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

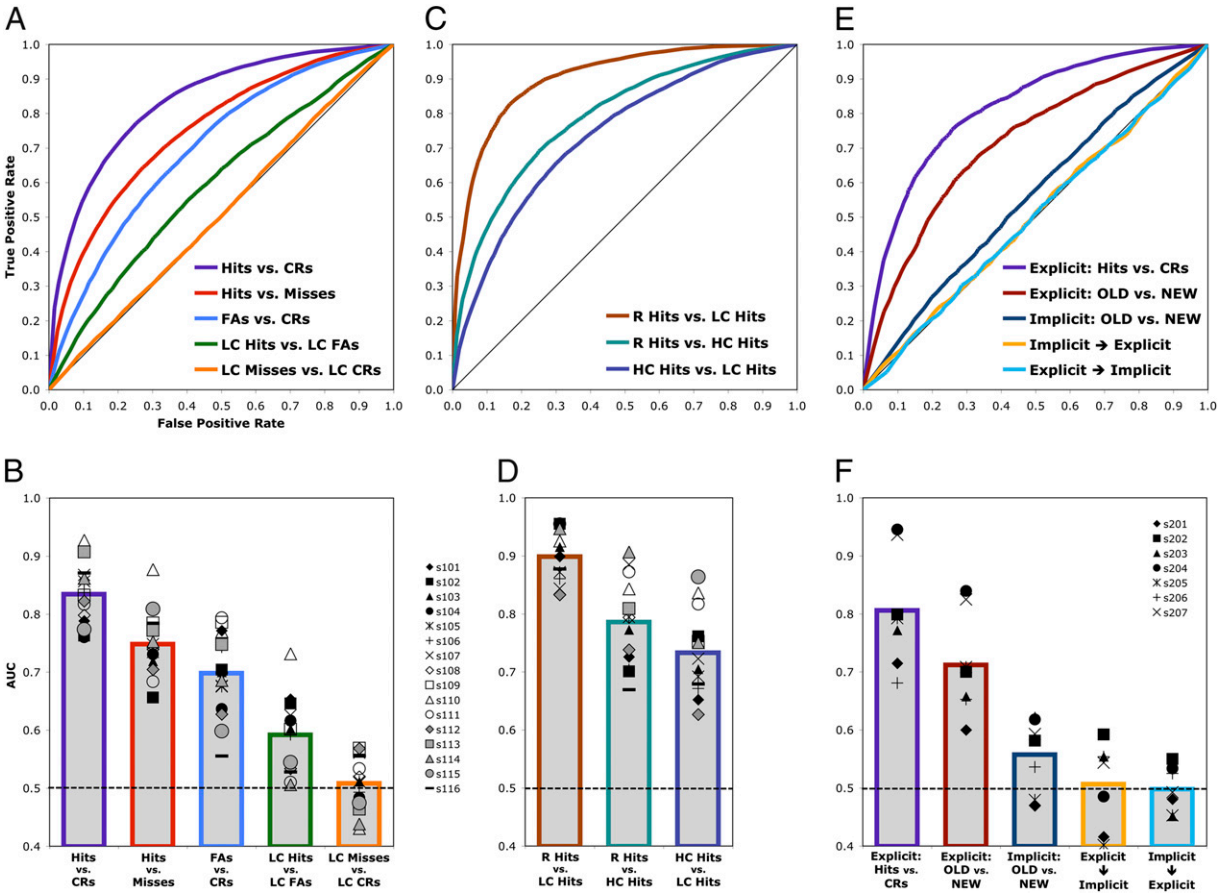

Fig. 1. Mnemonic decoding results. Mean ROC curves (*A*, *C*, and *E*) and their corresponding AUC values (*B*, *D*, and *F*) summarize classifier performance for various classification schemes in Exp. 1 (*A*–*D*) and Exp. 2 (*E* and *F*). AUC values are plotted for each participant's data using unique 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-

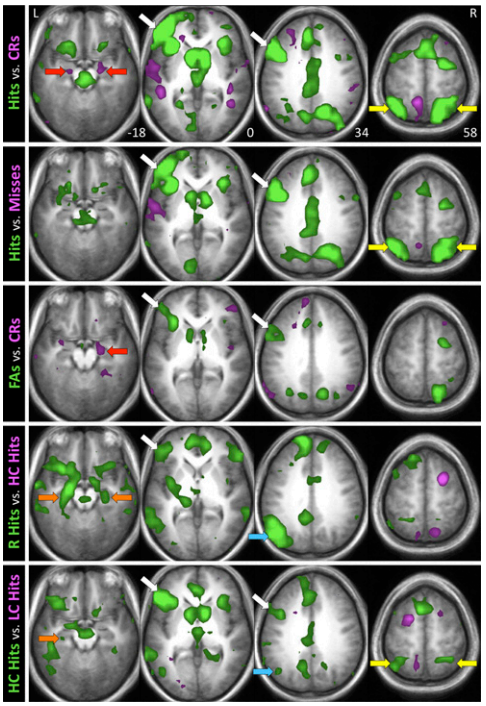

Fig. 2. Classification importance maps. For each classification scheme, group mean importance maps highlight voxels wherein activity increases drive the classifier toward a Class A prediction (green) or Class B prediction (violet). Importance values were arbitrarily thresholded at ±0.0002 and overlaid 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.

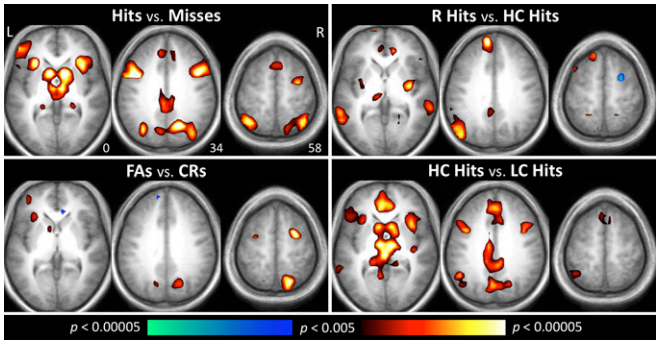

Fig. 3. Univariate contrast maps. Group *t* tests on activity parameter estimates (derived from a standard voxel-level general linear model-based analysis) illustrate regions with greater activity for trials from Class A (warm colors) or Class B (cool colors). The general correspondence between these 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.