

Decoding Near-Threshold Perception of Fear from Distributed Single-Trial Brain Activation

Luiz Pessoa and Srikanth Padmala

Department of Psychology, Brown University,
Providence, RI, USA

Instead of contrasting functional magnetic resonance imaging (fMRI) signals associated with 2 conditions, as customarily done in neuroimaging, we reversed the direction of analysis and probed whether brain signals could be used to “predict” perceptual states. We probed the neural correlates of perceptual decisions by “decoding” brain states during near-threshold fear detection. Decoding was attempted by using support vector machines and other related techniques. Although previous decoding studies have employed relatively “blocked” data, our objective was to probe how the “moment-to-moment” fluctuation in fMRI signals across a population of voxels reflected the participant’s perceptual decision. Accuracy increased from when 1 region was considered (~64%) to when 10 regions were used (~78%). When the best classifications per subject were averaged, accuracy levels ranged between 74% and 86% correct. An information theoretic analysis revealed that the information carried by pairs of regions reliably exceeded the sum of the information carried by individual regions, suggesting that information was combined “synergistically” across regions. Our results indicate that the representation of behavioral choice is “distributed” across several brain regions. Such distributed encoding may help prepare the organism to appropriately handle emotional stimuli and regulate the associated emotional response upon the conscious decision that a fearful face is present. In addition, the results show that challenging brain states can be decoded with high accuracy even when “single-trial” data are employed and suggest that multivariate analysis strategies have considerable potential in helping to elucidate the neural correlates of visual awareness and the encoding of perceptual decisions.

Keywords: awareness, decision making, emotion, fear, fMRI

Introduction

Neuroimaging studies typically adopt a subtractive methodology to determine brain regions engaged by specific perceptual, motor, or cognitive conditions. For instance, in a particularly successful example, the contrast of viewing faces versus viewing nonface objects has been used to reveal the neural substrates of face perception. Although the interpretation of the related findings has generated heated debate, the general subtractive strategy has been largely the same. In the present study, however, we took a different approach. Instead of contrasting the functional magnetic resonance imaging (fMRI) signals associated with 2 conditions, we reversed the direction of the analysis and probed whether brain signals could be used to “predict” perceptual states (Fig. 1A), much as proposed by Haxby and others (2001) and further developed by others (Cox and Savoy 2003; Hanson and others 2004; Mitchell and others 2004; Haynes and Rees 2005; Kamitani and Tong 2005; Mourao-Miranda and others 2005; O’Toole and others 2005).

The perception of emotion-laden visual stimuli engages a network of brain regions (Haxby and others 2000; Adolphs 2002; Pessoa and others 2002), including the fusiform gyrus and superior temporal sulcus in visual cortex, as well as regions more directly involved in affective processing per se, such as the amygdala, insula, and anterior cingulate cortex, among others. Previous research has largely focused on how the physical properties of the stimuli affect brain responses (e.g., contrasting responses evoked by fearful and happy faces). The goal of the present study was to probe the neural correlates of perceptual decisions during near-threshold fear detection. Participants performed a difficult fear-detection task in which an initial target face was briefly presented for 67 or 33 ms and immediately followed by a neutral-face mask. In each trial, subjects indicated whether or not they perceived a fearful face. We investigated how “moment-to-moment” fluctuations in fMRI signals were correlated with “behavioral choice,” namely, whether a subject reported “fear present” or “fear absent” on a given trial. In particular, we evaluated the extent to which signals from multiple brain regions would provide a better prediction of choice than a single region. We reasoned that if the representation of the perceptual decision was localized, signals from one brain region should predict behavioral response no worse than signals from multiple regions. Alternatively, if multiple regions predicted decisions better than a single one, our results would favor the interpretation that the representation of the decision is distributed across the brain. We investigated the above questions by using machine learning techniques. Although most previous “brain-decoding” studies have employed relatively “blocked” data (Haxby and others 2001; Cox and Savoy 2003), our objective was to probe how the “trial-by-trial” fluctuation in fMRI signals across a population of voxels from one or multiple brain regions reflected the participant’s perceptual decision (Fig. 1B). In this manner, we probed whether “single-trial” fMRI data would support robust prediction of difficult behavioral decisions and sought to establish how multivariate analysis strategies could be employed to provide insights about the encoding of perceptual decisions at conditions near the threshold of visual awareness. Overall, we believe that investigating how distributed patterns of activation across multiple brain regions are linked to moment-to-moment fluctuations in “perception” offers the potential to go beyond the type of information that is obtained with more standard univariate, subtractive techniques (Haxby and others 2001).

Materials and Methods

Subjects

Nine volunteers (6 women) aged 23 (mean) \pm 8 (standard deviation) years participated in the study, which was approved by the Institutional

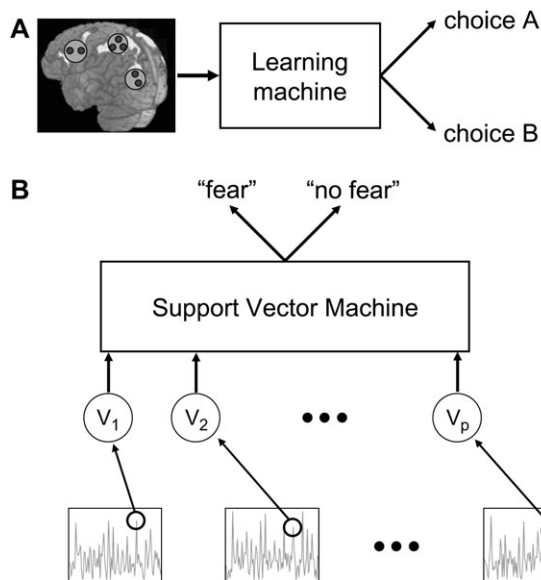


Figure 1. Predicting perceptual decisions from distributed single-trial activation. (A) fMRI responses from voxels (indicated by dots) from multiple regions (indicated by the larger circles) were used by a machine learning algorithm to predict behavioral choice (reported “fear” or “no fear”) during a near-threshold fear-detection task. (B) The input to the SVM comprised “single-trial” responses from multiple voxels from one or multiple regions. Thus, the SVM learned an input-output mapping from the input space of fMRI voxel responses to the output space of behavioral choices. The plots at the bottom show simulated time series from individual voxels V_i . Given the distributed pattern of activation at time t (see circles), after learning, the algorithm predicted the perceptual decision.

Review Board of both Brown University and Memorial Hospital of Rhode Island. All subjects were in good health with no past history of psychiatric or neurological disease and gave informed consent. Subjects had normal or corrected-to-normal vision.

Stimuli

Face stimuli were obtained from the Ekman set (Ekman and Friesen 1976), a set recently developed by Ohman and others (KDEF, Lundqvist D, Flykt A, and Ohman A; Karolinska Hospital, Stockholm, Sweden), as well as a set developed and validated by Ishai and others (2004) at the National Institute of Mental Health (Bethesda, MD). Forty instances of identity-matched fearful, happy, and neutral faces were employed.

Stimuli and Procedure

Each trial began with a white fixation cross shown for 1000 ms on a black background, followed by a green fixation shown for 300 ms on a black background, followed by the presentation of a fearful, happy, or neutral “target” face, and immediately followed by a neutral face, which served as a “mask” (Fig. 2). The identities of the target and mask stimuli were always different. Faces subtended 4 degrees of visual angle. Two target durations were employed: 67 and 33 ms. Mask faces were shown such that the target plus mask stimuli always lasted 133 ms (i.e., 66 and 100 ms, respectively). Target presentation durations were confirmed by employing a photodiode and an oscilloscope. Subjects were instructed that the stimulus would always comprise 2 faces and to respond “fear” if they perceived fear, however, briefly. Following the presentation of each face pair, subjects indicated “fear” or “no fear” with a button press. On each trial, subjects also rated the confidence in their response on a scale of 1–4 (low to high confidence). The total trial duration was approximately 12 s (11920 ms). Each subject performed 320 trials. In this paper, we mostly focused our analysis on fearful-neutral and neutral-neutral target-mask trials; additional analyses of happy-neutral trials were investigated elsewhere (Pessoa and Padmala 2005). Although in our previous report only 67-ms trials were probed, in the present study, we investigated both 67- and 33-ms trials. Moreover, because of the importance of the amygdala in the processing of fear, here we also investigated fMRI responses evoked by this

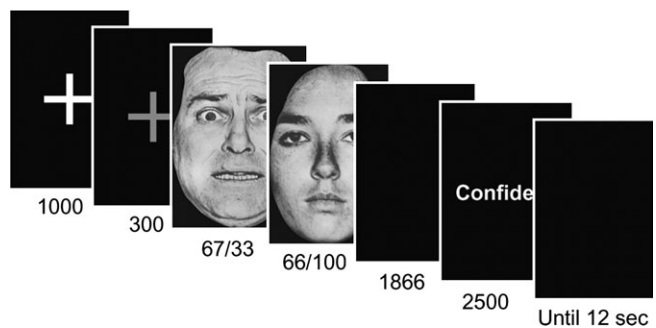


Figure 2. Experimental paradigm. For every trial, after the target-mask face pair, subjects indicated whether they saw a fearful face or not and then indicated their confidence in the response. The initial target face was a fearful, happy, or neutral face, and the mask was always a neutral face. Target stimuli were shown for 67 or 33 ms, and masks were shown for 66 or 100 ms, respectively (such that the target plus mask duration totaled 133 ms). Trials occurred every 12 s in a slow event-related design. The 300-ms fixation cross was actually green in the experiment. All durations are in milliseconds, unless noted.

region. Finally, the predictive power of both “choice-related” regions and “stimulus-responsive” regions (see below) was studied.

fMRI Data Acquisition and Analysis

fMRI data were collected using a Siemens 1.5-T scanner. Each scanning session began with the acquisition of a high-resolution magnetization prepared rapid gradient echo anatomical sequence (time repetition [TR] = 1900 ms, echo time [TE] = 4.15 ms, time to inversion = 1100 ms, 1-mm isotropic voxels, 256-mm field of view). Each subject performed 7–8 experimental runs, each lasting 8 min and 10 s. During each functional scan, 162 gradient-echo echo-planar volumes were acquired with a TE of 38 ms and TR of 2980 ms. Each volume consisted of 37 axial slices with slice thickness of 3 mm and in-plane resolution of 3×3 mm.

Regions of Interest

In a recent study, we investigated how the trial-by-trial variability in fMRI response could be used to predict behavioral choice, namely, the subject’s decision as to whether a fearful face had appeared or not (Pessoa and Padmala 2005). We identified 5 key “choice-related” brain regions that had strong predictive power and were not driven by stimulus properties (i.e., fear-containing vs. neutral trials) or behavioral performance (i.e., correct vs. incorrect trials): posterior cingulate cortex (PCC) ($x = 0, y = -32, z = 32$), medial prefrontal cortex (MPFC) ($x = -2, y = 50, z = 12$), right inferior frontal gyrus (IFG) at both more posterior ($x = 44, y = 19, z = -10$) and more anterior sites ($x = 44, y = 46, z = -4$), and left insula ($x = -60, y = 2, z = 4$); all coordinates follow the Montreal Neurological Institute convention. In the present paper, in several of our analyses, we utilized these regions of interest (ROIs) in decoding brain states from distributed patterns of activation given that one of our goals was to test the idea that the representation of the behavioral choice in our task is better described as distributed, instead of localized. However, we also considered additional brain regions that predicted behavioral performance and exhibited stimulus-driven responses to further probe the issue of decoding and the representation of behavioral choice (here called stimulus-responsive regions): left/right fusiform gyrus (left: $x = -37, y = -51, z = -13$; right: $x = 44, y = -50, z = -23$), left/right superior temporal sulcus (left: $x = -50, y = -53, z = 14$; right: $x = 56, y = -45, z = 23$), and anterior cingulate cortex ($x = 2, y = 26, z = 42$). Although responses in these regions were not exclusively choice related, like choice-related regions, they also predicted behavioral choice (Pessoa and Padmala 2005). Voxels from the above ROIs were employed in our decoding analyses, which considered only voxels with significant choice probabilities (CPs) (see below). Given this selection criterion, on average (across all ROIs and subjects) 9.7 voxels were used per ROI.

Classification with Support Vector Machines

For classification, we employed linear support vector machines (SVMs) (Boser and others 1992; Vapnik 1995; Burgas 1998), as implemented

in the Ohio State University SVM toolbox, which is based on the LIBSVM package (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>). The penalty/cost parameter C was 1.0. Inputs to the SVM consisted of the fMRI response strengths for individual trials from one or more voxels, which were classified in terms of the binary output class “fear” and “no fear” corresponding to the behavioral choice. Response strength was indexed by the average of the raw fMRI signal (after linear detrending) at times 3, 6, and 9 s relative to stimulus onset.

To estimate classification rates, a variant of the standard leave-one-out cross-validation scheme was employed. Specifically, the SVM was trained on all the data except for 2 trials (1 from each class), and prediction was attempted for those trials only (Mitchell and others 2004). Moreover, because classification was based on behavioral responses that could in principle differ in overall numbers (e.g., 40 “fear” responses and 30 “no fear” responses), during cross validation, we maintained the 2 sets balanced (in this case, 30 trials for each class) to guarantee that chance classification levels were 50% correct. Overall, we bootstrapped (Efron and Tibshirani 1993) our classification rates by averaging 400 random-trial samplings—for each one, we computed our leave-one-from-each-class-out cross validation.

Two types of classification were investigated: single region and multiple region. For the 2 types, the starting point was to consider voxels from the choice-related ROIs, stimulus-responsive ROIs, or the complete set of 10 ROIs. For all analyses, we only employed voxels that significantly predicted behavioral choice as assessed via “CPs” (Pessoa and Padmala 2005), with the exception of control analyses, as indicated in the text. Briefly, this method gives the probability that a so-called “ideal observer,” given only access to the fMRI amplitude in a trial, would be able to accurately identify which behavioral response was made in that particular trial (“fear” or “no fear” response). Whereas in monkey physiology research, spike data are used as a measure of response (Britten and others 1996; Dodd and others 2001; Grunewald and others 2002), in the present case, fMRI amplitude was used as an index of response strength. For present purposes, CP values were used simply to 1) assess whether individual voxels significantly predicted behavioral choice and 2) rank voxels from most to least predictive.

Given our voxel selection criterion, we expected classifications to be better than 50% (chance). The questions that we addressed were 1) whether classification rates would benefit from the number of voxels considered as the basis for classification (i.e., the dimensionality of the input vectors), 2) whether classification rates would benefit from considering voxels from more than one region, and 3) how choice-related and stimulus-responsive ROIs contributed to classification. For single-region classification, for every region, one or more voxels were used for classification. To provide a strict test of potential improvements in classification, voxels were initially ordered in terms of the strength of choice-related activation as indexed by CP values—that is, the higher the CP value, the greater the probability that an ideal observer would be able to predict the behavioral response based on the observation

of fMRI responses from that voxel. Classification based on the highest CP voxel was then attempted. Additional voxels were then considered in “decreasing” order of CP. Such analysis was performed for all regions, and the number of voxels associated with the highest classification rate was determined (labeled “maximum” in Fig. 3A).

We also investigated how classification accuracy improved as a function of the number of regions considered and the number of voxels per region. For example, for choice-related ROIs, for each subject, all possible 2-, 3-, 4-, and 5-region combinations were tested (10, 10, 5, and 1 combination(s), respectively). In all cases, initially 1 voxel for each region was considered, then 2 voxels, and so on, until all significant voxels were exhausted. Because the number of voxels with significant predictive activation differed from region to region and participant to participant, when the voxels for a given region were exhausted, we continued increasing the number of voxels from other regions, until all available voxels were finished. For all multiple-region combinations, we determined the average number of voxels for each region that led to the highest classification rate (labeled “maximum” in Fig. 3A). As in the case of single-region classification, voxels were considered in decreasing order of CP to provide a strict test of potential improvements in classification accuracy.

For nonlinear SVMs, we employed the following parameters. Polynomial: $C = 1$ (cost), $\gamma = 1$ (gain), $r = 0$ (offset). Radial basis function (RBF): $C = 1$, $\gamma = 1$. Grid searches indicated that the particular parameter choice adopted was not critical. For instance, for the RBF case, we performed a grid search in which $C = \{2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2^0, 2^1, 2^3, 2^5, 2^7\}$ and $\gamma = \{2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2^0, 2^{-1}, 2^{-3}, 2^{-5}, 2^{-7}\}$. Prediction accuracy with nonchoice regions did not differ from the values reported ($C = 1$, $\gamma = 1$) by more than 2% for any of the 81 (9×9) crossed parameter choices above. These results suggest that our findings are representative and not a simple consequence of poor parameter selection. Note, however, that grid searches may miss solutions that require finer grids.

Virtual Lesions

To quantify the “relative” importance of a given region for prediction accuracy, we performed a “lesion” analysis. Our goal was to determine the reduction in classification accuracy when a given region was not available (i.e., “lesioned”) to be combined with the remaining regions in the prediction of behavioral choice. Such analysis was performed by considering the set of choice-related regions plus the left amygdala. The starting point of the analysis was to consider the maximum classification values for 1- through 6-region combinations as assessed via linear SVMs (similar to the results shown in Fig. 3A). Then, for each n -region combination, we removed one specific region from the pool of available regions, determined the classification accuracy obtained without that region, and determined the reduction in classification accuracy with respect to the accuracy obtained when that region was actually available. The final value plotted in Figure 5 was the overall mean

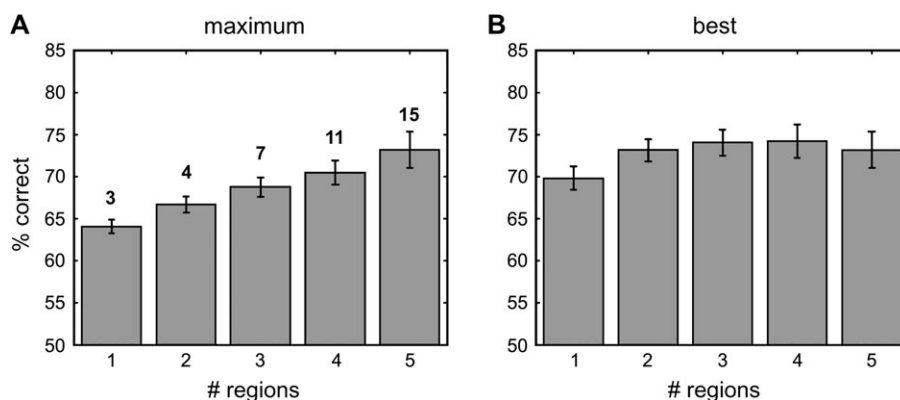


Figure 3. Decoding of brain states linked to “fear” and “no fear” behavioral choices from distributed voxelwise activation with SVMs. (A) Average classification rates increased linearly when multiple voxels from multiple regions were considered, suggesting a distributed representation of behavioral choice. The values over individual bars show the average number of voxels (summed across regions) associated with the maximum classification rate. (B) The potential for decoding was further assessed by considering the “best” classification per participant, for each n -way region combination. Error bars indicate the standard error of the mean.

reduction in prediction accuracy when each region was lesioned (averaged across all region combinations and subjects). For example, suppose the left amygdala was lesioned. The value plotted in Figure 5 (6.7% reduction) considered all the n -region combinations without the amygdala relative to when this region was available to be pooled with the remaining ones. Although other strategies of quantifying the contribution of a region are certainly possible, the present method provides the “relative” reduction in prediction accuracy for each particular region as a starting point in assessing “importance.”

Neural Network Classification

For our tests with neural networks, we employed standard architectures largely inspired by the results of Hanson and others (2004) and implemented them in the Matlab Neural Networks toolbox. Briefly, we employed a feedforward 3-layer network with one output node with logistic activation and cross-entropy error function (the functions `cross_entropy.m` and `cross_entropy_deriv.m` were obtained from the multivoxel pattern analysis toolbox at <http://www.csmbm.princeton.edu/mvpa/>). We tested networks with 5, 10, 15, and 20 hidden nodes (i.e., nodes in the intermediate layer), which used a hyperbolic tangent activation transfer function. Learning (i.e., weight updating) employed the scaled conjugate gradient method (Moller 1993), which is a fast variant of the conjugate gradient method. Training proceeded until the error was less than 0.0001 or 1000 iterations.

Information Theory

We can quantify the relationship between fMRI responses and behavioral choices by treating the voxel as a communication channel (Cover and Thomas 1991) and by measuring the information conveyed by the fMRI response about the decisions made during the fear-detection task. One way to formalize this relationship is in terms of the “mutual information” between choices and fMRI amplitude, which is the reduction in uncertainty of choice given fMRI amplitude:

$$I(C; \text{fMRI}) = H(C) - H(C | \text{fMRI}),$$

where $H(C)$ is the entropy of choice and $H(C | \text{fMRI})$ is the conditional entropy of choice given fMRI signals. Shannon entropy is a measure of uncertainty (or “average surprise”) of a random variable and is defined by

$$H(X) \equiv - \sum_x P(x) \log_2 P(x).$$

Conditional entropy measures the uncertainty associated with a conditional probability (Ash 1965) and can be written as

$$H(Y|X) = - \sum_i P(x_i) \sum_j P(y_j|x_i) \log P(y_j|x_i).$$

In the case of interest, Y refers to behavioral choice and X refers to fMRI amplitude. The choice index is discrete and assumes only 2 values in the present task (“fear” and “no fear”). fMRI amplitude indexes response strength, as defined previously. Thus, computing $I(C; \text{fMRI})$ requires estimating simple and conditional probabilities. Because employing limited data samples may result in biased estimates, it is important to employ bias correction procedures developed for neuronal data (Golomb and others 1997), which have been shown to yield adequate results for low-dimensional cases (e.g., 1- and 2-dimensional codes, as when 1 or 2 voxel time series are considered).

To investigate how signals from one voxel are potentially combined with those from another voxel so as to allow better prediction accuracy, we computed the “joint information” carried by a pair of voxels together and compared that with the transmitted information calculated for each voxel separately. We defined an index

$$J = \frac{I_{AB}}{I_A + I_B},$$

where I_A and I_B are the information about the behavioral choice carried by voxels A and B and I_{AB} is the information carried by both voxels “jointly.” This index was previously employed by Richmond and others to investigate neural coding in inferior temporal cortex in monkeys (Gawne and others 1996). Thus, J should be greater than 1 if the in-

formation carried by the 2 voxels considered together is greater than the sum of the information carried out by each voxel alone, 1 if the 2 voxels are independent, and less than 1 if there is redundant information (1/2 if the 2 voxels carry identical information). Although the index J was described above for voxels, it can also be used to investigate joint information from 2 regions; in this case, a representative time series would be used for each ROI. Note that in the present case, information cannot exceed 1 bit given that behavioral choice is a binary variable. Thus, if each region transmitted, for example, 0.6 bits about the subject’s choice, the joint information index would always indicate redundancy (given that $I_A + I_B$ would be greater than 1 and that I_{AB} cannot exceed 1). In such cases, J would underestimate the potential for synergistic interactions. In the present study, however, this was not a major concern because information carried by an individual region did not exceed 0.5 bits.

Results

In a recent study, we showed that fMRI signals of several brain regions predicted behavioral choice during near-threshold fear perception (Pessoa and Padmala 2005). Two types of regions were observed: regions in which voxels predicted choice and were driven by stimulus differences (i.e., fear-containing trials evoked stronger responses relative to neutral trials) and regions in which voxels predicted choice, but were not stimulus driven. We will refer to the latter group of regions as “choice related,” as they may be more directly related to the representation of behavioral choice per se; this group consisted of 5 ROIs: PCC, MPFC, right IFG (at both posterior and anterior sites), and left anterior insula (for coordinates, see Materials and Methods). The former group of stimulus-responsive regions also comprised 5 ROIs: left/right fusiform gyrus, left/right superior temporal sulcus, and anterior cingulate cortex. Overall, these regions reliably predicted the perceptual decision when individual voxels were considered.

In the analyses below, we attempted to decode brain states separately considering the set of choice-related ROIs or the set of stimulus-responsive ROIs, as well as when all 10 ROIs were considered together. Furthermore, we investigated decoding both for trials containing 67- and 33-ms target faces. Analyses utilized SVMs (Boser and others 1992; Vapnik 1995; Burges 1998), which have been extensively used in recent years and have been successful in a range of machine learning and classification applications, including genetic (Guyon and others 2002) and neuroimaging data (Mitchell and others 2004; Mourao-Miranda and others 2005).

Decoding Choice with SVMs: 67-ms Target Faces

Participants performed the task with 81% correct accuracy, indicating that although the task was relatively difficult, they could reliably detect 67-ms target fearful faces—the average nonparametric sensitivity measure A' was 0.89 (a value of 0.5 indicates change performance, and a value of 1.0 indicates perfect sensitivity; see Macmillan and Creelman 1991). We tested whether the distributed trial-to-trial variability in fMRI magnitude could be used to robustly predict the perceptual decision made on each trial by the subject (report of “fear” or “no fear”). As a strong test of the potential improvements obtained by adding voxels from one or multiple regions, we employed voxels in descending order of predictability (see Materials and Methods). Thus, the improvements in classification accuracy that we observed were not simply due to having initially considered poorly predictive voxels and then employing more predictive ones. Instead, improvements in predictability when

considering voxels whose individual time series were less predictive when they were considered individually revealed how they contributed to classification when the entire multivariate activation pattern was considered.

Initially, we performed an analysis to evaluate whether classification based on a single region would improve as a function of the number of voxels utilized. The average single-region classification rate across the 5 choice-related ROIs was 64.1% when on average 3 voxels were employed from a region. Next, we considered classification rates when “multiple” voxels from 2, 3, 4, or 5 regions were considered. Maximum multiple-region classification rates (Fig. 3) were 66.7%, 68.7%, 70.5%, and 73.2% correct, respectively (averaged over all possible multiple-region combinations and subjects). A trend analysis revealed a significant linear trend ($P < 0.0001$) as a function of the number of ROIs, revealing that although relatively modest average increases were observed when a single region was added, extra regions provided further information for classification. In addition, it should be pointed out that classification rates for “fear present” and “fear absent” reports were very similar, with differences of less than 2.5% in all cases. Thus, the values given above are representative of both types of perceptual decision.

The above results revealed that classification rates improved when additional regions were considered. In fact, planned comparisons revealed that, in all cases, average classification improved when utilizing n regions instead of $n - 1$ regions (all P values < 0.05). It is important to note, however, that increases in classification were not just due to increases in the total number of voxels employed in classifications involving additional regions. The average number of voxels (summed across regions) when classification reached the maximum value was 3 for single-region classification, 4 for 2-region classification, 7 for 3-region classification, 11 for 4-region classification, and 15 for 5-region classification (values correspond to averages across subjects). Critically, classifications when considering fewer regions did not improve when additional voxels were utilized (from those regions). In addition, note that improvements in classification accuracy were not due to the fact that considering additional regions entailed using voxels that were potentially more individually discriminative. The average CPs for the voxels from 1- through 5-region combinations leading to the results shown in Figure 3A were nearly indistinguishable (66.6, 67.1, 67.1, 67.1, and 67.6, respectively).

The classification rates shown in Figure 3A were based on averages of many individual values (all n -way region combinations, for all participants). As illustrated for the case of 3-region combinations, there was considerable variation in classification accuracy (10 possible combinations \times 9 subjects = 90 values). Thus, the overall average classification rate may have underestimated the potential for decoding perceptual decisions from distributed fMRI data. To further probe this issue, we determined the best classification rate per subject for all n -way region groupings and averaged the results across subjects (we called such summary statistic “best”). In this case, best classification accuracy exceeded 70% in all cases, except for 1 region, and peaked for 4-region combinations at 74.2% (Fig. 3B).

To further probe the ability to decode brain states from distributed voxelwise patterns of activation, we determined classification rates when voxels were selected from the set of 5 stimulus-responsive ROIs. Interestingly, classification rates when considering stimulus-responsive regions were nearly identical to the ones observed when we employed choice-

related regions: 65.1%, 67.5%, 70.0%, 71.7%, and 73.5% correct, for 1 through 5 regions, respectively (significant linear trend, $P < 0.0001$); the best classification peaked for 3-region combinations at 75.9%. These results suggest that such regions may be involved not only in encoding stimulus-based features but also in the representation of the perceptual decision.

We then investigated whether all 10 regions combined would improve classification accuracy. Classification rates improved as a function of the number of ROIs (significant linear trend, $P < 0.0001$) and peaked at 77.8% when all 10 regions were considered (Fig. 4). As in the results shown in Figure 3, classification rates were based on averages of many individual values (e.g., for 5 ROIs: 252 combinations \times 9 subjects = 2268 values). In this case, the best classification rate consistently exceeded 80% correct (average 81%) and peaked for 5 regions at 85.6%. In addition, several individual classification rates exceeded 90% correct (3/9 subjects), further demonstrating the method’s ability to decode brain activation with high accuracy. Finally, note that when 10 regions were employed, a near-significant quadratic trend was also observed ($P < 0.08$), indicating that improvements in accuracy tended to “saturate” with increased number of regions.

Decoding Choice with SVMs: 33-ms Target Faces

Thirty-three milliseconds targets were very challenging to detect, and the average performance was 61% correct—the average sensitivity measure A' for 33-ms target fearful faces was 0.68 (again, a value of 0.5 indicates chance performance, and a value of 1.0 indicates perfect sensitivity; see also Discussion). Average classification rates for choice-related regions for 1 through 5 ROIs were, respectively, 64.9%, 68.2%, 70.7%, 72.8%, and 74.1% correct (significant linear trend, $P < 0.005$); values that were very similar to the rates obtained with 67-ms targets. Classification accuracy was a little lower when the set of stimulus-responsive regions was investigated and peaked at 71.0% for the 5-region combination (linear trend not significant). Likewise, classification rates for the complete set

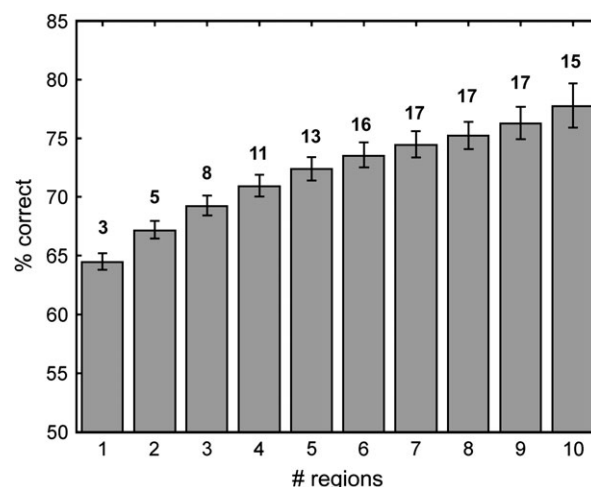


Figure 4. Decoding perceptual decisions when choice-related and stimulus-responsive regions (total of 10 ROIs) were considered together. Classification accuracy increased linearly as a function of the number of regions and approached 80% correct for the 10-region combination. The values over individual bars show the average number of voxels (summed across regions) associated with the maximum classification rate. Error bars indicate the standard error of the mean.

of 10 ROIs were slightly lower than the corresponding values when 67-ms targets were considered and peaked at 75.5%.

Decoding Activation when Happy Faces Were Considered

So far, we have focused our analysis on decoding fMRI signals in terms of “fear” and “no fear” target classes when fear-neutral and neutral-neutral target-mask pairs were considered. In additional analyses, we investigated decoding when fear-neutral and happy-neutral target-mask pairs were considered (see also Pessoa and Padmala 2005). For 67-ms targets, classification rates for 1 through 5 ROIs were 63.3%, 65.4%, 66.7%, 68.9%, and 70.7% correct, when choice-related ROIs were considered (significant linear trend, $P < 0.0001$). Values were similar when stimulus-responsive regions were used and peaked at 73.1% for the 5-region combination (significant linear trend, $P < 0.001$). For 33-ms targets, values were comparable with those obtained with 67-ms targets and peaked at 69.7% for 5 choice-related ROIs and at 67.1% for 3-region combinations when stimulus-responsive regions were considered. These results demonstrate that reliable predictions can be obtained regardless of the type of nontarget stimulus considered (neutral or happy), corroborating the notion that decoding ability reflects the prediction of behavioral choice per se and not other aspects of the data (see also Decoding the Stimulus below).

Decoding and the Amygdala

In the previous analyses, we did not consider the amygdala, a region that has been implicated in emotion in general, and the perception of fearful faces in particular (Aggleton 2000; Adolphs 2002). In our previous study (Pessoa and Padmala 2005), although responses of voxels in the left amygdala predicted behavioral choice, we conservatively did not label this region as “choice related” because, for example, this region did not exhibit significant CPs as consistently across subjects (possibly due to difficulties of scanning this region with fMRI due to susceptibility artifacts). Nevertheless, because of the theoretical significance of the amygdala in emotional perception in general and near-threshold fear detection in particular (Pessoa 2005), we investigated classification accuracy when the left amygdala ($x = -21, y = -1, z = -22$) was included in the set of choice-related ROIs. For this analysis, we considered 6 participants for whom classification accuracy for the (single-region) amygdala averaged 64% correct (this value was chosen to be similar to the 1-region classification rate shown in Fig. 3); for the remaining 3 subjects, average classification was lower (57.3%), possibly due to lower signal-to-noise ratios in these subjects. Classification accuracy increased linearly ($P < 0.0001$) as more regions were considered and peaked when all 6 regions were pooled at 74.4% correct (“best” classification was observed for 4-region combinations, 77.8%). Thus, voxels in the left amygdala were capable of supporting decoding at levels similar to those observed for other regions (at least when 6/9 subjects were considered), and when this region was combined with other choice-related ROIs, robust classification levels were obtained.

Ranking Regions in Terms of Decoding Importance via “Virtual Lesions”

Our results revealed that many different coalitions of regions supported robust classification. One question of considerable interest is to develop methods to quantify the “relative importance” of a given region for a given function given the

multivariate analysis framework adopted here (cf., Hanson and others 2004). We took an initial step in this direction by determining the reduction in classification accuracy when a given region was “lesioned” and thus not available to be combined with the remaining regions in the prediction of behavioral choice (Materials and Methods). The results of our analysis are illustrated in Figure 5 and reveal that although the “lesion” of specific choice-related regions (plus the left amygdala) did not have huge effects on classification, regions did not appear to be “equipotential.” Overall, the left amygdala had the largest effect on accuracy when it was not employed, causing a 6.7% reduction in classification rate.

Decoding the Stimulus

If fMRI signals could be used to predict behavioral choice, could they be used instead to predict the “physical stimulus” presented on a given trial (cf., Haxby and others 2001), independent of the choice? To investigate this question, classification of voxelwise fMRI signals was attempted in terms of the stimulus (fear-containing vs. neutral target faces). For each ROI, instead of ordering individual voxels in terms of how well they predicted behavioral choice, we ordered them in terms of stimulus predictability (computed by employing a signal detection theory analysis analogous to our computation of CP). Stimulus prediction when considering the set of stimulus-responsive ROIs was in general poorer than the prediction of choice and reached 68.7% for 67-ms targets (value obtained for the 5-region combination) and 65.8% for 33-ms targets (value obtained for 4-region combinations). These results support the notion that previous predictions of “fear” and “no fear” behavioral choice reflected the encoding of choice per se and not other stimulus-based features.

Decoding Behavioral Choice from Nonchoice-Related Regions

We performed a control analysis in which accuracy of prediction of behavioral choice was determined for a set of 5 regions that did not exhibit robust choice-related activation. The 5 regions comprised sites with strong task-related activation at the group level (i.e., task vs. rest) and included left posterior intraparietal sulcus ($x = -29, y = -74, z = 25$), right anterior

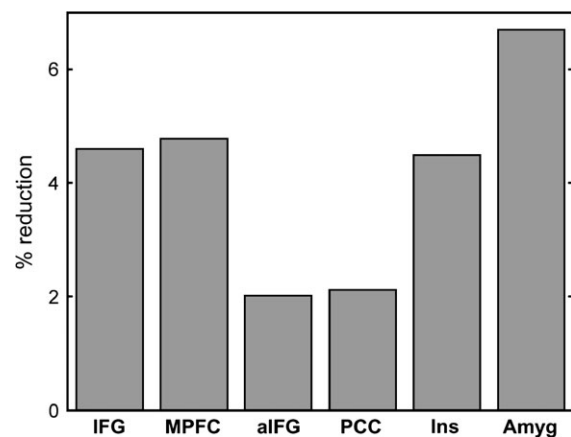


Figure 5. Percentage reduction in classification accuracy when specific regions were “lesioned.” For this analysis, we considered the set of choice-related ROIs in addition to the left amygdala. Accuracy was most strongly affected when the amygdala was not employed in multiple-region combinations (i.e., it was “lesioned”). Amyg, (left) amygdala; aIFG, (right) anterior IFG; Ins, (left) insula.

intraparietal sulcus ($x = 29, y = -58, z = 47$), bilateral frontal eye fields (left: $x = -37, y = -5, z = 44$; right: $x = 37, y = -5, z = 44$), and supplementary eye fields ($x = -3, y = 2, z = 56$). For the present analysis, we employed voxels from these regions but excluded voxels that significantly predicted behavioral choice at the individual level (based on the computation of CPs)—choice-related activation in these regions was sparse and not consistent across subjects. Classification accuracy for 1 through 5 regions was 54.7%, 54.4%, 54.6%, 54.5%, and 53.4% correct. Such results reveal that accurate prediction was not possible when voxels from nonchoice-related regions were considered. Moreover, considering multiple regions did not lead to improvements in classification accuracy. Note also that adding nonchoice regions to an individual choice-related region did not improve classification either. When 1 nonchoice region was added to 1 choice region, on average prediction accuracy was 63.5% (the average single-region accuracy for choice-related regions was 64.1%); when 5 nonchoice regions were added to 1 choice region, prediction accuracy decreased to 59.8%. These results argue against the interpretation that increases in prediction accuracy were simply obtained by considering multiple regions and argue that only the combination of choice-related regions was effective.

Feature Selection and Prediction Accuracy

A key problem in multivariate classification concerns the question of “feature selection” (see also Discussion): how does one select the input features (voxels in the present case) that optimize classification accuracy? An exhaustive search of all possible subsets is only viable when the number of variables is not “too large”; such strategy quickly becomes computationally prohibitive. Given that recent pattern classification applications routinely employ a relatively large number of features (>50), research on feature selection methods has grown considerably (Guyon and Elisseeff 2003). Here, we investigated classification accuracy when we employed a recent method called recursive feature elimination (Guyon and others 2002), an iterative procedure that is an instance of backward feature elimination (Kohavi and John 1997). The basic idea is to start with all available features and iteratively eliminate the “least informative” feature (i.e., the smallest weight) at each cycle. This process is repeated until no features are available. In the present context, for each region, all voxels with significant CPs were initially employed and iteratively eliminated until no further voxels were available. At each iterative step, classification accuracy was determined by the leave-one-from-each-class-out cross-validation procedure described in Materials and Methods. Figure 6 illustrates the potential of recursive feature elimination when all 10 regions were employed for classification of 67-ms targets. Prediction accuracy started at around 64% (as before) and increased until it reached a value of 84.9% (significant linear trend, $P < 0.0001$). For 33-ms targets, a very similar pattern of classification accuracy was observed; prediction peaked at 87.4% for 10 regions. These results suggest that improved methods of feature selection have considerable potential in boosting prediction accuracy; in the present examples, we observed improvements on the order of ~10%.

Decoding with Nonlinear SVMs, Neural Networks, and Additional Results

Linear SVMs of the type considered so far perform classification based on a “separating hyperplane” that is constructed based

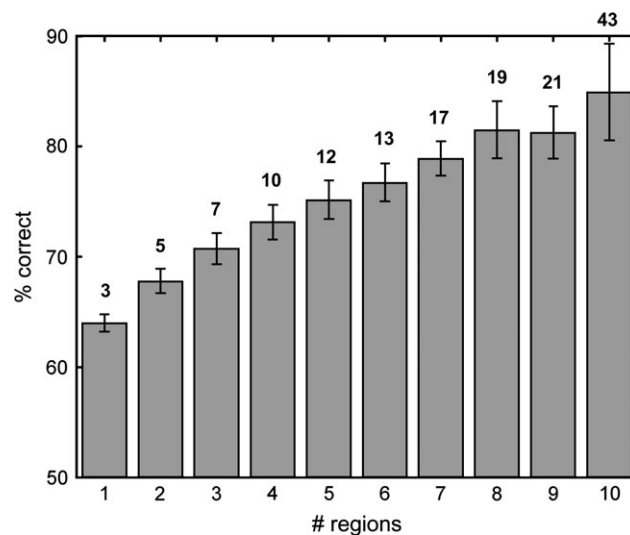


Figure 6. Decoding perceptual decisions when recursive feature elimination was employed for feature selection (both choice-related and stimulus-responsive regions were considered). Prediction accuracy increased linearly as a function of the number of regions and approached 85% correct for the 10-region combination. The values over individual bars show the average number of voxels (summed across regions) associated with the maximum classification rate. Error bars indicate the standard error of the mean.

on the training data. In many cases, nonlinear classifiers are extremely beneficial as they allow for classifications based on nonlinear decision surfaces (e.g., the exclusive-OR problem). An extremely convenient and powerful feature of SVMs is that one can perform nonlinear classification in a relatively straightforward way—technically, by considering nonlinear kernel functions. Although the objective of this paper was not to perform an extensive benchmark of alternative machine learning architectures (i.e., to investigate the difficult problem of model selection), we investigated 2 types of nonlinear SVMs: polynomial (defined by the order of the polynomial) and RBF. For a third-degree polynomial SVM, classification accuracy for the 5 choice-related ROIs was 62.8%, 66.5%, 69.2%, 70.4%, and 70.3% when 1 through 5 regions were considered; for a fifth-degree polynomial, prediction accuracy only reached 68.3% for 4-region combinations. For the RBF classifier, classification values were 63.7%, 68.4%, 71.8%, 74.2%, and 72.0%, respectively. Therefore, considering nonlinear classifiers did not improve classification accuracy relative to the levels obtained with linear architectures. Likewise, considering a neural network (Materials and Methods) did not improve prediction accuracy. Classification for a network with 10 hidden nodes in the intermediate layer ranged between 60.8% and 72.1% correct for 1- through 5-region combinations (peak accuracy was observed with 4-region combinations). Very similar values were obtained with 5, 15, or 20 hidden nodes (peak classification ranged between 69.8% and 72.2% correct); in all cases, the training error was essentially 0 (i.e., <0.0001).

In a previous section, we showed that the decoding of choice from nonchoice-related regions was poor and did not improve when considering multiple regions (values ranged between 53% and 54% correct). We revisited this question by considering nonlinear SVMs, as in the preceding paragraph. In all cases, prediction accuracy was low and did not exceed 55.3%; moreover, accuracy did not improve as a function of the number of regions considered (a grid search on SVM parameters did not lead to

changes greater than 2% in prediction accuracy; see Materials and Methods). Likewise, the neural network values did not exceed 55.1% and did not improve with the number of regions. Thus, the poor prediction accuracy obtained with linear classifiers was not simply due to the choice of a suboptimal architecture (although it is conceivable that other architectures and/or parameters could improve classification). Taken together, the results strengthen the view that accurate prediction is not feasible when only voxels from nonchoice regions are considered.

Thus far, we have attempted classification when only a relatively small subset of the total voxels within the brain was considered. Alternative strategies would be to consider all voxels inside the brain or all task-responsive voxels (at a certain threshold, see also Discussion). We attempted “near full volume” classification with a linear SVM by employing 16 160 voxels (average across subjects) after “lightly” thresholding individual task-activation maps ($P < 0.1$). In this manner, a significant portion of the voxels within the brain was employed, including “all” voxels in our ROIs. In such case, prediction accuracy was poor at a level of 56.0%.

Information Theory and Distributed Encoding of Perceptual Decisions

The results of the preceding sections revealed that prediction accuracy increased as additional regions were considered. To further investigate how signals from one region are potentially combined with those from another region so as to allow better prediction accuracy, we carried out an information theoretic analysis. We computed the joint information carried by a pair of ROIs together and compared that with the transmitted information calculated for each ROI separately. Following Richmond and others (Gawne and others 1996), we defined an index $J = I_{AB}/(I_A + I_B)$, which should be greater than 1 if the information carried by the 2 ROIs considered together is greater than the sum of the information carried out by each region alone. The average value of J across subjects was 1.23 ± 0.15 (standard error of the mean), a value that was significantly greater than 1 (1-tailed t -test, $P < 0.01$), which indicates a “synergistic” combination.

Another way of stating the above findings is to consider that mutual information when “pairs” of regions were considered reduced uncertainty on average by 29.6% (i.e., mutual information as a percentage of the uncertainty of behavioral choice). At the same time, when “individual” regions were considered, the percentage reduction in uncertainty was on average 12.2%. It should be stressed that given the conservative nature of the bias correction procedure adopted for the estimation of information (Golomb and others 1997), these values are likely underestimates of the true values of the information carried by fMRI responses.

Discussion

In the present article, we showed that perceptual decisions can be reliably predicted, or “decoded,” from single-trial fMRI data. Prediction accuracy was investigated in terms of the number of voxels utilized per ROI and the total number of ROIs. Average classification accuracy ranged between 64% and 78% correct and increased from when 1 ROI was employed (~64%) to when 10 ROIs were used (~78%). When more sophisticated methods of feature selection were employed, accuracy improved and ranged from 64% to 87% when recursive

feature elimination was utilized. When the best classifications per subject were averaged, accuracy levels ranged between 74% and 86% correct. Finally, and somewhat surprisingly, classification rates were very similar for 67- and 33-ms target durations. Overall, these results demonstrate that challenging brain states can be decoded with high accuracy at conditions near the threshold of visual awareness from “distributed single-trial” data.

Previous research addressing the perception of fear has largely focused on the role of the amygdala, and possibly a few other structures, such as the fusiform gyrus and the orbitofrontal cortex (Dolan 2003). This is especially the case for masking studies similar to the present one in which target faces are presented briefly and backwardly masked in order to manipulate visual awareness. In the present paper, our approach was, instead, to probe how signals from multiple regions “simultaneously” predict trial-by-trial perceptual decisions. Our analysis initially focused on a set of regions that we previously showed to contain “individual voxels” that were predictive of choice in the masking task (Pessoa and Padmala 2005): PCC, MPFC, right IFG (at both more anterior and more posterior sites), and left insula. Interestingly, all these regions are involved in affective processing and are interconnected with the amygdala, which was also predictive of choice when individual voxels were employed, although less consistently (Pessoa and Padmala 2005). The set of choice-related regions is also known to be involved in the control of autonomic functions. For instance, the MPFC is connected to the periaqueductal gray, nucleus accumbens, and hypothalamus and has outflow to autonomic and visceromotor centers (Barbas 2000). Thus, the MPFC is well positioned to modulate brain responses according to the emotional significance of the incoming stimulus. Overall, the PCC, MPFC, right IFG, and left insula, in conjunction with the amygdala, may be part of a network of brain regions involved in emotional processing that participate in the “decision” that a fearful face has been seen.

In addition, prior work on the perception of emotion-laden visual stimuli has focused on how the physical properties of the stimuli affect brain responses. The emphasis of the present investigation was, instead, to probe how moment-to-moment fluctuations in the perceptual choices reported by the subject were correlated with brain responses. As stated, we investigated how multiple voxels from one or more regions were correlated with behavioral choice. In particular, our goal was to employ machine learning techniques to investigate how perceptual decisions were linked to distributed patterns of fMRI activation.

Distributed Encoding of Behavioral Choice

For regions containing voxels that were individually predictive of behavioral choice, prediction accuracy increased when signals from several ROIs were combined, suggesting that extra information was available when additional regions were considered (see below). Thus, “distributed” patterns of activation predicted behavioral choice better than individual voxels alone.

What does it mean for a behavioral choice to be represented in a distributed fashion? It does not mean, for instance, that the entire network is directly responsible for “deciding” that a fearful target is present (Schall 2005). For instance, choice-related signals in one region may reflect computations made at another site in the brain (Kim and Shadlen 1999). We suggest that the conscious decision that a fearful face is present is “represented” across a network of brain regions that prepare the organism to appropriately handle emotionally challenging stimuli and

that regulate the associated emotional response. In this regard, further elucidation of the neural representation of the perception of fear would greatly benefit from the direct comparison of electrophysiological measures and blood oxygen level-dependent (BOLD) measures obtained during fMRI (Logothetis and others 2001; Heeger and Ress 2002). In particular, understanding how BOLD-BOLD correlations are linked to correlations at the neural level is important to constrain population code models (Panzeri and others 2003; Schneidman and others 2003; Nevado and others 2004).

Interestingly, stimulus-responsive regions, such as the fusiform gyrus and superior temporal sulcus, predicted behavioral choice at similar levels compared with those obtained with choice-related regions. In fact, stimulus-responsive regions did not predict the physical stimulus (fear-containing vs. neutral stimuli) at the same levels that they predicted perceptual decisions (values reached 69% for 67-ms stimuli and 65% for 33-ms stimuli). Moreover, classification improved when we considered the complete set of 10 regions, instead of just considering choice-related regions. These results reveal that in the case of emotional perception, stimulus-responsive regions encode more than stimulus features. We suggest that choice-related signals in these regions may contribute to further preparing the organism to handle an emotional event.

Thus, in general, the prediction of the stimulus was less effective than the prediction of choice (especially for 33-ms targets). Such results support the view that the decision as to whether a fearful face is present in the visual field instates a robust representation across several brain regions that are involved in affective processing. Such representation appears to be much more robust than the traces linked to the physical stimulus itself. Overall, the present findings are consistent with our previous results that revealed that responses evoked by briefly presented and masked faces are strongly linked to subjective reports and less so to the physical characteristics of the stimulus (Pessoa and others 2006).

Why do more regions and more voxels per region improve classification? Although all ROIs were selected on the basis of exhibiting choice-related activation at the group level, their signals were not strongly correlated. Indeed, the overall average interregion voxel-to-voxel correlation was 0.2 (averaged across all voxel pairs). As would be expected given spatial smoothing of the data and other factors, the within-region correlation was substantially higher (0.57). Thus, additional information is available when voxels from distinct regions are combined and, to some extent, even when voxels from the same region are pooled. This view is strengthened when we consider our information theoretic analysis. The information carried by one region increased when a second region was considered. Note that if the information carried by 2 regions were “independent,” no such reliable increase would have been expected. Critically, if the information carried by the 2 regions had been “redundant,” our index J would have been less than 1 (0.5 if the 2 regions were completely redundant). These results suggest that the encoding of the perceptual decision is not simply “rerepresented” across several regions. Instead, utilizing multiple regions for classification significantly improved accuracy, revealing that extra information was available when additional regions were considered.

Until recently, it was largely assumed that 33-ms masked fearful faces did not reach visual awareness (Pessoa 2005). However, we showed that there is great interindividual vari-

ability in sensitivity to fearful faces and that a sizeable percentage of subjects is actually able to reliably detect them, when awareness is assessed via objective methods (Pessoa and others 2005, 2006). In the present experiment, 7/9 subjects detected 33-ms target fearful faces better than chance. Behavioral choices during such just-above-threshold emotional perception could be predicted with high accuracy from distributed fMRI activation. Interestingly, classification rates for the 2 subjects who could not reliably detect fearful targets were within the same range as the remaining participants. Again, these results are consistent with our recent findings that activations in several brain regions during the detection of fearful faces, including the amygdala and the fusiform gyrus, are largely driven by “perception” (as reflected by behavioral choice) and not the physical stimulus (Pessoa and others 2006)—for example, evoked signals in these regions are greater for hits (correctly detected fearful faces) than misses (missed fearful faces), even though the physical stimulus is identical. Although subsequent studies employing larger subject pools are needed, our current findings appear to indicate that behavioral choice can be predicted even during conditions in which subjects are objectively unaware of fearful-face targets.

Previous Work on Distributed Representations and Decoding

In recent years, multivariate techniques for fMRI data analysis have been increasingly used to investigate neural processes. In an early study, Haxby and others (2001) showed that the distributed pattern of activation across ventrotemporal cortex could be used to predict the visual object viewed by the participant (see also Spiridon and Kanwisher 2002). Recent approaches have employed more sophisticated processing strategies, including the use of SVMs and other machine learning techniques (Cox and Savoy 2003; Hanson and others 2004; Mitchell and others 2004; O’Toole and others 2005). In a comprehensive study, Mitchell and others (2004) showed the feasibility of training pattern classifiers to distinguish cognitive states, such as the perception of pictures or sentences and the perception of nouns or verbs. They showed, for example, that it is possible to classify pictures versus sentences with greater than 90% accuracy, although the classification of noun versus verb proved more challenging (~77%).

One critical variable for fMRI-based classification is the amount of data that is employed. In one extreme, one can employ an entire experimental run or sets of runs (Haxby and others 2001). More recent approaches have utilized much less data. For example, Cox and Savoy (2003) employed short blocks (20 s) containing 10 stimulus repetitions. Recent approaches have employed either 30-s blocks (Haynes and Rees 2005) or 16-s “trials” in which stimulation flashed on and off every 250 ms (Kamitani and Tong 2005). Single-trial fMRI data are notoriously noisy, and only simple motor acts had been decoded in the past (Dehaene and others 1998). The work of Mitchell and others (2004) revealed encouraging results when trial-related data were used. Short temporal segments from single trials were utilized, for instance, an 8-s input linked to the viewing of a picture or a sentence. The approach adopted in the present paper was to take the extreme strategy of performing classification of challenging perceptual decisions based on single-trial data. Our average classification rates ranged from 64% to 78% correct depending on how many regions were

utilized (or higher with recursive feature elimination; see below) and exceeded 80% when the “best” classification was assessed.

Another key consideration in fMRI-based classification is the issue of “feature selection.” This general pattern classification problem refers to the question of selecting informative input attributes (instead of using all possible attributes) that aid in discriminating the target classes (Guyon and Elisseeff 2003). A directly related question in the context of fMRI concerns the selection of the voxels to be used in the classification process. A viable strategy is to employ all task-responsive voxels (at some threshold value of the task vs. rest contrast), thereby focusing the analysis on voxels with large signal-to-noise ratios. Such selection procedure does not consider a priori whether voxels (i.e., features) distinguish the target classes. Naturally, such information can be utilized when choosing the input features. For instance, one could train a classifier on single voxels and utilize classification accuracy as a measure of the voxel’s discriminating power. Next, the n voxels with the highest discriminability could be selected and employed for classification (Mitchell and others 2004). Our own method can be viewed as an instance of such strategy. Initially, we determined the most predictive voxels according to their CP values. Only voxels with significant CPs were then considered and were added to the input vector in decreasing order of predictability. Intriguingly, Mitchell and others (2004) reported that utilizing highly “active” voxels led to higher classification accuracy relative to classification based on “discriminative” voxels—because, for instance, discriminative voxels may “overfit” the data due to noise. In our study, however, adopting such strategy produced weak predictability. As reported in Results, when we considered 5 task-responsive regions that did not exhibit choice-related activation, classification was quite poor (~54%; although, as stated before, it is conceivable that other architectures and/or parameters could improve classification in such case). In an additional analysis (not reported in Results), when we selected voxels based on how active they were (indexed by their t value in the task vs. rest contrast) and employed the 5 stimulus-responsive ROIs (which exhibited robust task-related activation), for 67-ms targets, classification was relatively poor (58.8%, 60.8%, 62.0%, 62.7%, and 62.2% correct, for 1–5 regions, respectively). Finally, when we attempted “near full volume” classification, prediction accuracy was only 56.0% correct, indicating that learning was not able to “tune out” uninformative features. At the same time, we illustrated that fMRI decoding potentially has much to gain from recent, sophisticated (albeit computationally expensive) methods of feature selection, such as recursive feature elimination (Guyon and others 2002). When such method was applied, prediction accuracy approached 84–87% when 10 regions were considered. Collectively, our results suggest that proper feature selection is instrumental in fMRI decoding.

In summary, our results suggest that multivariate analysis strategies have considerable potential in helping to elucidate the neural correlates of visual awareness (Beck and others 2001; Pessoa and Ungerleider 2004) and the encoding of perceptual decisions (Heekeren and others 2004) in the brain and lead us closer to the “dream of a single image for a single event” (Grabowski and Damasio 1996).

Notes

This work was supported in part by NIMH grant 1R01 MH071589-02, a “Research Seed Funds” award by Brown University, a “Pilot grant”

award by the Brain Science Program at Brown University, and by the Ittleson Foundation. We thank the anonymous reviewers for valuable feedback on earlier versions of the manuscript. We also thank S. Panzeri for providing code and assisting in the bias correction procedure for the computation of mutual information. *Conflict of Interest:* None declared.

Address correspondence to Luiz Pessoa, Department of Psychology, 89 Waterman Street, Brown University, Providence, RI 02912, USA. Email: pessoa@brown.edu.

References

- Adolphs R. 2002. Neural systems for recognizing emotion. *Curr Opin Neurobiol* 12:169–177.
- Aggleton J, editor. 2000. *The amygdala: a functional analysis*. Oxford: Oxford University Press.
- Ash R. 1965. *Information theory*. New York: Wiley.
- Barbas H. 2000. Connections underlying the synthesis of cognition, memory, and emotion in primate prefrontal cortices. *Brain Res Bull* 52:319–330.
- Beck DM, Rees G, Frith CD, Lavie N. 2001. Neural correlates of change detection and change blindness. *Nat Neurosci* 4:645–650.
- Boser B, Guyon I, Vapnik V. 1992. A training algorithm for optimal margin classifiers. *Fifth Annual Workshop on Computational Learning Theory*; Pittsburgh, PA: ACM. p 144–152.
- Britten KH, Newsome WT, Shadlen MN, Celebrini S, Movshon JA. 1996. A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Vis Neurosci* 13:87–100.
- Burges CJC. 1998. A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discov* 2:121–167.
- Cover TM, Thomas JA. 1991. *Elements of information theory*. New York: Wiley.
- Cox DD, Savoy RL. 2003. Functional magnetic resonance imaging (fMRI) “brain reading”: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* 19:261–270.
- Dehaene S, Le Clecq HG, Cohen L, Poline JB, van de Moortele PF, Le Bihan D. 1998. Inferring behavior from functional brain images. *Nat Neurosci* 1:549–550.
- Dodd JV, Krug K, Cumming BG, Parker AJ. 2001. Perceptually bistable three-dimensional figures evoke high choice probabilities in cortical area MT. *J Neurosci* 21:4809–4821.
- Dolan R. 2003. Emotion, cognition, and behavior. *Science* 298:1191–1194.
- Efron B, Tibshirani RJ. 1993. *An introduction to the bootstrap*. New York: Chapman & Hall.
- Ekman P, Friesen WV. 1976. *Pictures of facial affect*. Palo Alto, CA: Consulting Psychologists Press.
- Gawne TJ, Kjaer TW, Hertz JA, Richmond BJ. 1996. Adjacent visual cortical complex cells share about 20% of their stimulus-related information. *Cereb Cortex* 6:482–489.
- Golomb D, Hertz J, Panzeri S, Treves A, Richmond B. 1997. How well can we estimate the information carried in neuronal responses from limited samples? *Neural Comput* 9:649–665.
- Grabowski TJ, Damasio AR. 1996. Improving functional imaging techniques: the dream of a single image for a single mental event. *Proc Natl Acad Sci USA* 93:14302–14303.
- Grunewald A, Bradley DC, Andersen RA. 2002. Neural correlates of structure-from-motion perception in macaque V1 and MT. *J Neurosci* 22:6195–6207.
- Guyon I, Elisseeff A. 2003. An introduction to variable and feature selection. *J Mach Learn Res* 3:1157–1182.
- Guyon I, Weston J, Barnhill S, Vapnik V. 2002. Gene selection for cancer classification using support vector machines. *Mach Learn* 46:389–422.
- Hanson SJ, Matsuka T, Haxby JV. 2004. Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a “face” area? *Neuroimage* 23:156–166.
- Haxby JV, Gobbini MI, Furey ML, Ishai A, Shouten JL, Pietrini P. 2001. Distributed and overlapping representations of faces and objects in the ventral temporal cortex. *Science* 293:2425–2430.
- Haxby JV, Hoffman EA, Gobbini MI. 2000. The distributed human neural system for face perception. *Trends Cogn Sci* 4:223–233.

- Haynes JD, Rees G. 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat Neurosci* 8:686-691.
- Heeger DJ, Ress D. 2002. What does fMRI tell us about neuronal activity? *Nat Rev Neurosci* 3:142-151.
- Heekeren HR, Marrett S, Bandettini PA, Ungerleider LG. 2004. A general mechanism for perceptual decision-making in the human brain. *Nature* 431:859-862.
- Ishai A, Pessoa L, Bickle PC, Ungerleider LG. 2004. Repetition suppression of faces is modulated by emotion. *Proc Natl Acad Sci USA* 101:9827-9832.
- Kamitani Y, Tong F. 2005. Decoding the visual and subjective contents of the human brain. *Nat Neurosci* 8:679-685.
- Kim JN, Shadlen MN. 1999. Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nat Neurosci* 2:176-185.
- Kohavi R, John G. 1997. Wrappers for feature selection. *Artif Intell* 97:273-324.
- Logothetis NK, Pauls J, Augath M, Trinath T, Oeltermann A. 2001. Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412:150-157.
- Macmillan NA, Creelman CD. 1991. *Detection theory: a user's guide*. New York: Cambridge University Press.
- Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X, Just M, Newman S. 2004. Learning to decode cognitive states from brain images. *Mach Learn* 57:145-175.
- Moller MF. 1993. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Netw* 6:525-533.
- Mourao-Miranda J, Bokde AL, Born C, Hampel H, Stetter M. 2005. Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. *Neuroimage* 28:980-995.
- Nevado A, Young MP, Panzeri S. 2004. Functional imaging and neural information coding. *Neuroimage* 21:1083-1095.
- O'Toole AJ, Jiang F, Abdi H, Haxby JV. 2005. Partially distributed representations of objects and faces in ventral temporal cortex. *J Cogn Neurosci* 17:580-590.
- Panzeri S, Pola G, Petersen RS. 2003. Coding of sensory signals by neuronal populations: the role of correlated activity. *Neuroscientist* 9:175-180.
- Pessoa L. 2005. To what extent are emotional visual stimuli processed without attention and awareness? *Curr Opin Neurobiol* 15:188-196.
- Pessoa L, Japee S, Sturman D, Ungerleider LG. 2006. Target visibility and visual awareness modulate amygdala responses to fearful faces. *Cereb Cortex* 16:366-375.
- Pessoa L, Japee S, Ungerleider LG. 2005. Visual awareness and the detection of fearful faces. *Emotion* 5:243-247.
- Pessoa L, Kastner S, Ungerleider LG. 2002. Attentional control of the processing of neural and emotional stimuli. *Cogn Brain Res* 15:31-45.
- Pessoa L, Padmala S. 2005. Quantitative prediction of perceptual decisions during near-threshold fear detection. *Proc Natl Acad Sci USA* 102:5612-5617.
- Pessoa L, Ungerleider LG. 2004. Neural correlates of change detection and change blindness in a working memory task. *Cereb Cortex* 14:511-520.
- Schall JD. 2005. Decision making. *Curr Biol* 15:R9-R11.
- Schneidman E, Bialek W, Berry MJ II. 2003. Synergy, redundancy, and independence in population codes. *J Neurosci* 23:11539-11553.
- Spiridon M, Kanwisher N. 2002. How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron* 35:1157-1165.
- Vapnik VN. 1995. *The nature of statistical learning theory*. New York: Springer-Verlag.