence"). Following this balancing procedure, the data from each voxel were *z*-scored again, such that each voxel's mean activity level for Class A trials was the inverse of its mean activity level for Class B trials. For any given classification, participants with fewer than 18 trials/class were excluded, because having an insufficient number of training examples can result in unstable classifier performance. Table S2 reports the mean number of trials contributing to each classification analysis. To achieve stable results, all classification analyses were repeated 20 times, using a different subset of trials (from the more plentiful class) each time, and the results were then averaged.

The two exceptions to our use of the 10-fold cross-validation procedure were (*i*) the across-participant classification analyses conducted on the Exp. 1 data and (*ii*) the across-task (within-participant) classification analyses conducted on the Exp. 2 data. The across-participant classification analyses were conducted using a leave-one-participant-out procedure, in which a classifier was trained on the combined set of data from all but one participant and tested on the data from the held-out participant (note that this analysis, like all of our classification analyses, still operated on the data from individual trials). In Exp. 2, across-task classification analyses were conducted using all trials from one task (e.g., the Explicit Recognition Task) for training and all trials from the other task (e.g., the Implicit Recognition Task) for testing.

A variety of machine learning algorithms have been successfully used to decode cognitive states from fMRI data (30). Here, we explored several algorithms, including two-layer back-propagation neural networks, linear support vector machines, and regularized logistic regression (RLR). Although all three performed well, we found that RLR generally outperformed the other techniques, if by only a small amount (see Fig. S6*A* for an example comparison of classification accuracy across these techniques). Thus, we elected to use RLR for all classification analyses reported in the manuscript. The RLR algorithm implemented a multiclass logistic regression function using a softmax transformation of linear combinations of the features, as described in (31), with an additional ridge penalty term as a Gaussian prior on the feature weights. This penalty term provided L_2 regularization, enforcing small weights, but not aggressively driving the majority of the weights to zero, as would be accomplished by L_1 regularization (see Fig. S6*B* for a demonstration that L_1 regularization failed to improve classification performance beyond the levels achieved with L_2 regularization). 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 (www.kyb.tuebingen.mpg.de/bs/people/carl/code/minimize) using the gradient of the log likelihood combined with the L_2 penalty.

The L_2 penalty was set to be half of the additive inverse of a user-specified λ parameter, multiplied by the square of the L_2 norm of the weight vector for each class, added over classes. We elected to set this free λ parameter to a fixed value of 1 for all within-participant classification analyses reported in this manuscript. This particular λ value was selected based on a cursory examination of several different fixed λ levels, as well an evaluation of classifier performance using a nested penalty optimization routine (i.e., subdividing each training set into new training and testing sets and determining the penalty parameter that maximizes classification accuracy, and then subsequently applying this parameter to the original held-out testing set). Performance gains, when present, were minimal when using higher fixed settings of λ or when applying the computationally intensive nested penalty optimization routine (Fig. S6*B*). For the across-participant classification schemes, we found that a more aggressive penalty setting of $\lambda = 10{,}000$ reduced over-fitting and improved generalization performance by 5–7% on average; thus this higher λ value was applied to all across-participant classification analyses.

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 concatenated across all cross-validation testing folds and then ranked. The classifier's true positive (hit) rate and false positive (false alarm) rate were calculated at 80 fixed cutoff thresholds along the probability continuum to generate ROC curves. The AUC values associated with these curves were computed as described in Fawcett (32) and can be formally interpreted 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. The ROC curves themselves provide further valuable information. For example, if one's goal is to sensitively detect examples of Class A, and one is willing to misclassify a certain proportion of Class B

examples to achieve this goal, the classifier's decision boundary (criterion) can be set toward the right of the curve. Conversely, if one desires greater specificity in labeling examples of Class A and is unwilling to tolerate many false positives, the decision boundary can be set toward the left of the curve.

Prior fMRI studies using pattern classification techniques have often implemented an initial feature selection step to eliminate uninformative voxels, because the inclusion of these voxels sometimes reduces classification performance. An exploratory analysis examining whether feature selection would impact our classifier performance revealed that performance did not decline when all 23,000 voxels within our mask were included as features, although classification performance was almost as good when nearly half this number of voxels were used (Fig. S6*C*). This outcome likely reflects the ability of the RLR classifier to effectively reduce the weight values of voxels that provide little relevant information to the classifier. Thus, we elected not to use feature selection (beyond our anatomical mask) for any of the classification analyses.

For each classification scheme, importance maps were constructed following the procedure described in previous MVPA studies (16, 17). A voxel's importance value provides an index of how much its signal increases or decreases influence the classifier's predictions. Following training, the logistic regression classification procedure yields a set of β weight values reflecting each voxel's predictive value (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 β weights were then multiplied by each voxel's mean activity level for Class A trials (which, owing to 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 (16, 17). Group-level summary maps were created by averaging the importance maps of the individual participants. Owing to poor levels of overall classification performance, importance maps for the analyses of objective recognition offer little informative value and thus are not reported. 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.

Univariate data analyses were conducted using SPM5, with trials modeled as events convolved with a canonical hemodynamic response function. The resulting functions were entered into a general linear model with motion parameters included as a covariate. Linear contrasts were used to obtain participant-specific activation parameter estimates for each condition of interest. These estimates were then entered into a second-level analysis, treating participants as a random effect, using a one-sample *t* test against a contrast value of zero at each voxel.

1. Tzourio-Mazoyer N, et al. (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15:273–289.
2. Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *Proc Natl Acad Sci USA* 103:3863–3868.
3. Bles M, Haynes JD (2008) Detecting concealed information using brain-imaging technology. *Neurocase* 14:82–92.
4. Schacter DL, Slotnick SD (2004) The cognitive neuroscience of memory distortion. *Neuron* 44:149–160.
5. 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.
6. Honts CR, Devitt MK, Winbush M, Kircher JC (1996) Mental and physical countermeasures reduce the accuracy of the concealed knowledge test. *Psychophysiology* 33:84–92.
7. Ben-Shakhar G, Dolev K (1996) Psychophysiological detection through the guilty knowledge technique: Effects of mental countermeasures. *J Appl Psychol* 81: 273–281.
8. Verschuere B, Prati V, Houwer JD (2009) Cheating the lie detector: Faking in the autobiographical Implicit Association Test. *Psychol Sci* 20:410–413.
9. Sartori G, Agosta S, Zogmaister C, Ferrara SD, Castiello U (2008) How to accurately detect autobiographical events. *Psychol Sci* 19:772–780.
10. Cabeza R, et al. (2004) Brain activity during episodic retrieval of autobiographical and laboratory events: An fMRI study using a novel photo paradigm. *J Cogn Neurosci* 16: 1583–1594.
11. McDermott KB, Szpunar KK, Christ SE (2009) Laboratory-based and autobiographical retrieval tasks differ substantially in their neural substrates. *Neuropsychologia* 47: 2290–2298.
12. Mertens R, Allen JJ (2008) The role of psychophysiology in forensic assessments: Deception detection, ERPs, and virtual reality mock crime scenarios. *Psychophysiology* 45:286–298.
13. Ben-Shakhar G, Elaad E (2003) The validity of psychophysiological detection of information with the Guilty Knowledge Test: A meta-analytic review. *J Appl Psychol* 88:131–151.
14. Schauer F (2010) Neuroscience, lie-detection, and the law: Contrary to the prevailing view, the suitability of brain-based lie-detection for courtroom or forensic use should be determined according to legal and not scientific standards. *Trends Cogn Sci* 14: 101–103.
15. Polyn SM, Natu VS, Cohen JD, Norman KA (2005) Category-specific cortical activity precedes retrieval during memory search. *Science* 310:1963–1966.
16. 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.
17. Johnson JD, McDuff SG, Rugg MD, Norman KA (2009) Recollection, familiarity, and cortical reinstatement: A multivoxel pattern analysis. *Neuron* 63:697–708.
18. Hassabis D, et al. (2009) Decoding neuronal ensembles in the human hippocampus. *Curr Biol* 19:546–554.
19. Chadwick MJ, Hassabis D, Weiskopf N, Maguire EA (2010) Decoding individual episodic memory traces in the human hippocampus. *Curr Biol* 20:544–547.
20. Loftus EF (2005) Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learn Mem* 12:361–366.
21. Farwell LA, Donchin E (1991) The truth will out: Interrogative polygraphy ("lie detection") with event-related brain potentials. *Psychophysiology* 28:531–547.
22. van Hooff JC, Brunia CH, Allen JJ (1996) Event-related potentials as indirect measures of recognition memory. *Int J Psychophysiol* 21:15–31.
23. Allen JJ, Iacono WG, Danielson KD (1992) The identification of concealed memories using the event-related potential and implicit behavioral measures: A methodology for prediction in the face of individual differences. *Psychophysiology* 29:504–522.
24. Lumer ED, Friston KJ, Rees G (1998) Neural correlates of perceptual rivalry in the human brain. *Science* 280:1930–1934.
25. Appelbaum PS (2007) The new lie detectors: Neuroscience, deception, and the courts. *Psychiatr Serv* 58:460–462.
26. Meegan DV (2008) Neuroimaging techniques for memory detection: scientific, ethical, and legal issues. *Am J Bioeth* 8:9–20.
27. Iacono WG (2008) The forensic application of "brain fingerprinting": Why scientists should encourage the use of P300 memory detection methods. *Am J Bioeth*, 8:30–32; discussion W31–W34.
28. Wais PE, Mickes L, Wixted JT (2008) Remember/know judgments probe degrees of recollection. *J Cogn Neurosci* 20:400–405.
29. Glover GH, Law CS (2001) Spiral-in/out BOLD fMRI for increased SNR and reduced susceptibility artifacts. *Magn Reson Med* 46:515–522.
30. Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fMRI: A tutorial overview. *Neuroimage* 45 (1, Suppl):S199–S209.
31. Bishop CM (2006) *Pattern Recognition and Machine Learning* (Springer, Berlin), p 209.
32. Fawcett T (2003) ROC graphs: Notes and practical considerations for data mining researchers. *Technical Report HPL-2003-4* (HP Laboratories, Palo Alto, CA).

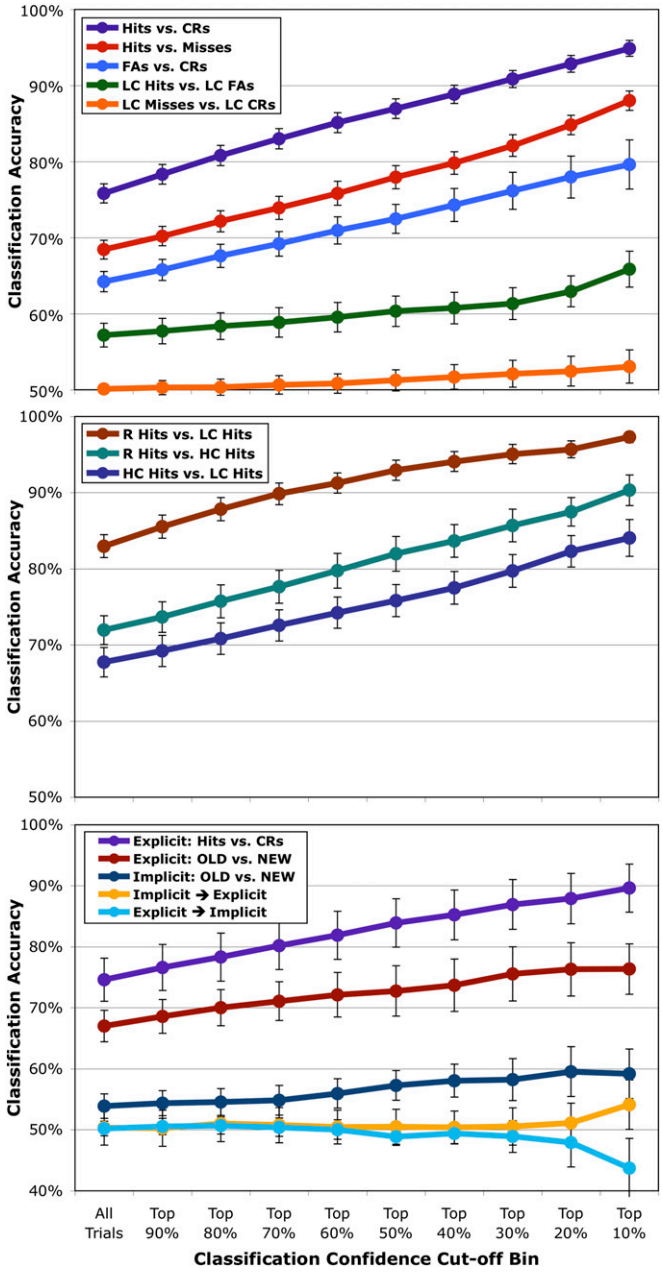

Fig. S1. Classification accuracy measures (mean ± SEM) for all classifications presented in Fig. 1. Overall classification accuracy is computed as the percentage of all trials for which the classifier's guess was correct. This measure, although commonly used to index classification performance in fMRI MVPA studies, effectively ignores the fact that each of the classifier's guesses is associated with a probability estimate. If one restricts the classifier's guesses to those test trials for which it has a higher degree of "confidence," classification accuracy will generally increase. Classifier confidence can be indexed by the absolute value of the difference of the two probability estimates for Class A and Class B. We ranked classification outputs according to this confidence metric and recomputed classification accuracy when the top N% most confidently classified trials were included (here, the N values represent deciles). At each of these inclusion thresholds, the trials that are excluded can be thought of as being assigned an "insufficient evidence" response from the classifier.

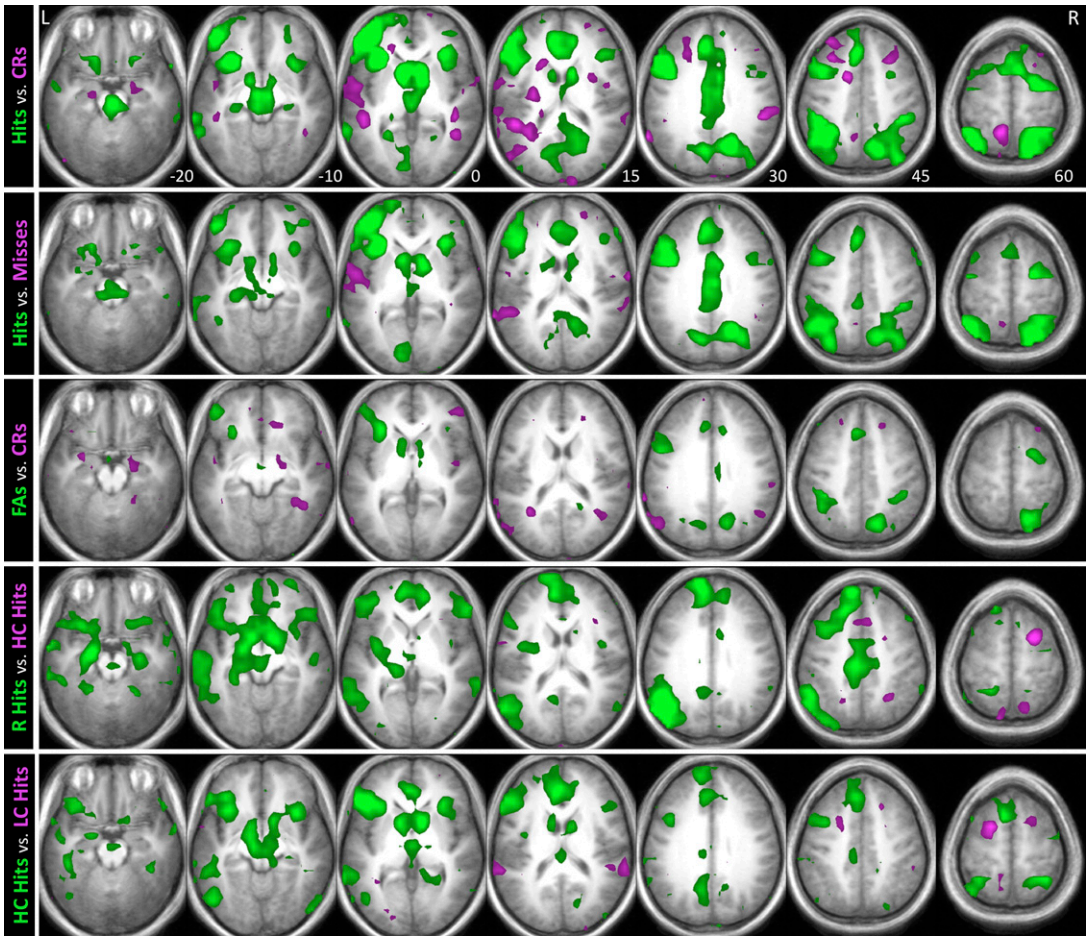

Fig. S2. Classification importance maps (expanded montage of axial slices). For each Exp. 1 classification scheme, group mean importance maps highlight voxels wherein activity increases drive the classifier toward a Class A prediction (green) or Class B prediction (violet). Importance values were arbitrarily thresholded at ±0.0002 and overlaid on axial slices of the mean normalized anatomical image. Coordinates indicate *z* axis position in MNI space.

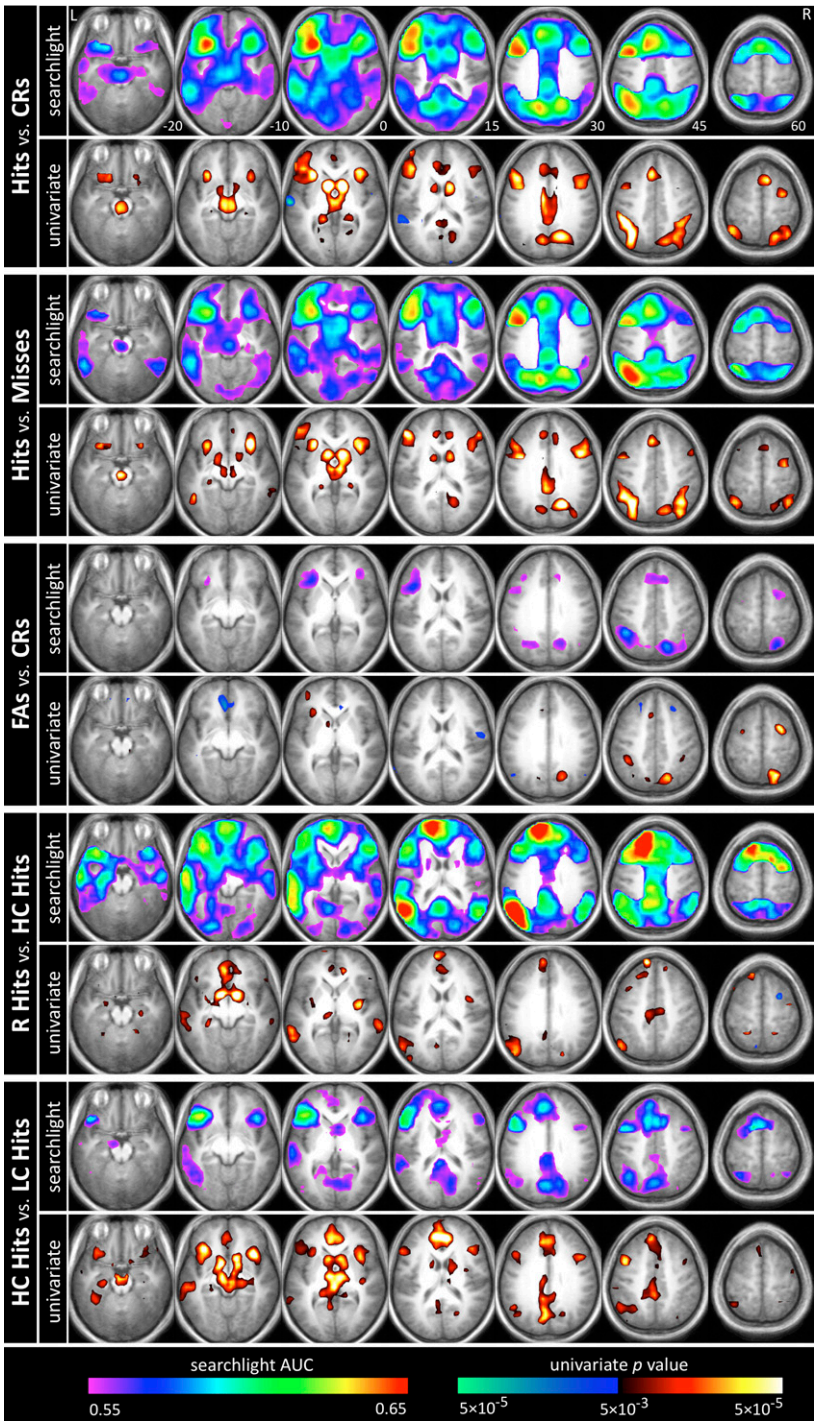

Fig. S3. Spherical searchlight classification maps and corresponding univariate activation contrast maps for Exp. 1 analyses of subjective recognition states. For each classification scheme, a group-averaged map of classifier AUC values from a spherical searchlight analysis are displayed in the upper panel (arbitrarily thresholded at AUC > 0.55), and the corresponding univariate contrast (*t* map) is displayed in the lower panel (thresholded at $P < 0.005$, two-tailed, uncorrected; warm colors indicate Condition A > Condition B and cool colors indicate the inverse pattern). Coordinates indicate *z* axis position in MNI space.

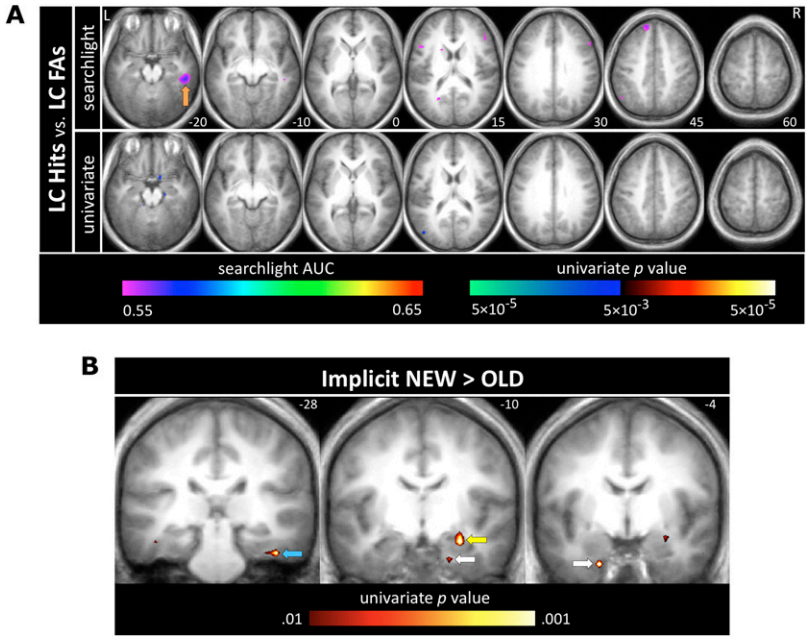

Fig. S4. Neural representations of objective recognition. (*A*) Spherical searchlight classification map and corresponding univariate activation contrast map for the Exp. 1 analysis of objective recognition (LC Hits vs. LC FAs), thresholded as described for Fig. S3. The orange arrow highlights a region of the right fusiform/inferotemporal cortex that showed the most robust performance in the searchlight analysis. (*B*) Repetition suppression effects during the Implicit Recognition Task (Exp. 2). A univariate statistical contrast of activity for NEW items > OLD items revealed repetition suppression effects in right fusiform cortex (blue arrow), right anterior hippocampus extending into the amygdala (yellow arrow), and bilateral regions of perirhinal cortex (white arrows), all regions where such effects are commonly observed for novel vs. repeated visual stimuli. Given the small sample size ($n = 7$), this statistical map is thresholded at a relatively liberal uncorrected threshold of $P < 0.01$ (two-tailed). No regions exhibited suprathreshold activity in the reverse contrast (OLD > NEW). Coordinates indicate *y* axis position in MNI space.

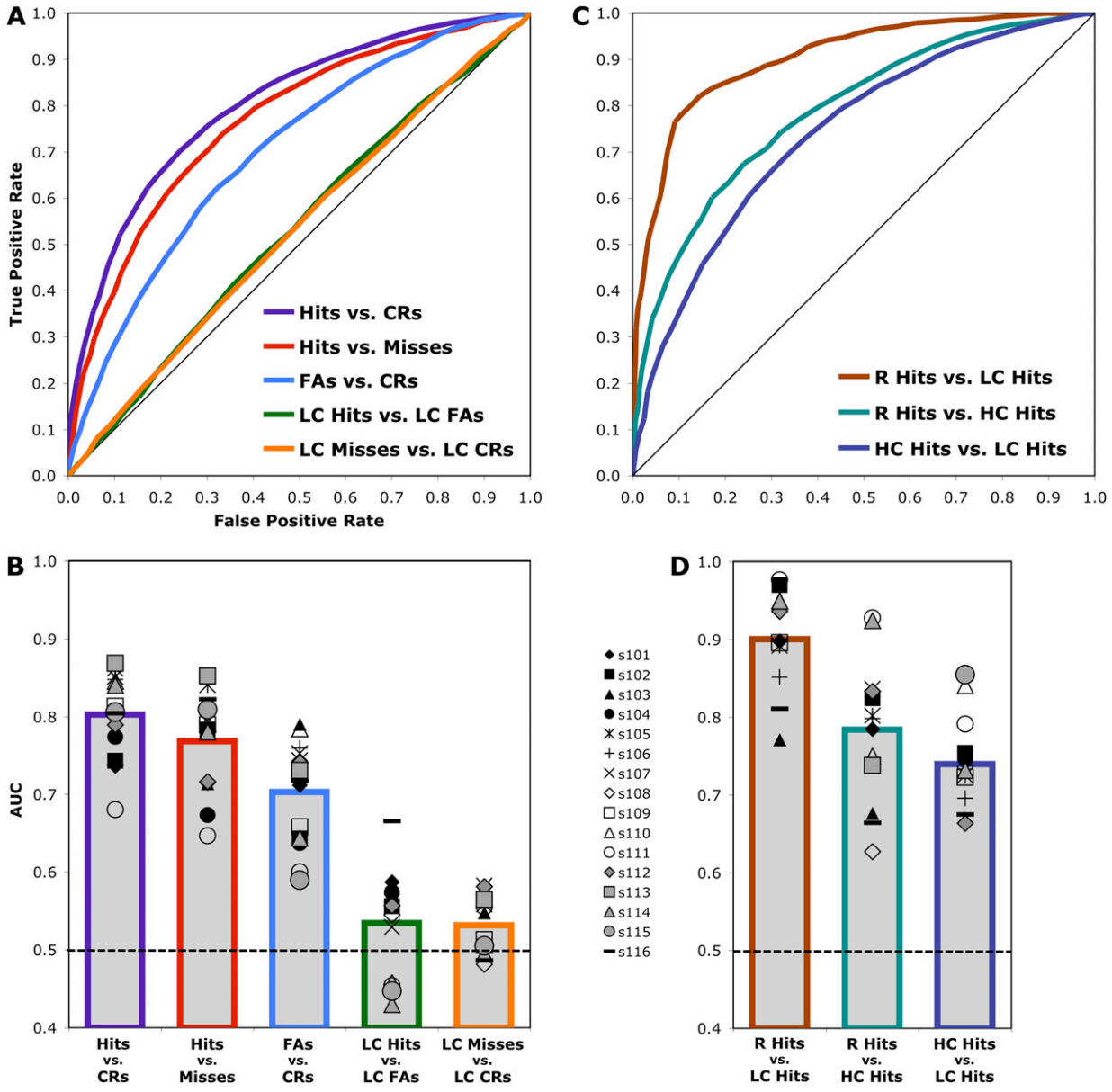

Fig. S5. Across-participant memory decoding. *A–D* are identical in form to those in Fig. 1 *A–D*, with the critical exception being that here classification performance indexes the ability to classify trials from each participant using a classifier that was trained exclusively on the brain data from the other participants. As can be seen by comparing these figures to those in Fig. 1 *A–D*, performance on across-participant classification was generally comparable to that on within-participant classification.

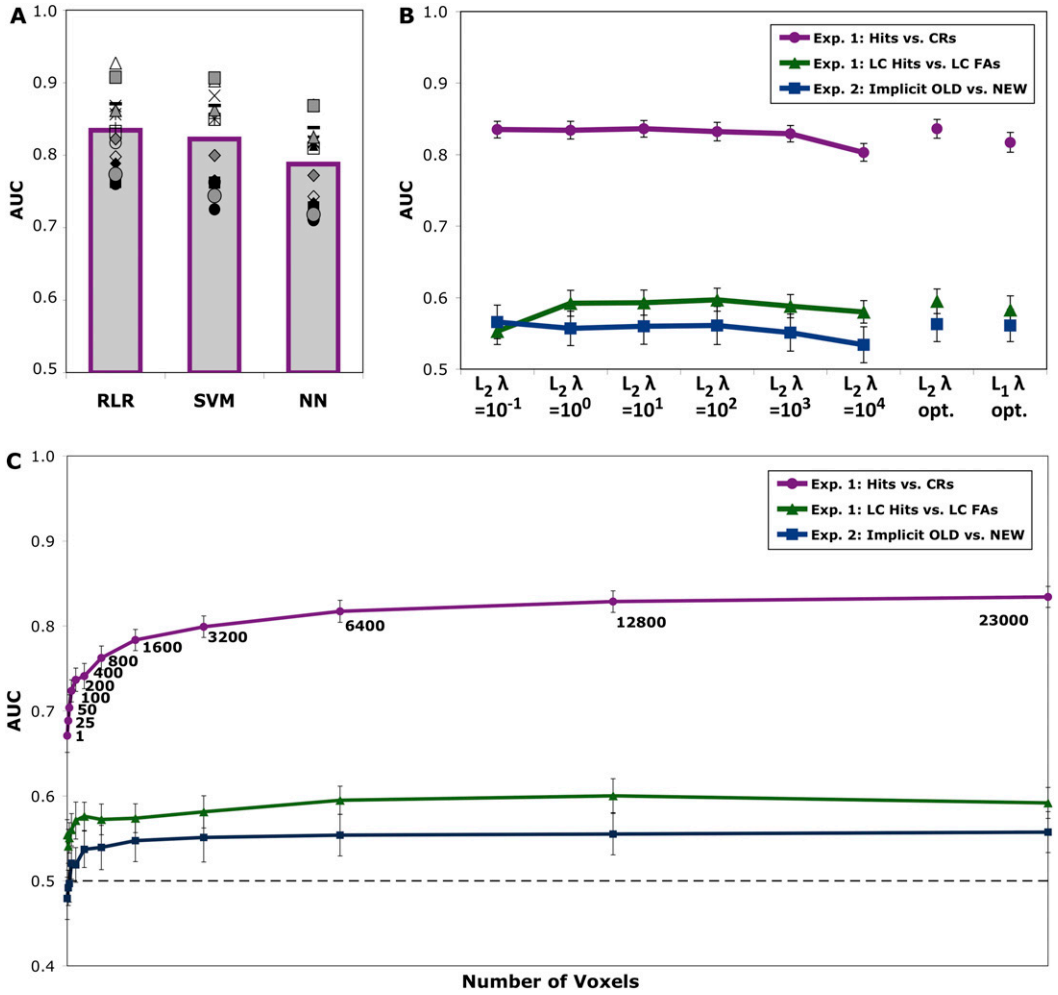

Fig. S6. Effects of classification algorithm, penalty parameter selection, and feature selection on classification performance. (*A*) Comparison of Hits vs. CRs classification performance across three different classification algorithms (RLR = L_2 regularized logistic regression; SVM = linear support vector machine; NN = two-layer back-propagation neural network). (*B*) Effects of penalty parameter settings for RLR classification performance. The linked data points represent classification performance for three distinct classification schemes as a function of the L_2 regularization parameter λ (as the value of this parameter was progressively increased by tenfold steps). For all three classification schemes, performance was relatively constant across all λ settings. The final two data points in each series represent classification performance when the value of the λ parameter was flexibly set for each participant using a nested cross-validation λ optimization routine (opt.). The first of these points used an L_2 regularization approach and the second used an L_1 regularization approach; neither method resulted in substantial performance changes. (*C*) Mean classification performance (AUC) as a function of the number of voxels included in the mask. Voxel inclusion (i.e., feature selection) at each level was determined by selecting the top N voxels whose univariate activity allowed the maximal differentiation of Class A vs. Class B, as assessed by an ANOVA (shown here for three different classification schemes). The numerical labels displayed for the uppermost data series indicate the number of voxels included in the mask, with each successive data point between 25 voxels and 12,800 voxels constituting a doubling of the number of features used for classification (the final data point represents the entirety of the 23,000 voxel mask). To avoid biasing the results, feature selection was done separately for each cross-validation fold. Classification accuracy initially rose rapidly with small increments in the voxel inclusion count, with the rate of improvement then diminishing as larger voxel counts were included. Error bars represent SEM.

Table S1. Recognition memory performance

	Recognition judgment				
	R old	HC old	LC old	LC new	HC new
Exp. 1					
Proportion of responses					
OLD	0.178 (0.126)	0.229 (0.104)	0.295 (0.131)	0.247 (0.107)	0.050 (0.083)
NEW	0.011 (0.019)	0.058 (0.049)	0.229 (0.098)	0.519 (0.161)	0.183 (0.155)
Reaction time					
OLD	1804 (410)	2138 (411)	2313 (404)	2219 (439)	n/a
NEW	n/a	n/a	2324 (408)	2143 (411)	1909 (497)
Exp. 2					
Proportion of responses					
OLD	0.152 (0.100)	0.334 (0.083)	0.313 (0.060)	0.163 (0.070)	0.039 (0.057)
NEW	0.004 (0.011)	0.030 (0.028)	0.181 (0.047)	0.527 (0.160)	0.257 (0.155)
Reaction time					
OLD	1729 (444)	1802 (301)	2035 (467)	1888 (401)	n/a
NEW	n/a	n/a	1987 (313)	1893 (422)	1744 (368)

For Exp. 1 and Exp. 2, the upper rows report the mean proportion of objectively OLD and NEW stimuli assigned each of the five subjective recognition judgments, and the lower rows report the mean RTs (ms) for each bin ("n/a" indicates that for most participants there were insufficient trials to reliably index RT). Exp. 2 data are exclusively from the Explicit Recognition Task. SDs are in parentheses.

Table S2. Numbers of trials (per class) contributing to each classification scheme in Exp. 1

	Hits vs. CRs	Hits vs. Misses	FAs vs. CRs	LC Hits vs. LC FAs	LC Misses vs. LC CRs	R Hits vs. LC Hits	R Hits vs. HC Hits	HC Hits vs. LC Hits
s101	88	n/a	62	38	n/a	38	34	34
s102	95	21	95	64	n/a	38	38	64
s103	127	47	29	24	44	42	39	39
s104	113	87	60	55	76	n/a	n/a	n/a
s105	125	70	32	n/a	37	24	24	24
s106	133	64	67	52	60	39	39	39
s107	131	68	56	54	65	19	19	25
s108	101	98	60	n/a	33	n/a	36	n/a
s109	118	72	72	63	48	n/a	n/a	42
s110	115	85	23	19	67	25	28	25
s111	105	33	93	71	29	19	19	72
s112	137	62	38	38	60	21	21	40
s113	139	32	60	n/a	32	n/a	59	n/a
s114	129	44	70	67	41	18	18	42
s115	93	86	33	29	83	n/a	n/a	37
s116	137	54	28	22	52	32	32	52
Mean	117.9	61.5	54.9	45.8	51.9	28.6	31.2	41.2
N	16	15	16	13	14	11	13	13

These values represent the total number of trials per class (training set + testing set) for each participant after the removal of outlier trials (i.e., excessive motion or global signal) and after artificially balancing the number of trials in each class. Values coded as "n/a" indicate that fewer than 18 trials from each class were available for classification, and thus the participant's data were not analyzed for that particular classification scheme. Summary statistics at bottom indicate mean number of trials per class and the number of participants (*N*) included in each classification analysis.

Table S3. Classification performance levels within individual anatomical ROIs

ROI name	No. vox	Hits vs. CRs	Hits vs. Misses	FAs vs. CRs	R Hits vs. HC Hits	HC Hits vs. LC Hits	LC Hits vs. LC FAs
L MFG	609	0.74 (4)	0.69 (1)	0.61 (8)	0.73 (2)	0.66 (3)	0.58 (2)
L SFG (lateral)	406	0.73 (5)	0.68 (3)	0.63 (4)	0.73 (4)	0.63 (14)	0.57 (7)
L SFG (medial)	356	0.74 (3)	0.68 (7)	0.62 (5)	0.72 (7)	0.65 (9)	0.57 (13)
L precuneus	384	0.71 (10)	0.68 (4)	0.62 (6)	0.73 (5)	0.65 (7)	0.56 (21)
L angular gyus	146	0.72 (7)	0.67 (9)	0.61 (13)	0.70 (15)	0.63 (17)	0.58 (4)
R SFG (lateral)	436	0.72 (6)	0.66 (12)	0.61 (10)	0.72 (8)	0.64 (12)	0.56 (22)
L inferior parietal	272	0.75 (2)	0.69 (2)	0.64 (3)	0.71 (12)	0.67 (2)	0.54 (54)
L superior parietal	210	0.71 (11)	0.68 (6)	0.62 (7)	0.71 (13)	0.68 (1)	0.54 (53)
R MFG	630	0.76 (1)	0.68 (5)	0.61 (9)	0.72 (9)	0.66 (4)	0.53 (65)
L mid-cingulate gyrus	266	0.69 (18)	0.67 (11)	0.59 (20)	0.73 (6)	0.65 (8)	0.55 (32)
R inferior parietal	144	0.69 (20)	0.68 (8)	0.60 (14)	0.74 (1)	0.65 (6)	0.54 (47)
L IFG (pars triangularis)	297	0.71 (12)	0.66 (16)	0.58 (26)	0.68 (21)	0.65 (10)	0.57 (14)
L IFG (pars opercularis)	131	0.70 (14)	0.66 (15)	0.59 (19)	0.66 (27)	0.63 (15)	0.57 (9)
R precuneus	373	0.71 (9)	0.67 (10)	0.64 (2)	0.71 (14)	0.62 (23)	0.54 (51)
L IFG (pars orbitalis)	203	0.70 (15)	0.64 (24)	0.58 (24)	0.68 (23)	0.65 (5)	0.56 (19)
R IFG (pars triangularis)	258	0.72 (8)	0.63 (25)	0.61 (11)	0.66 (28)	0.62 (21)	0.56 (18)
L mid-temporal gyrus	609	0.66 (30)	0.63 (28)	0.58 (23)	0.71 (11)	0.62 (19)	0.58 (5)
R superior parietal	224	0.69 (19)	0.66 (13)	0.66 (1)	0.70 (17)	0.64 (11)	0.53 (69)
L mid-occipital	415	0.66 (29)	0.63 (26)	0.58 (30)	0.73 (3)	0.60 (34)	0.57 (10)
R supramarginal gyrus	190	0.66 (33)	0.64 (19)	0.59 (17)	0.71 (10)	0.63 (16)	0.54 (39)
R IFG (pars opercularis)	161	0.67 (24)	0.59 (55)	0.59 (22)	0.66 (26)	0.64 (13)	0.57 (6)
L post-cingulate gyrus	60	0.66 (32)	0.64 (23)	0.57 (32)	0.63 (41)	0.59 (42)	0.59 (1)
R cuneus	184	0.68 (21)	0.64 (18)	0.59 (21)	0.60 (57)	0.58 (50)	0.57 (8)
R fusiform gyrus	311	0.61 (60)	0.57 (70)	0.56 (48)	0.60 (55)	0.57 (62)	0.58 (3)

Separate classification analyses were run using the voxels within each of 80 ROIs selected from the AAL library. For each of six classification schemes, mean classification performance levels (AUC; displayed as the decimal value in each cell) based on the 80 ROIs were ranked (ordinal rankings displayed in parentheses). The top 10 performing ROIs for each classification scheme are included in this table, resulting in a total of 25 ROIs (ordered by mean rank across the six schemes). For the LC Misses vs. LC CRs classification, no ROIs exhibited AUC performance levels above 0.54, and thus those data are excluded from this table. L, left; R, right; SFG, superior frontal gyrus; MFG, middle frontal gyrus; IFG, inferior frontal gyrus.