HHS Public Access

Author manuscript

Annu Rev Psychol. Author manuscript; available in PMC 2021 February 08.

Published in final edited form as:

Annu Rev Psychol. 2012 ; 63: 483–509. doi:10.1146/annurev-psych-120710-100412.

Decoding Patterns of Human Brain Activity

Frank Tong, **Michael S. Pratte**

Psychology Department and Vanderbilt Vision Research Center, Vanderbilt University, Nashville, Tennessee 37240, USA

Abstract

Considerable information about mental states can be decoded from non-invasive measures of human brain activity. Analyses of brain activity patterns can reveal what a person is seeing, perceiving, attending to, or remembering. Moreover, multidimensional models can be used to investigate how the brain encodes complex visual scenes or abstract semantic information. Such feats of “brain reading” or “mind reading”, though impressive, raise important conceptual, methodological, and ethical issues. What does successful decoding reveal about the cognitive functions performed by a brain region? How should brain signals be spatially selected and mathematically combined, to ensure that decoding reflects inherent computations of the brain rather than those performed by the decoder? We will highlight recent advances and describe how multivoxel pattern analysis (MVPA) can provide a window into mind-brain relationships with unprecedented specificity, when carefully applied. However, as brain-reading technology advances, issues of neuroethics and mental privacy will be important to consider.

Keywords

fMRI; multivoxel pattern analysis; MVPA

INTRODUCTION

Imagine that it is the future, an unknown year in the 21st century. A participant is brought into a neuroimaging lab and asked to lie back comfortably on a padded bed table, which is slowly glided into a brain scanner. The participant watches a brightly colored display as it provides a virtual tour of every painting in the *Musée d’Orsay*. All the while, non-invasive measures of that person’s brain activity are discretely taken, the arrays of numbers quickly transferred to the memory banks of a high-speed digital computer. After hours of brain scanning and computer analysis, the real scientific test begins. A randomly drawn painting is shown again to the observer. The computer analyzes the incoming patterns of brain activity from the participant’s visual cortex and makes the following prediction with 99% confidence: *She is looking at Painting #1023, Cezanne’s Still Life with Apples and Oranges*.

Correspondence and requests for materials should be addressed to Frank Tong (frank.tong@vanderbilt.edu).

Key terms

Decoding – neural decoding involves determining what stimuli or mental states are represented by an observed pattern of neural activity

Encoding – how a stimulus or mental state is encoded or represented by specific patterns of neural activity

The experimenter turns to look at the computer screen, and indeed, the participant is looking at a plateful of pastel-colored red and yellow apples, ripe oranges stacked in a porcelain bowl, all carefully arranged in the thick folds of a tousled white tablecloth. Another randomly drawn picture is shown, and the computer correctly predicts *Landscape with Green Trees by Maurice Denis*.

What does this remarkable scientific demonstration reveal — successful mind reading? Have the neuroscientists effectively cracked the brain's internal code for vision, such that they now understand how features and objects are represented in the mind's internal eye? We will refer to this as *Science Fiction Story #1*.

The lab volunteer has kindly offered to participate in a second experiment. This time she is shown two paintings in quick succession (*Bedroom in Arles, The White Horse*), and then is asked to pick one and hold that image in mind for several seconds. She imagines a horse standing in a shallow river, head bent low as if looking at its own reflection in the slowly flowing stream. The computer quickly scans the matrix of numbers streaming in. Although brain activity levels are substantially weaker as she gazes steadily at the blank screen, compared to moments ago, a pattern begins to emerge from her visual cortex. The computer announces, with 85% confidence, that the participant is imagining the second painting, *The White Horse*. Would successful decoding in this case indicate that the neural codes for imagination and internal visual thoughts have been successfully decoded? More generally, what would such a demonstration reveal about the visual and cognitive functions performed by the brain? We will refer to this as *Science Fiction Story #2*.

In reality, these stories represent more fact than fiction. A simplified version of *Science Fiction Story #1* was carried out at the start of the 21st century in a pioneering study by Haxby and colleagues (2001). The authors used functional magnetic resonance imaging (fMRI) to measure patterns of BOLD activity, focusing on object-responsive regions in the ventral temporal cortex. By comparing the similarity of brain activity patterns between the first and second half of the experiment, the authors showed that these high-level object areas could accurately predict whether participants were viewing pictures of faces, houses, chairs, cats, bottles, shoes, scissors, or scrambled stimuli (Figure 1a). The use of more sophisticated pattern classification algorithms (Figure 1b) greatly improved researchers' ability to predict what object categories people were viewing (Carlson et al 2003; Cox & Savoy 2003). Subsequently, Kamitani & Tong (2005) discovered that it was possible to decode orientation- and direction-selective responses with surprising accuracy (Figure 2), even though such feature-selective information is primarily organized at the scale of submillimeter columns in the visual cortex. Thus, fMRI pattern analysis could reveal cortical information that would otherwise fail to be detected. Perhaps the most striking demonstration of *Science Fiction Story #1* comes from the work of Kay et al (2008). They presented over a thousand natural images to observers and then characterized the response preferences of each voxel in the visual cortex, specifying their selectivity for retinotopic position, spatial frequency, and orientation. When the observers were shown a new set of 120 pictures, each of a different real-world scene, the authors could accurately predict which new image was being viewed by finding the best match between the observed pattern of activity and the predicted activity of these modeled voxels.

These studies reveal an unprecedented ability to predict the basic visual features, complex objects, or natural scenes that are being viewed by the participant. By combining fMRI with sensitive pattern analysis methods, accurate predictions about the viewed stimulus can be made. Yet it would be mistake to consider such feats as examples of mind reading. Why? Because the experimenter does not need a mind-reading device to achieve this performance. The same result could be achieved by simply looking over the participant's shoulder, "*Oh, she is looking at painting #1023, Cezanne's still life with apples and oranges*." Put another way, one could perform these same feats by reading out the activity patterns formed on the retina, even though conscious processing of the image has yet to take place. Activity patterns on the retina would remain robust even if the person were anaesthetized or fell into a deep coma. So instead, *Science Fiction Story #1* should be considered an example of *brain reading.*

Science Fiction Story #2 can be better justified as a demonstration of *mind reading.* Here, information that is fundamentally private and subjective is being decoded from the person's brain; the only alternative would be to ask the participant directly about what she is thinking and to hope for an honest reply. Ongoing research is just beginning to probe the possibilities and limits of reading out subjective information from the human brain.

In this review, we will discuss recent advances in brain reading and mind reading, and consider important conceptual and methodological issues regarding how to apply these techniques to the study of human cognition. The brain reading approach has revealed how different types of stimulus information are represented in specific brain areas, and some studies provide clues to the functional organization of these representations. Pattern analysis of brain activity can also be adapted to perform feats of mind reading, to extract information about a person's subjective mental state or cognitive goal. We will consider whether such feats of mind reading should be likened to fancy parlor tricks that require the assistance of a brain scanner, or whether these methods can be used to genuinely advance our understanding of brain function. Studies employing this mind-reading approach have revealed how particular representations are activated or called upon during conscious perception, attentional selection, imagery, memory maintenance and retrieval, and decision making. As will be seen, careful consideration of experimental design, analysis, and interpretation of the data is essential when adopting powerful pattern analysis algorithms to probe the functions that might be carried out by a brain area. As these methodologies continue to advance, it will become increasingly important to consider the ethical implications of this technology.

There have been previous reviews on the topic of fMRI decoding (sometimes called multivoxel pattern analysis or MVPA, (Haynes & Rees 2006; Norman et al 2006) as well as more in-depth reviews on the technical aspects of decoding and encoding (Kriegeskorte 2011; Naselaris et al 2011; O'Toole et al 2007; Pereira et al 2009). In this review, we will highlight recent studies and discuss key issues regarding how fMRI pattern analysis can be used to advance understanding of the bases of human cognition.

BRIEF TUTORIAL ON MULTIVOXEL PATTERN ANALYSIS

Traditional methods of fMRI analysis treat each voxel as an independent piece of data, using statistical tests to determine whether that voxel responded more in some experimental conditions than others. Such analyses are *univariate*: the analysis of one voxel has no impact on the analysis of any other. By contrast, *multivariate* pattern analysis extracts the information contained in the patterns of activity among multiple voxels, so that the *relative differences* in activity between voxels can provide relevant information. Whereas univariate statistical analyses are designed to test whether some voxels respond more to one condition than another, multivariate analyses are designed to test whether two (or more) experimental conditions can be distinguished from one another based on the activity patterns observed in a set of voxels. Critically, multivariate methods might be able to tell apart the activity patterns for two different conditions, even if the average level of activity does not differ between conditions.

Figure 1b illustrates the simplest example of multivariate pattern analysis involving two experimental conditions (shown in red and green) and just two voxels, with the response amplitude of each voxel shown on separate axes. Each dot corresponds to a single *activity pattern* or *data sample*, with its position indicating the strength of the response for voxels 1 and 2. The Gaussian density plots in the margins indicate that either voxel alone does a rather poor job of separating the two experimental conditions. Nevertheless, the two conditions can be well separated by considering the pattern of responses to both voxels, as indicated by the separating boundary line. In this particular example, the responses of voxels 1 and 2 are positively correlated and the classification boundary helps to remove this correlated “noise” to better separate the two experimental conditions. If there were three voxels, a third dimension would be added; the red dots and greens dots would form two largely separated (but still overlapping) clouds of points and the classification boundary would consist of a linear plane that best divides those two clouds. Typically, anywhere from a few dozen to several thousand voxels might be used for fMRI pattern analysis, so an activity pattern with N voxels would be represented in an N-dimensional space, and “clouds” of dots representing the two classes would be separated by a linear *hyperplane*. (Multiclass classification analysis involves calculating multiple hyperplanes to carve up this multidimensional space among three or more conditions.)

The goal of linear pattern classification algorithms, such as support vector machines (SVM), linear discriminant analysis (LDA), or logistic regression, is to find the linear hyperplane that best separates the two (or more) conditions in this multi-dimensional voxel space. The accuracy of classification performance is usually assessed using cross-validation, which involves dividing the full set of data samples into separate sets for training and testing the classifier. Typically an entire fMRI run or perhaps just one sample from each condition is reserved for the test set. The classifier is trained with the remaining data to obtain the classification boundary, which is then used to predict the class of each data sample (e.g., “red” or “green”) in the test set. This procedure can be done iteratively, so that every sample in the data set is tested and an overall measure of classification accuracy is obtained. Classification accuracy reflects the amount of information available in a set of voxels for discriminating between the experimental conditions tested.

Here, we focus on linear pattern classification, since the performance of nonlinear classifiers applied to a brain region could potentially reflect computations performed by the classifier, rather than by brain itself (Kamitani & Tong 2005). For example, if one were to apply sufficiently complex nonlinear classifiers to the patterns of activity observed on the retina, it would be possible to construct the functional equivalent of receptive fields with position-invariant tuning to say visual orientation, curved lines, sharp corners, or even a smiley face cartoon of Bart Simpson, despite the lack of any such pattern detectors in the human retina. All brain processes essentially reflect a series of nonlinear computations; therefore, to characterize the information processed by a brain region, we believe it is important to avoid adding additional nonlinear steps.

The reliability of linear classification performance depends on several factors: i) the degree of separation between the two classes of data samples (i.e., pattern separability or signal-to-noise ratio), ii) the number of data samples available for analysis, since having more samples will allow for better estimation of the optimal classification hyperplane, iii) the choice of classification algorithm and its suitability for the data set to be analyzed (Misaki et al 2010), and iv) the voxels used for pattern analysis. Adding more voxels should lead to better classification performance if those voxels contain some relevant information that can be used to better distinguish between the two conditions. However, if these additional voxels are uninformative, they may simply add noise or unwanted variability to the activity patterns and could thereby impair classification performance (Yamashita et al 2008).

REVIEW OF fMRI STUDIES

Decoding Visual Features

In their original study of orientation decoding, Kamitani & Tong (2005) found that activity patterns in early visual areas could predict which of several oriented gratings was being viewed with remarkable accuracy (Figure 2a). How was this possible, given that BOLD responses were sampled from the visual cortex using 3mm-wide voxels whereas orientation columns are organized at submillimeter spatial scales (Obermayer & Blasdel 1993; Yacoub et al 2008)? The authors performed simulations to show that random local variations in cortical organization could lead to weak orientation biases in individual voxels. By pooling the information available from many independent voxels, a pattern classifier could achieve robust predictions of what orientation was being presented in the visual field. In subsequent work, high-resolution functional imaging studies of the cat and human visual cortices have provided support for this hypothesis (Swisher et al 2010). These experiments show that orientation information exists at multiple spatial scales, extending from that of submillimeter cortical columns to several millimeters across the cortex (Figure 2b). In effect, variability in columnar organization at a submillimeter scale appears to lead to modest feature biases at coarser spatial scales on the order of millimeters. It should be noted that studies find the presence of some global preference for orientations radiating outward from the fovea as well (Freeman et al 2011; Sasaki et al 2006), but when such radial biases are controlled for, substantial orientation information can still be extracted from the visual cortex (Harrison & Tong 2009; Mannion et al 2009).

These orientation decoding studies suggest that pattern analysis can be used to detect signals of columnar origin, by pooling weakly feature-selective signals that can be found at the scale of millimeters, presumably due to variability in the organization of the columns. Thus, fMRI pattern analysis could be used to reveal hidden signals originating from fine-scale cortical columns, which would otherwise be difficult or impossible to isolate with non-invasive imaging. Previously, researchers had to rely on fMRI measures of visual adaptation to assess the feature selectivity of responses in the human visual cortex (Boynton & Finney 2003; Engel & Furmanski 2001).

This decoding approach has been used to investigate cortical responses to many basic visual features. Studies have revealed how the human visual system responds selectively to motion direction (Kamitani & Tong 2006), color (Brouwer & Heeger 2009; Goddard et al 2010; Sumner et al 2008), eye-of-origin information (Haynes et al 2005; Shmuel et al 2010) and binocular disparity (Preston et al 2008). The reliability of feature decoding depends on the strength of the sensory signal; for example, the orientation of high-contrast gratings can be decoded more readily than low-contrast gratings (Smith et al 2011). Moreover, the amplitude of the stimulus-driven BOLD response serves as a good predictor of how much feature-selective information can be extracted from the detailed pattern of activity found in a given visual area. Pattern classification has also revealed sensitivity to more complex visual features. For example, sensitivity to orientations defined by motion boundaries and by illusory contours has been found in early visual areas, including the human primary visual cortex (Clifford et al 2009). It has also been used to show that motion patterns that are more difficult to see, namely second-order texture-defined motion, lead to similar direction-selective patterns of activity in the human visual cortex as basic first-order motion (Hong et al 2012).

The feature decoding approach has also been used to test for selectivity to conjunctions of features (Seymour et al 2009; 2010). For example, Seymour et al. (2009) tested for sensitivity to conjunctions of color and motion, by presenting observers with compound displays consisting of red dots moving clockwise overlapping with green dots moving counterclockwise, or green dots moving clockwise paired with red dots moving counterclockwise. Activity patterns in early visual areas could discriminate between these different combinations of color and motion, implying that these areas contain neurons sensitive to the conjunction of these features. These findings inform current theories of perceptual binding, which have debated whether top-down attentional processes are required to represent conjunctions of features (Treisman 1996).

What are the underlying neural sources of these feature-selective responses in the human visual cortex? In the case of orientation or eye-of-origin signals, these feature-selective responses appear to reflect local biases in columnar organization to a considerable extent (Shmuel et al 2010; Swisher et al 2010). In other cases, feature selectivity might reflect random variations in the distribution of feature-selective neurons (Kamitani & Tong 2006) or more global biases such as a preference for radial patterns or radial motions across the retinotopic visual cortex (Clifford et al 2009; Sasaki et al 2006). For example, optical imaging has revealed the presence of ocular dominance columns, orientation columns, and color-sensitive blobs in the primary visual cortex (V1) of monkeys, but no evidence of

direction-selective columns (Lu et al 2010). Nonetheless, it is possible to decode strong direction-selective responses from human V1 (Kamitani & Tong 2006). Multiple factors can contribute to the spatial distribution of these feature preferences in the cortex, and these factors could have a strong impact on the efficacy of fMRI pattern analysis. In many cases, future studies using high-resolution fMRI in humans or optical imaging in animals will be required to map the feature-selective properties of the visual cortex.

Decoding Visual Perception

In the study by Kamitani & Tong (2005), a major goal was to extend the pattern classification approach from the problem of brain reading to that of mind reading, which had not been demonstrated before. They reported the results of a visual mind-reading experiment, showing that it was possible to decode whether an observer was covertly attending to one set of oriented lines or the other when viewing an ambiguous plaid display. Activity patterns in early visual areas (V1–V4) allowed for reliable prediction of the observer's attentional state (~80% accuracy). Moreover, decoding of the attended orientation was successful even in V1 alone, indicating that feature-based attention can bias orientation processing at the earliest possible cortical site.

Encouraged by these findings, several research groups began to pursue fMRI pattern classification methods to investigate the neural underpinnings of subjective perceptual and cognitive states. Haynes & Rees (2005b) showed that fMRI pattern classification can effectively decode which of two stimuli are perceptually dominant during binocular rivalry, with perceptual alternations occurring every several seconds. Similarly, they found that orientation-selective responses were disrupted by backward visual masking, although a small amount of orientation information could still be detected in V1 for unseen visual orientations (Haynes & Rees 2005a). Perhaps most striking, they were able to apply these methods to extract monocular responses in the lateral geniculate nucleus (Figure 3), and showed that binocular rivalry leads to modulations at this very early site of visual processing (Haynes et al 2005; Wunderlich et al 2005). This latter study provided novel evidence to inform neural models of binocular rivalry (Blake & Logothetis 2002; Tong et al 2006). Other research groups demonstrated that the perception of ambiguous motion displays could be decoded at greater than chance levels from human motion area MT+ and other dorsal visual areas (Brouwer & van Ee 2007; Serences & Boynton 2007b).

A intriguing study by Scolari & Serences (2010) revealed that these feature-selective responses can also be linked to the accuracy of behavioral performance. The researchers first characterized the very modest orientation preference of every voxel in the visual cortex. Next, they tested whether voxel responses to a particular orientation might be boosted on trials in which observers correctly discriminate a small change in visual orientation, as compared to incorrect trials. When observers correctly discriminated a change in orientation centered around say 45°, responses in V1 were not enhanced for voxels tuned specifically to 45°; instead, they were enhanced for voxels that preferred neighboring orientations (~10° and 80°). This counterintuitive result is predicted by models of optimal visual coding, which propose that discrimination performance will be most improved by enhancing neighboring off-channel responses.

Decoding Visual Objects

The pioneering work by Haxby et al. (2001) suggested that categorical information about objects is represented in a distributed manner throughout the ventral temporal lobe. Activity patterns in this region could accurately discriminate between multiple object categories, even when the most strongly category-selective voxels were removed from the analysis (but see also Spiridon & Kanwisher 2002). In effect, the authors could perform “virtual lesions” on these activity patterns, and thereby revealed the distributed nature of object information. Curiously however, subsequent studies found that activity patterns in low-level visual areas could outperform high-level object areas at telling apart viewed objects (Cox & Savoy 2003). How was this possible, given that low-level visual areas are primarily tuned to the retinotopic position of low-level features? These results indicate that the images in each object category differed in some of their low-level properties, and that these low-level confounds can persist even when multiple images are shown in a stimulus block. Although low-level confounds can be reduced by manipulations of object size or 3D vantage point, they might not be eliminated, as indicated by the fact that early visual areas can still classify an object across changes in size and 3D viewpoint (Eger et al 2008).

These findings reveal a core challenge for fMRI decoding studies. Pattern classifiers are quite powerful and will try to leverage any discriminating information that is present in brain activity patterns. Even if a brain area can distinguish between certain object images, how can one go further to show that a brain area is genuinely sensitive to object properties and not simply the low-level features of those objects?

Work by Kanwisher and colleagues has provided several lines of evidence linking the activity patterns in object-selective areas to object perception. In a study of backward visual masking, they found that activity patterns in object-selective areas were severely disrupted on trials in which the observer failed to recognize a briefly presented target (Williams et al 2007). By contrast, activity patterns remained stable in early visual areas, despite the participant’s impaired performance. Another study manipulated the physical similarity of simple 2D shapes and estimated the perceptual similarity between pairs of stimuli based on the confusion errors that participants’ made with visually masked stimulus presentations. Multivariate pattern analysis revealed a striking dissociation: activity patterns in the lateral occipital area reflected the physical similarity of the images, whereas those in the ventral temporal cortex correlated with perceptual similarity (Haushofer et al 2008). However, other studies have found that activity patterns in the lateral occipital area reflect the perceived 3D shape of “bumps” and “dimples” conveyed by shape-from-shading cues, even when the physical image is greatly altered by changes in the source of illumination (Gerardin et al 2010).

Activity patterns in the lateral occipital and ventral temporal cortices show strong position-invariant selectivity, and remain quite stable for a particular object across changes in retinal position (Schwarzlose et al 2008). However, these areas show some evidence of position selectivity as well. Face- and body-selective areas can better discriminate between pictures of different body parts if those parts are presented at a familiar location (Chan et al 2010). For example, a front-view image of a person’s right shoulder will lead to more reliable activity patterns if the stimulus appears to the left of fixation, as it would if one were looking

at the head or chest, than if it appears to the right of fixation. It is also possible to decode the retinotopic position of an object from activity patterns in high-level object areas. Moreover, perceptual illusions that lead to shifts in apparent position are better predicted by the position information contained in the activity patterns in high-level object areas than those in the early visual areas (Fischer et al 2011).

When objects are subliminally presented to an observer, activity in object-selective areas is greatly attenuated, but somewhat greater than chance-level decoding is still possible, indicating the presence of some unconscious visual information in these areas (Sterzer et al 2008). Subliminal stimuli also appear to evoke more variable patterns of activity in object-selective areas across repeated presentations, which partly accounts for the poorer decoding of subliminal stimuli (Schurger et al 2010).

A major challenge in object recognition concerns the ability to distinguish a particular exemplar from other items in the same category. In an ambitious study, Kriegeskorte and colleagues (2008) presented 92 images of different real-world objects, and assessed which images tended to evoke more similar patterns of activity. Images of animate and inanimate stimuli led to broadly distinctive patterns of activity in the human ventral temporal cortex, and a similar animate/inanimate distinction was observed when analyzing neuronal activity patterns obtained from single-unit recordings in monkeys (Kiani et al 2007). This study also found evidence of exemplar-specific activity. Activity patterns in the human inferotemporal cortex were better at discriminating between images of different human faces than between the faces of non-human primates, while a trend towards the opposite pattern of results was observed in the monkey data.

Attempts to isolate exemplar-specific information from small cortical regions have met with limited success, with decoding performance reaching levels just slightly greater than chance (Kaul et al 2011; Kriegeskorte et al 2007). When large portions of the ventral temporal cortex are pooled for analysis, then considerably better decoding of specific faces can be obtained (Kriegeskorte et al 2008; Natu et al 2010). However, it remains to be seen whether these large-scale distributed representations are truly important for representing individual faces, or whether the diverse shape codes throughout this region simply provide more information for the classifier to capitalize upon when performing these subtle discriminations. Single-unit recordings from isolated face-selective patches in the monkey indicate that a cluster of a few hundred neighboring neurons can provide remarkably detailed information for distinguishing between individual faces (Tsao et al 2006; Freiwald et al 2009; Freiwald & Tsao 2010). However, current fMRI technology cannot readily isolate information at this level of detail.

Identifying and Reconstructing Novel Visual Scenes

Decoding algorithms can classify a person’s brain state as belonging to the same category as a previously recorded brain state, but these methods lack the flexibility to identify novel brain states. To address this, Kay, Gallant and colleagues (2008) devised a visual encoding model to predict how early visual areas should respond to novel pictures of complex real-world scenes. First, they presented 1750 different images to observers, and from the resulting fMRI data, they could characterize the response preferences of each voxel in visual

cortex, specifying its preference for particular retinotopic locations, spatial frequencies, and orientations. When the observers were later shown a new set of 120 pictures, the model predicted how these voxels should respond to each new image. By comparing the predicted and actual patterns of activity, the model correctly identified 110 out of 120 test images for one participant. In a follow-up experiment, the observer was tested with 1000 new images, of which 820 were correctly identified.

This level of identification performance is akin to *Science Fiction Story #1*, identifying which painting the participant is viewing at the *Musée d'Orsay*. An even loftier goal would be to *reconstruct* the painting, using only the brain activity that results from viewing that work of art. An early attempt at reconstruction met with some success at reconstructing fragments of simple shapes (Thirion et al 2006). In a more recent fMRI study, observers were presented with hundreds of different random patterns of flickering checks placed within a 10×10 square grid, and pattern analysis was used to predict whether any given tile of the grid was flickering or not (Miyawaki et al 2008). Using this model, the authors could effectively reconstruct novel stimuli shown to the participant, including simple shapes and letters (Figure 4a). Moreover, the authors could reconstruct the viewed stimulus from single brain volumes to show how this information evolved over the time course of the BOLD response (Figure 4b). Extending the work of Kay et al (2008), Naselaris et al (2009) attempted to reconstruct complex natural scenes using local-feature models, and could capture regions of high contrast and some of the "blurry" low spatial frequency components of the image (Figure 4c). By incorporating the category-specific information available in higher-level object areas, they could also select an image (from a set of 6 million possible images) that best matched the visual features and category properties evoked by the original viewed image (*natural image prior* condition).

Decoding Top-Down Attentional Processes

The ability to decode feature-selective responses has helped advance the study of visual attention, and in particular, feature-based attention. Kamitani & Tong (2005) showed that the activity patterns evoked by single orientations can predict which of two overlapping orientations is being attended by an observer. Similar results were obtained in studies of attention to overlapping motion stimuli (Kamitani & Tong 2006). These findings indicate that top-down attention can bias the strength of feature-selective responses in early visual areas, consistent with models of early attentional selection. Serences & Boynton (2007a) demonstrated that attending to one of two overlapping sets of moving dots leads to biased direction-selective responses not only at the site of the attended stimulus, but also in unstimulated portions of the visual field. Such spatial spreading of feature-based attention is consistent with neurophysiological studies in monkeys (Treue & Maunsell 1996). A recent study found that spatial and feature-based attention can lead to distinct effects in the visual cortex (Jehee et al 2011). When spatial attention was directed to one of two laterally presented gratings, overall BOLD activity was enhanced for the attended stimulus and yet the orientation-selective component of these responses improved only when observers focused on discriminating the orientation of the stimulus, rather than its contrast. This may suggest that enhanced processing of a specific visual feature may depend more on feature-

based attention than on spatial attention (Jehee et al 2011; but see also Saproo & Serences 2010).

Recent studies have also investigated the possible top-down sources of these attentional signals. Activity patterns in posterior parietal areas and the frontal eye fields contain reliable information about whether participants are attending to features or spatial locations (Greenberg et al 2010), and can even discriminate which of two features or locations is being attended (Liu et al 2011). These parietal and frontal areas could serve as plausible sources of attentional feedback to early visual areas.

Multivariate pattern analysis has also been used to quantify the extent to which spatial attention can bias activity in category-selective object areas, for example when face and house stimuli are simultaneously presented in different locations (Reddy et al 2009). When observers view overlapping face-house stimuli, it is possible to decode the focus of object-based attention from activity patterns in high-level object areas as well as in early visual areas, indicating that top-down feedback serves to enhance the local visual features belonging to the attended object (Cohen & Tong, *under review*). Interestingly, attending to objects in the periphery leads to pattern-specific bias effects in the foveal representation of early visual areas, perhaps suggesting some type of remapping of visual information or reliance on foveal representations to recognize peripheral stimuli (Williams et al 2008). Pattern classification has also been used to investigate visual search for objects in complex scenes. Activity patterns in the lateral occipital complex can reveal what object category participants are actively searching for, as well as those occasions when the target object briefly appears at an attended or unattended location (Peelen et al 2009). Overall, fMRI pattern classification has greatly expanded the possibilities for studies of visual attention by providing an effective tool to measure attention-specific signals in multiple brain areas, including parietal and frontal areas.

Decoding Imagery and Working Memory

In an early fMRI study of mental imagery, (O'Craven & Kanwisher 2000) showed that it was possible to predict with 85% accuracy whether a person was imagining a famous face or place by inspecting the strength of activity in the fusiform face area and parahippocampal place area. A more recent study used multivoxel pattern analysis, and found that activity patterns in the ventral temporal cortex could predict whether participants were imagining famous faces, famous buildings, tools or food items with reasonable accuracy (Reddy et al 2010). Similar results have been reported in studies of working memory for faces, places, and common objects (Lewis-Peacock & Postle 2008). It is also possible to decode the imagery of simple shapes, such as an 'X' or 'O', from these object-sensitive visual areas (Stokes et al 2009). In these studies, the activity patterns observed during imagery or working memory were very similar to those observed during perception, consistent with perception-based theories of imagery (Kosslyn et al 2001). Interestingly, it is also possible to distinguish silent clips of movies that imply distinctive sounds (e.g., howling dog, violin being played) from activity patterns in the auditory cortex, presumably because these visual stimuli elicit spontaneous auditory imagery (Meyer et al 2010).

Although early visual areas have been implicated in visual imagery (Kosslyn & Thompson 2003), these areas typically show little evidence of sustained BOLD activity during visual working memory tasks (Offen et al 2008). However, recent fMRI decoding studies have provided novel evidence to suggest that early visual areas are important for retaining visually precise information about visual features (Harrison & Tong 2009; Serences et al 2009). Serences and colleagues cued participants in advance to remember either the color or orientation of a grating, and after a 10s delay, presented a second grating to evaluate working memory for the cued feature. They found that activity patterns in V1 allowed for prediction of the task-relevant feature (~60% accuracy) but not for the task-irrelevant features; information in extrastriate visual areas proved unreliable. Harrison & Tong used a postcuing method to isolate memory-specific activity by presenting two near-orthogonal gratings at the beginning of each trial, followed by a cue indicating which orientation to retain in working memory (see Figure 5a for timeline of trial events). Activity patterns in areas V1–V4 allowed for reliable decoding of the remembered orientation (mean accuracy of 83%) and reliable working memory information was found in each visual area, including V1 (~70–75% accuracy). Moreover, they found evidence of a striking dissociation between the overall amplitude of BOLD activity and the decoded information contained at individual fMRI time points. Whereas BOLD activity fell over time (Figure 5a), information about the remembered grating was sustained throughout the delay period (Figure 5b). In half of their participants, activity in V1 fell to baseline levels, equivalent to viewing a blank screen, yet decoding of the retained orientation proved as effective for these participants as for those who showed significantly elevated activity late in the delay period. These results suggest that visually precise information can be retained in early visual areas with very little overall change in metabolic activity, due to subtle shifts in the patterns of activity in these areas.

Decoding Episodic Memory

Although long-term memories are stored via modified synaptic connections in the hippocampus and cortex in their inactive state, it is possible to decode these memories when they are actively recalled or reinstated by the participant (for an in-depth review, see Rissman & Wagner, this issue). Polyn et al (2005) had participants study images of famous faces, famous places and common objects in the MRI scanner, and trained pattern classifiers on whole-brain activity to discriminate between these categories. When participants were later asked to freely recall these items, the classifier readily tracked the category that was being recalled from memory (Figure 5c). Remarkably, this category-selective activity emerged several seconds before participants switched to reporting items from a new category, suggesting that this categorical information might have served as a reinstated contextual cue to facilitate memory retrieval (Howard & Kahana 1999; Tulving & Thomson 1973). Evidence of contextual reinstatement has even been observed when participants fail to recall the studied context (Johnson et al 2009). Whole-brain activity patterns could predict which of three of different encoding tasks was performed on an item at study, based on the reinstated patterns of activity that were later observed during a recognition memory test. Task-specific patterns of activity were found for correctly recognized items, and this proved true even for items that were rated as merely familiar, despite participants' reports that they could not recollect any details surrounding the time of studying the target item. These findings argue against proposed dissociations between conscious recollection and feelings of

familiarity, and further suggest that cortical reinstatement of the studied context might not be sufficient for experiencing explicit recollection (McDuff et al 2009). Decoding can also reliably predict whether an item will be judged as old or new. When participants performed a recognition memory task involving faces, multiple brain regions responded more strongly to items judged as old than new, including the lateral and medial prefrontal cortex and posterior parietal cortex (Rissman et al 2010). The pooled information from these regions could reliably distinguish between correctly recognized or correctly rejected items with 83% mean accuracy, but failed to distinguish missed items from correctly rejected items. Explicit performance of these recognition memory judgments was necessary for decoding, as the classifier could no longer distinguish between old and new items when participants instead performed a gender discrimination task. The studies described above reveal how fMRI pattern analysis can provide a powerful tool for investigating item-specific memory processing at the time of study and test, and how such data can be used to address prevalent theories of memory function.

Decoding can also be used to isolate content-specific information from fine-scale activity patterns in the human hippocampus. After participants learn the spatial layout of a virtual environment, decoding applied to hippocampal activity can reveal some reliable information about the participant's current location in that learned environment (Morgan et al 2011; Rodriguez 2010). It has also been shown that activity patterns in the hippocampus can predict which of three short movie clips a participant is engaged in recalling from episodic memory (Chadwick et al 2010). Although decoding performance for the hippocampus was modest (~60% accuracy), activity patterns in this region were found to perform significantly better than neighboring regions of the entorhinal cortex or the posterior parahippocampal gyrus. The ability to target specific episodic memories in the hippocampus may greatly extend the possibilities for future studies of human long-term memory.

Extracting Semantic Knowledge

Semantic knowledge is fundamentally multidimensional and often multimodal, consisting of both specific sensory-motor associations and more abstracted knowledge. For example, we know that a rose is usually red, has soft petals but sharp thorns, smells sweetly fragrant, and that the flowers of this plant make for an excellent gift on Valentine's Day. Given the multidimensional nature of semantic information, multivariate pattern analysis might be well suited to probe its neural bases.

An early fMRI study demonstrated that it was possible to decode whether participants were viewing words belonging to 1 of 12 possible semantic categories, such as four-legged animals, fish, tools, or dwellings (Mitchell et al 2003). Subsequent studies have consistently found that animate and inanimate visual objects lead to highly differentiated patterns of activity in the ventral temporal cortex (Kriegeskorte et al 2008; Naselaris et al 2009). Remarkably, people who have been blind since birth exhibit a similar animate/inanimate distinction in the ventral temporal cortex when presented with tactile objects (Mahon et al 2009; Pietrini et al 2004), leading to the proposal that this semantic differentiation might be innately determined rather than driven by visual experience (Mahon & Caramazza 2011).

How might one characterize the broader semantic organization of the brain or predict how the brain might respond to any item based on its many semantic properties? (Mitchell et al 2008) developed a multidimensional semantic feature model to address this issue. They tried to predict brain responses to novel nouns by first quantifying how strongly these nouns were associated with a basis set of semantic features, consisting of 25 verbs (e.g., see, hear, touch, taste, smell, eat, run). In essence, these semantic features served as intermediate variables to map between novel stimuli and predicted brain activity (cf. Kay et al 2008). The strength of the semantic association between any noun and these verbs could be estimated based on their frequency of co-occurrence, from analyzing a trillion-word text corpus provided by Google Inc. Using fMRI activity patterns elicited by 60 different nouns, the authors characterized the distinct patterns of activity associated with each verb, and could then predict brain responses to novel nouns by assuming that the resulting pattern of activity should reflect a weighted sum of the noun's association to each of the verbs. Using this method, Mitchell et al. could predict which of two nouns (excluded from the training set) was being viewed with 77% accuracy, and could even distinguish between two nouns belonging to the same semantic category with 62% accuracy. The activity patterns for particular verbs often revealed strong sensorimotor associations. For example, "eat" predicted positive activity in frontal regions associated with mouth movements and taste, whereas "run" predicted activity in the superior temporal sulcus associated with the perception of biological motion. These findings are quite consistent with the predictions of neural network models of semantic processing, in which specific items are linked to multiple associated features through learning, and semantically related items are represented by more similar patterns of activated features (McClelland & Rogers 2003).

Decoding has also been applied to other domains of knowledge such as numerical processing. One study found that activity patterns in the parietal cortex reflected not only spatial attention directed to the left or right side of space, this spatial bias could be used to predict whether participants were engaged in a subtraction or addition task (Knops et al 2009). Another study found that activity patterns in the parietal cortex could distinguish between different numbers, whether conveyed by digit symbols or dot patterns (Eger et al 2009). In general, these studies are consistent with the proposal that number representations are strongly associated with the parietal lobe and may be represented according to an implicit spatial representation of a number line (Hubbard et al 2005).

Decoding Phonological Representations and Language Processing

Some recent studies have begun to use fMRI decoding methods to investigate the neural underpinnings of phonological and language processing. In one study, participants were presented with audio clips of three different speakers uttering each of three different vowel sounds (Formisano et al 2008). Activity patterns in the auditory cortex could successfully discriminate which vowel was heard, even when the classifier was tested on a voice not included in the training set. Likewise, pattern classifiers could identify the speaker at above-chance levels, even when tested with vowels not included in the training set. Another study showed that activity patterns in the auditory cortex can distinguish between normal speech and temporally reordered versions of these stimuli, implying sensitivity to speech-specific content (Abrams et al 2011).

Another fruitful approach has been to investigate the role of experience in the development of phonological representations. An analysis of activity patterns in the auditory cortex revealed better discrimination of the syllables /ra/ or /la/ in native English speakers than in Japanese participants who often have difficulty distinguishing between these phonemes (Raizada et al 2010). Moreover, the authors found evidence of a correlation within each group, between an individual's decoding performance and his or her behavioral ability to distinguish between these phonemes, suggesting that fMRI decoding may be sensitive to individual differences in language processing. A recent study of reading ability provides further evidence for this view (Hoeft et al 2011). The authors instructed children with dyslexia to perform a phonological processing task in the scanner, and later assessed whether or not their reading skills had improved two and a half years later. Although purely behavioral measures taken in the first session failed to predict which children would improve in reading skills over time, a pattern classifier trained on the whole-brain data was able to predict improvement with over 90% accuracy. These results raise the exciting possibility of using fMRI pattern analysis for diagnostic purposes with respect to language processing.

Decoding Decisions in the Brain

Decoding has revealed that it is possible to predict the decisions that people are likely to make, even in advance of their actual choices. For example, activity in the anterior cingulate cortex, medial prefrontal cortex, and the ventral striatum is predictive of the participant's choices in a reward-learning paradigm (Hampton & O'Doherty J 2007). Here, one of two stimuli is associated with a higher likelihood of reward and the other with a lower likelihood, but these reward probabilities are reversed at unpredictable times. Activity in these areas is highly predictive of whether a participant will switch their choice of stimulus on a given trial, and activity on the trial prior to a switch is also somewhat predictive, indicating an accrual of information over time regarding whether the current regime should be preferred or not. Such valuation responses can also be observed in the insula and medial prefrontal cortex for unattended stimuli, and these decoded responses correspond quite well to the participants' valuation of that item, such as a particular model of car (Tusche et al 2010). fMRI decoding can even predict participants' choices of real-world products at greater-than-chance levels. In these experiments, participants were offered the opportunity to purchase or decline to purchase a variety of discounted items ranging in value from $8–80, with the foreknowledge that two of their purchase choices would be realized at the end of the experiment (Knutson et al 2007). In studies of arbitrary decisions, such as deciding to press a button with one's left or right hand at an arbitrary time, participants show evidence of preparatory activity in motor and supplementary motor areas a few seconds in advance of their action. Remarkably, however, a small but statistically reliable bias in activity can be observed in the frontopolar cortex up to 10 seconds prior to the participant's response, suggesting some form of preconscious bias in the decision making process (Soon et al 2008).

CONCEPTUAL AND METHODOLOGICAL ISSUES

Whenever a new methodology is developed, important conceptual and methodological issues can emerge regarding how the data should be analyzed, interpreted and understood. Pattern

classification algorithms are statistically powerful and quite robust. However, these very strengths can pose a challenge, as the algorithms are designed to leverage whatever information is potentially available in a brain region to make better predictions about a stimulus, experimental condition, mental state, or behavioral response. An example of unwanted leveraging was apparent in one of the reported results of the 2006 Pittsburgh Brain Competition (http://pbc.lrdc.pitt.edu/), an open competition that was designed to challenge research groups to develop state-of-the-art analytic methods for the purposes of brain reading and mind reading. This competition assessed the accuracy of decoding the presence of particular actors, objects, spatial locations, and periods of humor from the time series of fMRI data collected while participants watched episodes of the TV series "Home Improvement". To decode scenes containing humorous events, it turned out that the ventricles proved to be the most informative region of the brain—this high-contrast region in the functional images tended to jiggle whenever the participant felt an urge to laugh. Despite the remarkable accuracy of decoding periods of mirth from this region, it would clearly be wrong to conclude that this brain structure has a functional role in the cognitive processing of humorous information. If the accuracy of decoding is not sufficient for establishing function, then how can one determine precisely what information is processed by a brain region? Below, we consider these and other conceptual and methodological issues.

What Is Being Decoded?

A long-standing problem in fMRI research concerns the potential pitfalls of *reverse inference*. As an example, it is well established that the human amygdala responds more strongly to fear-related stimuli than to neutral stimuli, but it does not logically follow that if the amygdala is more active in a given situation that the person is necessarily experiencing fear (Adolphs 2010; Phelps 2006). If the amygdala's response varies along other dimensions as well, such as the emotional intensity, ambiguity or predictive value of a stimulus, then it will be difficult to make strong inferences from the level of amygdala activity alone.

A conceptually related problem emerges in fMRI decoding studies, when one identifies a brain region that can reliably discriminate between two particular sensory stimuli or two cognitive tasks. For example, Haxby et al. (2001) showed that activity patterns in the human ventral temporal cortex were reliably different when participants viewed images of different object categories. The authors interpreted this decoding result to suggest that the ventral temporal object areas are sensitive to complex object properties. However, subsequent studies revealed that early visual areas could discriminate between the object categories just as well as or better than the high-level object areas, because of the pervasiveness of low-level differences between the object categories (Cox & Savoy 2003). Therefore, successful decoding of a particular property from a brain region, such as object category, does not necessarily indicate that the region in question is truly selective for that property. The inferences one can make with multivariate pattern analysis still depend on strong experimental design, and in many cases multiple experiments may be needed to rule out potential confounding factors.

One approach for determining the functional relevance of a particular brain area is to test for links between behavioral performance and decoding performance. For example, if one

compares correct versus incorrect trials in a fine-grained orientation discrimination task, greater activity in the primary visual cortex is found specifically in those voxels tuned to orientations neighboring the target orientation (Scolari & Serences 2010). Month-long perceptual training at discriminating a specific orientation in the visual field can also lead to more reliable orientation-selective activity patterns in human V1, specifically around the trained orientation (Jehee et al., *submitted*). Decoding of object-specific information from the lateral occipital complex is much better on trials with successful than unsuccessful recognition (Williams et al 2007). Related studies have found that functional activity patterns in the ventral temporal object areas are more reliable and reproducible when a stimulus can be consciously perceived than when it is subliminally presented (Schurger et al 2010). Interestingly, when participants must study a list of items on multiple occasions, items that evoke more similar activity patterns across repeated presentations are also more likely to be remembered (Xue et al 2010).

Because of the high-dimensional nature of visual input, it is possible to investigate the similarity of cortical activity patterns across a variety of stimulus conditions to assess the properties they might be attuned to. For example, similar orientations evoke more similar activity patterns in early visual areas (Kamitani & Tong 2005), and similar colors have been found to do so in visual area V4 (Brouwer & Heeger 2009). However, the similarity relationships of responses to objects are quite different in early visual areas and high-level object areas, with the object areas exhibiting a sharp distinction in their activity patterns for animate and inanimate objects (Kriegeskorte et al 2008; Naselaris et al 2009). Studies of olfactory perception have revealed comparable findings in the posterior piriform cortex, with more similar odors leading to more similar patterns of fMRI activity (Howard et al 2009). Thus, if neural activity patterns share the similarity structure of perceptual judgments, this can provide strong evidence to implicate the functional role of a brain area.

One can further investigate the functional tuning properties of a brain area by assessing generalization performance: do the activity patterns observed in a brain area generalize to very different stimulus conditions or behavioral tasks? In Harrison & Tong's (2009) study of visual working memory, the authors trained a classifier on visual cortical activity patterns elicited by unattended gratings, and tested whether these stimulus-driven responses might be able to predict which of two orientations was being maintained in working memory while participants viewed a blank screen. Successful generalization was found despite the differences in both stimulus and task across the experiments, thereby strengthening the inference that orientation-specific information was being maintained in the visual cortex during the working memory task. In a study of auditory perception, classifiers trained using phonemes pronounced by one speaker could successfully generalize to the corresponding phonemes spoken by another speaker, despite changes in the auditory frequency content (Formisano et al 2008). Perhaps the most rigorous test of generalization performance comes from demonstrations of the ability to predict brain responses to novel stimuli, as has been shown by Kay & Gallant's visual encoding model and Mitchell et al.'s semantic encoding model (Kay et al 2008; Mitchell et al 2008). Successful generalization can be an effective tool for ruling out potential low-level stimulus confounds or task-related factors.

In studies of high-level cognition, isolating the specific function of a brain area may be more challenging if the experimental design focuses on discriminating between two cognitive tasks. When participants perform cognitive tasks differing in the stimuli, task demands, and behavioral judgments required, almost the entire cerebral cortex can show evidence of reliable discriminating activity (Poldrack et al 2009). Differential activity can result from many factors, including differences in low-level sensory stimulation, working memory load, language demands, or the degree of response inhibition required for the task. Even when two tasks are quite closely matched, such as performing addition or subtraction (Haynes et al 2007) or directing attention to features or spatial locations (Greenberg et al 2010), it is important to consider potential confounding factors. If one task is slightly more difficult or requires a bit more processing time for a given participant, then larger or longer fMRI amplitudes could occur on those trials, which could allow decoding to exceed chance-level performance. This potential confound has sometimes been addressed by performing decoding on the average amplitude of activity in a brain region, to see if overall activity is predictive or whether more fine-grained information is needed for reliable decoding. Another approach might be to attempt to assess decoding of fast vs. slow reaction times using the same brain region, and to test whether these activity patterns resemble those that distinguish the two tasks.

Where in the Brain to Decode From?

Many fMRI decoding studies have focused on the human visual system, which contains many well-defined visual areas. In addition, it is common to map the particular region of visual space that will be stimulated in an experiment, so that only the corresponding voxels in the retinotopic visual cortex are used for decoding analysis. There are several advantages to applying pattern analysis to well-defined functional areas. First, localization of function is possible, and the information contained in each functional region can be independently assessed and compared to other regions. Second, there is reduced concern that decoding performance might reflect information combined across functionally distinct areas. Finally, decoding performance can be compared to other known functional properties of that brain area to ask whether the results seem reasonable and readily interpretable. Focused investigations of the human hippocampus have also benefitted from having a targeted anatomical locus (Chadwick et al 2010; Hassabis et al 2009).

In studies of higher-level cognition, predefined regions of interest usually are not available and multiple distributed brain areas might be involved in the cognitive task. Many of these studies rely on decoding of whole-brain activity, sometimes first selecting the most active voxels in the task or applying a method to reduce the dimensionality of the data (e.g., principal components analysis) prior to classification analysis. (When selecting a subset of voxels prior to the decoding analysis, it is important to ensure that the selection process is independent of the property to be decoded, so it will not bias decoding performance to be better than it should (Kriegeskorte et al 2009).) The advantage of the whole-brain approach lies in its ability to reveal a majority of the information available throughout the brain. Moreover, it is possible to inspect the pattern of “weights” in the classifier and to project these onto the cortex to reveal how this information is distributed throughout the brain. For example, Polyn and colleagues (2005) found that that fusiform face area was one of the

regions most active during the free recall of famous faces whereas the parahippocampal place area and retrosplenial cortex were most active during the recall of famous places. Thus, decoding of whole brain activity can reveal what information is present in the brain and where in the brain such information is most densely concentrated.

However, classification analysis implicitly assumes a “readout mechanism”, in which relative differences between the strengths of particular brain signals are calculated and leveraged to compute useful information. It is not clear whether the brain is actually comparing or combining the neural signals that are being analyzed by the classifier, especially when information from distinct brain regions are combined. For example, a semantic model might find that the word “rose” leads to whole brain activity that is well predicted by the patterns associated with “smell”, “plants”, and “seeing” vivid colors such as red. Should each of the respective components of this activity be considered part of a single unified representation or several independent components that are being unified outside of the brain by the classifier (Mahon & Caramazza 2009; Mitchell et al 2008)? This distinction can be made more vivid with a somewhat different example. Assume it is possible to decode whether something smells *floral* or *citrus* from activity patterns in the olfactory piriform cortex, and it is also possible to decode whether the color *red* or *yellow* is being perceived from the visual cortex. Now, if decoding of whole-brain activity can tell apart a floral-scented red rose from one that smells like lemon or has lemon-colored petals, can it be argued that the brain contains a unified representation of the color and smell of roses? According to a recent fMRI study of perceptual binding (Seymour et al 2009), establishing evidence of a conjoint representation of color and smell would require demonstrating that brain activity patterns can distinguish between a floral-scented red rose paired with a citrus-scented yellow rose as distinct from a citrus-scented red rose paired with a floral-scented yellow rose. This issue also points to a longstanding debate regarding whether the brain relies on modular or distributed representations for information processing (Haxby et al 2001; Op de Beeck et al 2008). Recent fMRI studies indicate that many types of information are distributed quite widely throughout the brain, but that there also exist highly stimulus-selective modules that may form a more local, exclusive network (Moeller et al 2008; Tsao et al 2006).

An alternative to decoding whole-brain activity is to perform a searchlight analysis, in which decoding is iteratively performed on local activity patterns sampled throughout the cortex (Kriegeskorte et al 2006). This typically involves using a moveable searchlight to sample a local “sphere” of voxels (say a $5 \times 5 \times 5$ voxel cubic region) from each point in the cortex. This approach reveals the information contained in local activity patterns, which reduces the extent to which information will be combined across distinct functional areas. A potential concern is that brain signals from disparate areas may sometimes be combined across a sulcus, so this approach could be further strengthened by analyzing activity patterns based on a flattened representation of the cortical surface. A disadvantage of this approach is the need to correct for multiple comparisons for each iteration of the search, which reduces statistical power. For these reasons, searchlight analyses are often combined with group-level statistical analyses to evaluate whether reliable information is consistently found in a particular region of the brain across participants.

At What Spatial Scales of Cortical Representation Is Decoding Most Useful?

Multivoxel pattern analysis may serve different purposes, depending on whether the sought after information resides at fine or coarse spatial scales in the brain. At the finest scale, multivoxel pattern classification may be particularly advantageous at detecting signals arising from variability in the spatial arrangement of cortical columns, which can lead to locally biased signals on the scale of millimeters (Swisher et al 2010). Pattern analysis of fine-scale signals has proven effective not only in the visual cortex but also in high-resolution fMRI studies of the hippocampus (Hassabis et al 2009). Such fine-grained information would otherwise be very difficult or impossible to detect using traditional univariate methods of analysis. At a somewhat coarser scale, pattern classifiers are also very effective at extracting category-selective information from the ventral temporal cortex, which reveals a strong functional organization at spatial scales of several millimeters to centimeters (Haxby et al 2001). These methods can be helpful for pooling distributed information about objects or semantic categories, particularly when there is no single "hotspot" of functional selectivity available in the broad cortical region to be analyzed. Decoding has also been applied to activity patterns of large spatial scale, including whole-brain activity, even when differentially activated regions can be seen using traditional univariate analyses such as statistical parametric mapping. For example, one can attain much better predictions of an observer's near-threshold perceptual judgments regarding fearful versus non-fearful faces by pooling information across multiple activated regions (Pessoa & Padmala 2007). Beyond the benefits of signal averaging, combining signals from multiple regions of interest can be beneficial if each region contains some unique information. Another example of whole-brain decoding comes from a recognition memory study, which compared participants' behavioral performance at old-new judgments with the discriminating performance of the pattern classifier (Rissman et al 2010). Although the patterns picked up by the classifier closely resembled the statistical maps, the decoding analysis revealed a compelling relationship between subjective ratings of memory confidence and differential brain responses to old versus new items on individual trials. These examples illustrate how decoding can be useful when applied at large spatial scales. Nevertheless, interpreting the combined results from disparate brain areas can be challenging, and may warrant careful consideration of exactly *what* is being decoded, as we have described above.

ETHICAL AND SOCIETAL CONSIDERATIONS

What are the potential implications of human neuroimaging and brain-reading technologies as this rapidly growing field continues to advance? Over the last decade, there has been steadily growing interest in *neuroethics*, which focuses on the current and future implications of neuroscience technology on ethics, society and law (Farah 2005; Roskies 2002). Although some had thought these concerns to be premature, the intersection between law and neuroscience (sometimes called *neurolaw*) has rapidly evolved in recent years (Jones & Shen, *submitted*).

In October 2009, Dr. Kent Kiehl appeared at a Chicago court hearing to find out whether the fMRI scans he had collected of Brian Dugan's brain might be admissible as evidence in a

high-profile death penalty case. Dugan, who had already served over 20 years in prison for two other murders, had recently confessed to murdering a 10-year-old girl in 1983, following the discovery of DNA evidence linking him to the crime.

On November 5, 2009, the fMRI scans of a defendant's brain were considered as evidence in the sentencing phase of a murder trial, for what appears to be the first time (Hughes 2010). Dr. Kiehl provided expert testimony, describing the results of two psychiatric interviews and the unusually low levels of activity in several regions of Dugan's brain, similar to many other criminal psychopaths when shown pictures of violent or morally wrong actions (Harenski et al 2010). He pointed to these regions on cartoon drawings of the brain, as the judge had decided that the presentation of actual brain pictures might unduly influence the jury (Weisberg et al 2008). Expert testimony from the prosecution refuted the brain imaging data on two grounds: Dugan's brain might have been very different 26 years ago, and Dr. Kiehl's neuroimaging studies of criminal psychopaths showed average trends in the data and were not designed for individual diagnosis. After less than an hour of deliberation, the jury initially reached a mixed verdict (10 for and 2 against the death penalty), but then asked for more time, switching to a unanimous verdict in favor of the death penalty the next day. Dugan's lawyer noted that although the verdict was unfavorable, Kiehl's testimony "turned it from a slam dunk for the prosecution into a much tougher case".

If courts are primarily concerned that neuroimaging evidence appears unreliable for individual diagnosis, then recent advances in brain classification methods for diagnosing neurological disorders could lead to the increasing prevalence of such evidence in courtrooms. Recent studies have shown that pattern classification algorithms applied to structural MRI scans or functional MRI scans can distinguish between whether an individual is a normal control or a patient suffering from either schizophrenia (Nenadic et al 2009), depression (Craddock et al 2009) or psychopathy (Sato et al 2011), with reported accuracy levels ranging from 80–95%. In the context of a court case, these accuracy levels might be high enough to influence a jury's decision. For example, a diagnosis of paranoid schizophrenia might influence decisions regarding whether a defendant was likely to be psychotic at the time of the crime. Although a diagnosis of psychopathy might be unlikely to affect the determination of whether a defendant should be considered guilty based on his or her actions, such evidence could prove to be an influential mitigating factor during the sentencing phase of the trial. As neuroscience continues to advance our understanding of the neural mechanisms that lead to decisions and actions, neuroscientists and perhaps society more generally may feel motivated to reconsider our traditional definitions of free will and personal responsibility (Greene & Cohen 2004; Roskies 2006; Sapolsky 2004).

Brain classification methods for individual diagnosis could have strong ethical implications in medical settings as well, especially concerning disorders of consciousness. Some patients, who partially recover from coma, are diagnosed as being in a vegetative state if they exhibit periods of wakefulness but appear to lack awareness or any purpose in their motor actions. Despite this apparent lack of awareness, it was recently discovered that some vegetative state patients are capable of voluntarily performing mental imagery tasks (Owen et al 2006). When asked to imagine either playing tennis or walking around a house, differential patterns of activity can be observed in their brains. Recently, this imagery paradigm has been

combined with fMRI decoding to obtain reliable yes/no responses from a patient to questions such as “Is you father’s name Alexander?” (Monti et al 2010). If highly reliable communication can be established with such patients, this could lead to uncharted territories in terms of the ethical and legal considerations regarding, for example, any medical requests made by the patient.

Perhaps the strongest ethical concerns have been raised regarding the potential application of fMRI decoding to detect lies or the presence of guilty knowledge (Bizzi et al 2009). Much attention has focused on recent studies of lie detection and their claims, as well as the efforts made by private companies to develop and market this nascent technology. In a study by Langleben and colleagues, participants were given two cards in an envelope and asked in advance to lie whenever they were asked if they had one card and to tell the truth about the other (Davatzikos et al 2005). Pattern classification applied to whole brain activity revealed that truths and lies could be distinguished in this task with 88% accuracy on individual trials, due to greater activity observed for lies in multiple areas including the prefrontal cortex, anterior cingulate and insula. On the basis of these findings, some rather bold claims were made about the prospects of future fMRI lie detection technology. However, it is critical to note that it is not lying *per se* that is being decoded from these brain areas, but rather the cognitive and emotional processes that are associated with lying (Spence et al 2004). Thus, lie detection technology suffers the same problem of *reverse inference* that we have discussed previously. Although lying typically leads to the activation of a certain set of brain areas, the activation of these brain areas does not necessarily indicate lying. In real world settings, such as when a defendant is strongly suspected of committing a crime or feels guilty for having witnessed the crime, any questions about the crime might elicit strong emotional and cognitive responses akin to those evoked by lying. It is also not clear whether criminals, especially those with psychopathy, would show the same activity patterns during lying. Other fMRI studies have shown that brain activity patterns differ for prepared lies and spontaneous lies (Ganis et al 2003), and that fMRI lie detection technology can be subverted by covertly engaging in an separate cognitive task during brain scanning (Ganis et al 2011). These major shortcomings bring into serious question whether it will be possible to develop an ecologically valid and reliable fMRI lie detector anytime in the near future.

However, this has not prevented the recent efforts of private companies to market such technology or to prepare for their use in courtrooms. In May 2010, the first Daubert hearing was held in Tennessee to determine whether fMRI lie detection might be considered admissible as scientific evidence (Miller 2010). Dr. Steven Laken, CEO of Cephos, a company that provides fMRI lie detection services, presented evidence in favor of admitting the brain scans he had performed on the defendant, which according to him, indicated innocence on the charges of fraud. The prosecution invited expert testimony from neuroscientist Marcus Raichle and statistician Peter Imrey to dispute the reliability of the current technology. In the end, the judge determined that fMRI lie detection technology was supported by peer-reviewed publications, but had not gained wide acceptance among scientists. Moreover, its reliability and accuracy had yet to be validated in real-world settings, and a well-standardized protocol for implementing such tests had yet to be established (Shen & Jones In press).

It remains to be seen whether fMRI lie detection will ever improve enough to meet general scientific acceptance or gain admission into courts. Nevertheless, it would be prudent to consider the potential ethical and societal ramifications of such technology, should it improve to the point that detection accuracy is no longer the primary concern. There would be obvious benefits in a legal setting if accuracy were extremely high. However, mental privacy could face enormous new challenges, in both legal settings and beyond, as there has been no precedent for being able to look into the mind of another human being. Although DNA can be obtained as evidence from a suspect based on court order, brain reading of thoughts might fall under the category of testimony, in which case defendants would be protected by the Fifth Amendment. Even so, if the technology were ever to develop to near-perfect levels of accuracy, a refusal to voluntarily submit to fMRI lie detection might be interpreted as an implicit admission of guilt by some juries even when instructed not to do so. In the worlds of business and personal relationships, the availability of such technology could have far-reaching consequences, especially in situations involving employers and employees, business partners, or even spouses. Just the existence of such technology and the pressure of being asked to undergo testing could lead people to disclose information that they otherwise would have declined to share.

Given the conceptual challenges of developing reliable fMRI lie detection and the fact that people can use countermeasures to alter their patterns of brain activity, we are doubtful that the technology will progress to being truly reliable and ecologically valid. Nonetheless, it is important to consider potential implications in case it ever does.

CONCLUDING REMARKS

In recent years, fMRI pattern classification has led to rapid advances in many areas of cognitive neuroscience, encompassing perception, attention, object processing, memory, semantics, language processing and decision making. These methods have allowed neuroimaging researchers to isolate feature-selective sensory responses, neural correlates of conscious perception, content-specific activity during attention and memory tasks, and brain activity patterns that are predictive of future decisions.

Furthermore, multivariate analyses can be used to characterize the multidimensional nature of neural representations, such as the functional similarity between object representations, scene representations or semantic representations, allowing one to predict how the brain should respond to novel stimuli. Looking forward, the enhanced sensitivity and information content provided by these methods should greatly facilitate the investigation of mind-brain relationships, by revealing both local and distributed representations of mental content, functional interactions between brain areas, and the underlying relationships between brain activity and cognitive performance.

Despite, or perhaps because of, the statistical power of these analytic tools, careful experimentation and interpretation is required when making inferences about successful decoding of a stimulus, task, or mental state from human brain activity. The extension of these methods into real-world applications could prove very useful for medical diagnosis and also neuroprosthesis (Hatsopoulos & Donoghue 2009). However, there are major

concerns regarding the reliability and ecological validity of current attempts to perform real-world lie detection. Much more research will be needed to determine whether such methods might be valid or not. Strong ethical considerations also revolve around the prospect of developing reliable lie detection technology, and it would be prudent to consider how mental privacy would be protected if such technology were allowed to gain prominent use.

Acknowledgments.

The authors would like to thank Owen Jones, Yukiyasu Kamitani, Sean Polyn, Elizabeth Counterman, and Jascha Swisher for helpful comments on earlier versions of this manuscript. The authors were supported by grants from the National Eye Institute (R01EY017082), the National Science Foundation (BCS-0642633), and the Defense Advanced Research Projects Agency.

References

Abrams DA, Bhatara A, Ryali S, Balaban E, Levitin DJ, Menon V. 2011 Decoding temporal structure in music and speech relies on shared brain resources but elicits different fine-scale spatial patterns. Cereb Cortex 21:1507–18 [PubMed: 21071617]

Adolphs R 2010 What does the amygdala contribute to social cognition? Ann N Y Acad Sci 1191:42–61 [PubMed: 20392275]

Bizzi E, Hyman SE, Raichle ME, Kanwisher N, Phelps EA, et al. 2009 Using imaging to identify deceit: Scientific and ethical questions. Cambridge, MA: American Academy of Arts and Sciences

Blake R, Logothetis NK. 2002 Visual competition. Nat Rev Neurosci 3:13–21. [PubMed: 11823801]

Boynton GM, Finney EM. 2003 Orientation-specific adaptation in human visual cortex. J Neurosci 23:8781–7 [PubMed: 14507978]

Brouwer GJ, Heeger DJ. 2009 Decoding and reconstructing color from responses in human visual cortex. J Neurosci 29:13992–4003 [PubMed: 19890009]

Brouwer GJ, van Ee R. 2007 Visual cortex allows prediction of perceptual states during ambiguous structure-from-motion. J Neurosci 27:1015–23 [PubMed: 17267555]

Carlson TA, Schrater P, He S. 2003 Patterns of activity in the categorical representations of objects. J. Cogn. Neurosci 15:704–17 [PubMed: 12965044]

Chadwick MJ, Hassabis D, Weiskopf N, Maguire EA. 2010 Decoding individual episodic memory traces in the human hippocampus. Curr Biol 20:544–7 [PubMed: 20226665]

Chan AW, Kravitz DJ, Truong S, Arizpe J, Baker CI. 2010 Cortical representations of bodies and faces are strongest in commonly experienced configurations. Nat Neurosci 13:417–8 [PubMed: 20208528]

Clifford CW, Mannion DJ, McDonald JS. 2009 Radial biases in the processing of motion and motion-defined contours by human visual cortex. J Neurophysiol 102:2974–81 [PubMed: 19759326]

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 19:261–70 [PubMed: 12814577]

Craddock RC, Holtzheimer PE 3rd, Hu XP, Mayberg HS. 2009 Disease state prediction from resting state functional connectivity. Magn Reson Med 62:1619–28 [PubMed: 19859933]

Davatzikos C, Ruparel K, Fan Y, Shen DG, Acharyya M, et al. 2005 Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. Neuroimage 28:663–8 [PubMed: 16169252]

Eger E, Ashburner J, Haynes JD, Dolan RJ, Rees G. 2008 fMRI activity patterns in human LOC carry information about object exemplars within category. J Cogn Neurosci 20:356–70 [PubMed: 18275340]

Eger E, Michel V, Thirion B, Amadon A, Dehaene S, Kleinschmidt A. 2009 Deciphering cortical number coding from human brain activity patterns. Curr Biol 19:1608–15 [PubMed: 19781939]

Engel SA, Furmanski CS. 2001 Selective adaptation to color contrast in human primary visual cortex. J Neurosci 21:3949–54 [PubMed: 11356883]

Farah MJ. 2005 Neuroethics: the practical and the philosophical. Trends Cogn Sci 9:34–40 [PubMed: 15639439]

Fischer J, Spotswood N, Whitney D. 2011 The emergence of perceived position in the visual system. J Cogn Neurosci 23:119–36 [PubMed: 20044886]

Formisano E, De Martino F, Bonte M, Goebel R. 2008 "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322:970–3 [PubMed: 18988858]

Freeman J, Brouwer GJ, Heeger DJ, Merriam EP. 2011 Orientation decoding depends on maps, not columns. J Neurosci 31:4792–804 [PubMed: 21451017]

Freiwald WA, Tsao DY. 2010 Functional compartmentalization and viewpoint generalization within the macaque face-processing system. Science 330:845–51 [PubMed: 21051642]

Freiwald WA, Tsao DY, Livingstone MS. 2009 A face feature space in the macaque temporal lobe. Nat Neurosci 12:1187–96 [PubMed: 19668199]

Ganis G, Kosslyn SM, Stose S, Thompson WL, Yurgelun-Todd DA. 2003 Neural correlates of different types of deception: an fMRI investigation. Cereb Cortex 13:830–6 [PubMed: 12853369]

Ganis G, Rosenfeld JP, Meixner J, Kievit RA, Schendan HE. 2011 Lying in the scanner: covert countermeasures disrupt deception detection by functional magnetic resonance imaging. Neuroimage 55:312–9 [PubMed: 21111834]

Gerardin P, Kourtzi Z, Mamassian P. 2010 Prior knowledge of illumination for 3D perception in the human brain. Proc Natl Acad Sci U S A 107:16309–14 [PubMed: 20805488]

Goddard E, Mannion DJ, McDonald JS, Solomon SG, Clifford CW. 2010 Combination of subcortical color channels in human visual cortex. J Vis 10:25

Greenberg AS, Esterman M, Wilson D, Serences JT, Yantis S. 2010 Control of spatial and feature-based attention in frontoparietal cortex. J Neurosci 30:14330–9 [PubMed: 20980588]

Greene J, Cohen J. 2004 For the law, neuroscience changes nothing and everything. Philos Trans R Soc Lond B Biol Sci 359:1775–85 [PubMed: 15590618]

Hampton AN, O'Doherty JP. 2007 Decoding the neural substrates of reward-related decision making with functional MRI. Proc Natl Acad Sci U S A 104:1377–82 [PubMed: 17227855]

Harenski CL, Harenski KA, Shane MS, Kiehl KA. 2010 Aberrant neural processing of moral violations in criminal psychopaths. J Abnorm Psychol 119:863–74 [PubMed: 21090881]

Harrison SA, Tong F. 2009 Decoding reveals the contents of visual working memory in early visual areas. Nature 458:632–5 [PubMed: 19225460]

Hassabis D, Chu C, Rees G, Weiskopf N, Molyneux PD, Maguire EA. 2009 Decoding neuronal ensembles in the human hippocampus. Curr Biol 19:546–54 [PubMed: 19285400]

Hatsopoulos NG, Donoghue JP. 2009 The science of neural interface systems. Annu Rev Neurosci 32:249–66 [PubMed: 19400719]

Haushofer J, Livingstone MS, Kanwisher N. 2008 Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity. PLoS Biol 6:e187 [PubMed: 18666833]

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001 Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425–30. [PubMed: 11577229]

Haynes JD, Deichmann R, Rees G. 2005 Eye-specific effects of binocular rivalry in the human lateral geniculate nucleus. Nature 438:496–9 [PubMed: 16244649]

Haynes JD, Rees G. 2005a Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat Neurosci 8:686–91 [PubMed: 15852013]

Haynes JD, Rees G. 2005b Predicting the stream of consciousness from activity in human visual cortex. Curr Biol 15:1301–7 [PubMed: 16051174]

Haynes JD, Rees G. 2006 Decoding mental states from brain activity in humans. Nat Rev Neurosci 7:523–34 [PubMed: 16791142]

Haynes JD, Sakai K, Rees G, Gilbert S, Frith C, Passingham RE. 2007 Reading hidden intentions in the human brain. Curr Biol 17:323–8 [PubMed: 17291759]

Hoeft F, McCandliss BD, Black JM, Gantman A, Zakerani N, et al. 2011 Neural systems predicting long-term outcome in dyslexia. Proc Natl Acad Sci U S A 108:361–6 [PubMed: 21173250]

Hong SW, Tong F, Seiffert AE. 2012 Direction-selective patterns of activity in human visual cortex reveal common neural substrates for different types of motion. Neuropsychologia 50:514–21 [PubMed: 21945806]

Howard JD, Plailly J, Grueschow M, Haynes JD, Gottfried JA. 2009 Odor quality coding and categorization in human posterior piriform cortex. Nat Neurosci 12:932–8 [PubMed: 19483688]

Howard MW, Kahana MJ. 1999 Contextual variability and serial position effects in free recall. J Exp Psychol Learn Mem Cogn 25:923–41 [PubMed: 10439501]

Hubbard EM, Piazza M, Pinel P, Dehaene S. 2005 Interactions between number and space in parietal cortex. Nat Rev Neurosci 6:435–48 [PubMed: 15928716]

Hughes V 2010 Science in court: head case. Nature 464:340–2 [PubMed: 20237536]

Jehee JF, Brady DK, Tong F. 2011 Attention improves encoding of task-relevant features in the human visual cortex. J Neurosci 31:8210–9 [PubMed: 21632942]

Johnson JD, McDuff SG, Rugg MD, Norman KA. 2009 Recollection, familiarity, and cortical reinstatement: a multivoxel pattern analysis. Neuron 63:697–708 [PubMed: 19755111]

Kamitani Y, Tong F. 2005 Decoding the visual and subjective contents of the human brain. Nat Neurosci 8:679–85 [PubMed: 15852014]

Kamitani Y, Tong F. 2006 Decoding seen and attended motion directions from activity in the human visual cortex. Curr Biol 16:1096–102 [PubMed: 16753563]

Kaul C, Rees G, Ishai A. 2011 The Gender of Face Stimuli is Represented in Multiple Regions in the Human Brain. Front Hum Neurosci 4:238 [PubMed: 21270947]

Kay KN, Naselaris T, Prenger RJ, Gallant JL. 2008 Identifying natural images from human brain activity. Nature 452:352–5 [PubMed: 18322462]

Kiani R, Esteky H, Mirpour K, Tanaka K. 2007 Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. J Neurophysiol 97:4296–309 [PubMed: 17428910]

Knops A, Thirion B, Hubbard EM, Michel V, Dehaene S. 2009 Recruitment of an area involved in eye movements during mental arithmetic. Science 324:1583–5 [PubMed: 19423779]

Knutson B, Rick S, Wimmer GE, Prelec D, Loewenstein G. 2007 Neural predictors of purchases. Neuron 53:147–56 [PubMed: 17196537]

Kosslyn SM, Ganis G, Thompson WL. 2001 Neural foundations of imagery. Nat Rev Neurosci 2:635–42 [PubMed: 11533731]

Kosslyn SM, Thompson WL. 2003 When is early visual cortex activated during visual mental imagery? Psychol Bull 129:723–46 [PubMed: 12956541]

Kriegeskorte N 2011 Pattern-information analysis: From stimulus decoding to computational-model testing. Neuroimage 56:411–21 [PubMed: 21281719]

Kriegeskorte N, Formisano E, Sorger B, Goebel R. 2007 Individual faces elicit distinct response patterns in human anterior temporal cortex. Proc Natl Acad Sci U S A 104:20600–5 [PubMed: 18077383]

Kriegeskorte N, Goebel R, Bandettini P. 2006 Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863–8 [PubMed: 16537458]

Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, et al. 2008 Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60:1126–41 [PubMed: 19109916]

Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI. 2009 Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 12:535–40 [PubMed: 19396166]

Lewis-Peacock JA, Postle BR. 2008 Temporary activation of long-term memory supports working memory. Journal of Neuroscience 28:8765–71 [PubMed: 18753378]

Liu T, Hospadaruk L, Zhu DC, Gardner JL. 2011 Feature-specific attentional priority signals in human cortex. J Neurosci 31:4484–95 [PubMed: 21430149]

Lu HD, Chen G, Tanigawa H, Roe AW. 2010 A motion direction map in macaque V2. Neuron 68:1002–13 [PubMed: 21145011]

Mahon BZ, Anzellotti S, Schwarzbach J, Zampini M, Caramazza A. 2009 Category-specific organization in the human brain does not require visual experience. Neuron 63:397–405 [PubMed: 19679078]

Mahon BZ, Caramazza A. 2009 Concepts and categories: a cognitive neuropsychological perspective. Annu Rev Psychol 60:27–51 [PubMed: 18767921]

Mahon BZ, Caramazza A. 2011 What drives the organization of object knowledge in the brain? Trends Cogn Sci 15:97–103 [PubMed: 21317022]

Mannion DJ, McDonald JS, Clifford CW. 2009 Discrimination of the local orientation structure of spiral Glass patterns early in human visual cortex. Neuroimage 46:511–5 [PubMed: 19385017]

McClelland JL, Rogers TT. 2003 The parallel distributed processing approach to semantic cognition. Nat Rev Neurosci 4:310–22 [PubMed: 12671647]

McDuff SG, 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 29:508–16 [PubMed: 19144851]

Meyer K, Kaplan JT, Essex R, Webber C, Damasio H, Damasio A. 2010 Predicting visual stimuli on the basis of activity in auditory cortices. Nat Neurosci 13:667–8 [PubMed: 20436482]

Miller G 2010 Science and the law. fMRI lie detection fails a legal test. Science 328:1336–7

Misaki M, Kim Y, Bandettini PA, Kriegeskorte N. 2010 Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage 53:103–18 [PubMed: 20580933]

Mitchell TM, Hutchinson R, Just MA, Niculescu RS, Pereira F, Wang X. 2003 Classifying instantaneous cognitive states from FMRI data. AMIA Annu Symp Proc:465–9 [PubMed: 14728216]

Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, et al. 2008 Predicting human brain activity associated with the meanings of nouns. Science 320:1191–5 [PubMed: 18511683]

Miyawaki Y, Uchida H, Yamashita O, Sato MA, Morito Y, et al. 2008 Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron 60:915–29 [PubMed: 19081384]

Moeller S, Freiwald WA, Tsao DY. 2008 Patches with links: a unified system for processing faces in the macaque temporal lobe. Science 320:1355–9 [PubMed: 18535247]

Monti MM, Vanhaudenhuyse A, Coleman MR, Boly M, Pickard JD, et al. 2010 Willful modulation of brain activity in disorders of consciousness. N Engl J Med 362:579–89 [PubMed: 20130250]

Morgan LK, Macevoy SP, Aguirre GK, Epstein RA. 2011 Distances between real-world locations are represented in the human hippocampus. J Neurosci 31:1238–45 [PubMed: 21273408]

Naselaris T, Kay KN, Nishimoto S, Gallant JL. 2011 Encoding and decoding in fMRI. Neuroimage 56:400–10 [PubMed: 20691790]

Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL. 2009 Bayesian reconstruction of natural images from human brain activity. Neuron 63:902–15 [PubMed: 19778517]

Natu VS, Jiang F, Narvekar A, Keshvari S, Blanz V, O'Toole AJ. 2010 Dissociable neural patterns of facial identity across changes in viewpoint. J Cogn Neurosci 22:1570–82 [PubMed: 19642884]

Nenadic I, Sauer H, Gaser C. 2009 Distinct pattern of brain structural deficits in subsyndromes of schizophrenia delineated by psychopathology. Neuroimage 49:1153–60 [PubMed: 19833216]

Norman KA, Polyn SM, Detre GJ, Haxby JV. 2006 Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci 10:424–30 [PubMed: 16899397]

O'Craven KM, Kanwisher N. 2000 Mental imagery of faces and places activates corresponding stiimulus- specific brain regions. J Cogn Neurosci 12:1013–23. [PubMed: 11177421]

O'Toole AJ, Jiang F, Abdi H, Penard N, Dunlop JP, Parent MA. 2007 Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. J Cogn Neurosci 19:1735–52 [PubMed: 17958478]

Obermayer K, Blasdel GG. 1993 Geometry of orientation and ocular dominance columns in monkey striate cortex. J Neurosci 13:4114–29 [PubMed: 8410181]

Offen S, Schluppeck D, Heeger DJ. 2008 The role of early visual cortex in visual short-term memory and visual attention. Vision Res

Op de Beeck HP, Haushofer J, Kanwisher NG. 2008 Interpreting fMRI data: maps, modules and dimensions. Nat Rev Neurosci 9:123–35 [PubMed: 18200027]

Owen AM, Coleman MR, Boly M, Davis MH, Laureys S, Pickard JD. 2006 Detecting awareness in the vegetative state. Science 313:1402 [PubMed: 16959998]

Peelen MV, Fei-Fei L, Kastner S. 2009 Neural mechanisms of rapid natural scene categorization in human visual cortex. Nature 460:94–7 [PubMed: 19506558]

Pereira F, Mitchell T, Botvinick M. 2009 Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45:S199–209 [PubMed: 19070668]

Pessoa L, Padmala S. 2007 Decoding near-threshold perception of fear from distributed single-trial brain activation. Cereb Cortex 17:691–701 [PubMed: 16627856]

Phelps EA. 2006 Emotion and cognition: insights from studies of the human amygdala. Annu Rev Psychol 57:27–53 [PubMed: 16318588]

Pietrini P, Furey ML, Ricciardi E, Gobbini MI, Wu WH, et al. 2004 Beyond sensory images: Object-based representation in the human ventral pathway. Proc Natl Acad Sci U S A 101:5658–63 [PubMed: 15064396]

Poldrack RA, Halchenko YO, Hanson SJ. 2009 Decoding the large-scale structure of brain function by classifying mental States across individuals. Psychol Sci 20:1364–72 [PubMed: 19883493]

Polyn SM, Natu VS, Cohen JD, Norman KA. 2005 Category-specific cortical activity precedes retrieval during memory search. Science 310:1963–6 [PubMed: 16373577]

Preston TJ, Li S, Kourtzi Z, Welchman AE. 2008 Multivoxel pattern selectivity for perceptually relevant binocular disparities in the human brain. J Neurosci 28:11315–27 [PubMed: 18971473]

Raizada RD, Tsao FM, Liu HM, Kuhl PK. 2010 Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: prediction of individual differences. Cereb Cortex 20:1–12 [PubMed: 19386636]

Reddy L, Kanwisher NG, VanRullen R. 2009 Attention and biased competition in multi-voxel object representations. Proc Natl Acad Sci U S A 106:21447–52 [PubMed: 19955434]

Reddy L, Tsuchiya N, Serre T. 2010 Reading the mind's eye: decoding category information during mental imagery. Neuroimage 50:818–25 [PubMed: 20004247]

Rissman J, Greely HT, Wagner AD. 2010 Detecting individual memories through the neural decoding of memory states and past experience. Proc Natl Acad Sci U S A 107:9849–54 [PubMed: 20457911]

Rissman J, Wagner AD. In press Distributed representations in memory: Insights from functional brain imaging. Annual Review in Psychology

Rodriguez PF. 2010 Neural decoding of goal locations in spatial navigation in humans with fMRI. Hum Brain Mapp 31:391–7 [PubMed: 19722170]

Roskies A 2002 Neuroethics for the new millenium. Neuron 35:21–3 [PubMed: 12123605]

Roskies A 2006 Neuroscientific challenges to free will and responsibility. Trends Cogn Sci 10:419–23 [PubMed: 16901745]

Sapolsky RM. 2004 The frontal cortex and the criminal justice system. Philos Trans R Soc Lond B Biol Sci 359:1787–96 [PubMed: 15590619]

Saproo S, Serences JT. 2010 Spatial attention improves the quality of population codes in human visual cortex. J Neurophysiol 104:885–95 [PubMed: 20484525]

Sasaki Y, Rajimehr R, Kim BW, Ekstrom LB, Vanduffel W, Tootell RB. 2006 The radial bias: a different slant on visual orientation sensitivity in human and nonhuman primates. Neuron 51:661–70 [PubMed: 16950163]

Sato JR, de Oliveira-Souza R, Thomaz CE, Basilio R, Bramati IE, et al. 2011 Identification of psychopathic individuals using pattern classification of MRI images. Soc Neurosci:1–13

Schurger A, Pereira F, Treisman A, Cohen JD. 2010 Reproducibility distinguishes conscious from nonconscious neural representations. Science 327:97–9 [PubMed: 19965385]

Schwarzlose RF, Swisher JD, Dang S, Kanwisher N. 2008 The distribution of category and location information across object-selective regions in human visual cortex. Proc Natl Acad Sci U S A 105:4447–52 [PubMed: 18326624]

Scolari M, Serences JT. 2010 Basing perceptual decisions on the most informative sensory neurons. J Neurophysiol 104:2266–73 [PubMed: 20631209]

Serences JT, Boynton GM. 2007a Feature-based attentional modulations in the absence of direct visual stimulation. Neuron 55:301–12 [PubMed: 17640530]

Serences JT, Boynton GM. 2007b The representation of behavioral choice for motion in human visual cortex. J Neurosci 27:12893–9 [PubMed: 18032662]

Serences JT, Ester EF, Vogel EK, Awh E. 2009 Stimulus-specific delay activity in human primary visual cortex. Psychol Sci 20:207–14 [PubMed: 19170936]

Seymour K, Clifford CW, Logothetis NK, Bartels A. 2009 The coding of color, motion, and their conjunction in the human visual cortex. Current Biology 19:177–83 [PubMed: 19185496]

Seymour K, Clifford CW, Logothetis NK, Bartels A. 2010 Coding and binding of color and form in visual cortex. Cereb Cortex 20:1946–54 [PubMed: 20019147]

Shen FX, Jones OD, (2 23, 2011)., Vol., 2011. In press Brain Scans as Evidence: Truths, Proofs, Lies, and Lessons Mercer Law Review 62

Shmuel A, Chaimow D, Raddatz G, Ugurbil K, Yacoub E. 2010 Mechanisms underlying decoding at 7 T: ocular dominance columns, broad structures, and macroscopic blood vessels in V1 convey information on the stimulated eye. Neuroimage 49:1957–64 [PubMed: 19715765]

Smith AT, Kosillo P, Williams AL. 2011 The confounding effect of response amplitude on MVPA performance measures. Neuroimage 56:525–30 [PubMed: 20566321]

Soon CS, Brass M, Heinze HJ, Haynes JD. 2008 Unconscious determinants of free decisions in the human brain. Nat Neurosci 11:543–5 [PubMed: 18408715]

Spence SA, Hunter MD, Farrow TF, Green RD, Leung DH, et al. 2004 A cognitive neurobiological account of deception: evidence from functional neuroimaging. Philos Trans R Soc Lond B Biol Sci 359:1755–62 [PubMed: 15590616]

Spiridon M, Kanwisher N. 2002 How distributed is visual category information in human occipito-temporal cortex? An fMRI study. Neuron 35:1157–65 [PubMed: 12354404]

Sterzer P, Haynes JD, Rees G. 2008 Fine-scale activity patterns in high-level visual areas encode the category of invisible objects. J Vis 8:10 1–2

Stokes M, Thompson R, Cusack R, Duncan J. 2009 Top-down activation of shape-specific population codes in visual cortex during mental imagery. J Neurosci 29:1565–72 [PubMed: 19193903]

Sumner P, Anderson EJ, Sylvester R, Haynes JD, Rees G. 2008 Combined orientation and colour information in human V1 for both L-M and S-cone chromatic axes. Neuroimage 39:814–24 [PubMed: 17964188]

Swisher JD, Gatenby JC, Gore JC, Wolfe BA, Moon CH, et al. 2010 Multiscale pattern analysis of orientation-selective activity in the primary visual cortex. J Neurosci 30:325–30 [PubMed: 20053913]

Thirion B, Duchesnay E, Hubbard E, Dubois J, Poline JB, et al. 2006 Inverse retinotopy: inferring the visual content of images from brain activation patterns. Neuroimage 33:1104–16 [PubMed: 17029988]

Tong F, Meng M, Blake R. 2006 Neural bases of binocular rivalry. Trends Cogn Sci 10:502–11 [PubMed: 16997612]

Treisman A 1996 The binding problem. Curr Opin Neurobiol 6:171–8 [PubMed: 8725958]

Treue S, Maunsell JH. 1996 Attentional modulation of visual motion processing in cortical areas MT and MST. Nature 382:539–41 [PubMed: 8700227]

Tsao DY, Freiwald WA, Tootell RB, Livingstone MS. 2006 A cortical region consisting entirely of face-selective cells. Science 311:670–4 [PubMed: 16456083]

Tulving E, Thomson DM. 1973 Encoding Specificity and Retrieval Processes in Episodic Memory. Psychological Review 80:352–73

Tusche A, Bode S, Haynes JD. 2010 Neural responses to unattended products predict later consumer choices. J Neurosci 30:8024–31 [PubMed: 20534850]

Weisberg DS, Keil FC, Goodstein J, Rawson E, Gray JR. 2008 The seductive allure of neuroscience explanations. J Cogn Neurosci 20:470–7 [PubMed: 18004955]

Williams MA, Baker CI, Op de Beeck HP, Shim WM, Dang S, et al. 2008 Feedback of visual object information to foveal retinotopic cortex. Nat Neurosci 11:1439–45 [PubMed: 18978780]

Williams MA, Dang S, Kanwisher NG. 2007 Only some spatial patterns of fMRI response are read out in task performance. Nat Neurosci 10:685–6 [PubMed: 17486103]

Wunderlich K, Schneider KA, Kastner S. 2005 Neural correlates of binocular rivalry in the human lateral geniculate nucleus. Nat Neurosci 8:1595–602 [PubMed: 16234812]

Xue G, Dong Q, Chen C, Lu Z, Mumford JA, Poldrack RA. 2010 Greater neural pattern similarity across repetitions is associated with better memory. Science 330:97–101 [PubMed: 20829453]

Yacoub E, Harel N, Ugurbil K. 2008 High-field fMRI unveils orientation columns in humans. Proc Natl Acad Sci U S A 105:10607–12 [PubMed: 18641121]

Yamashita O, Sato MA, Yoshioka T, Tong F, Kamitani Y. 2008 Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. Neuroimage 42:1414–29 [PubMed: 18598768]

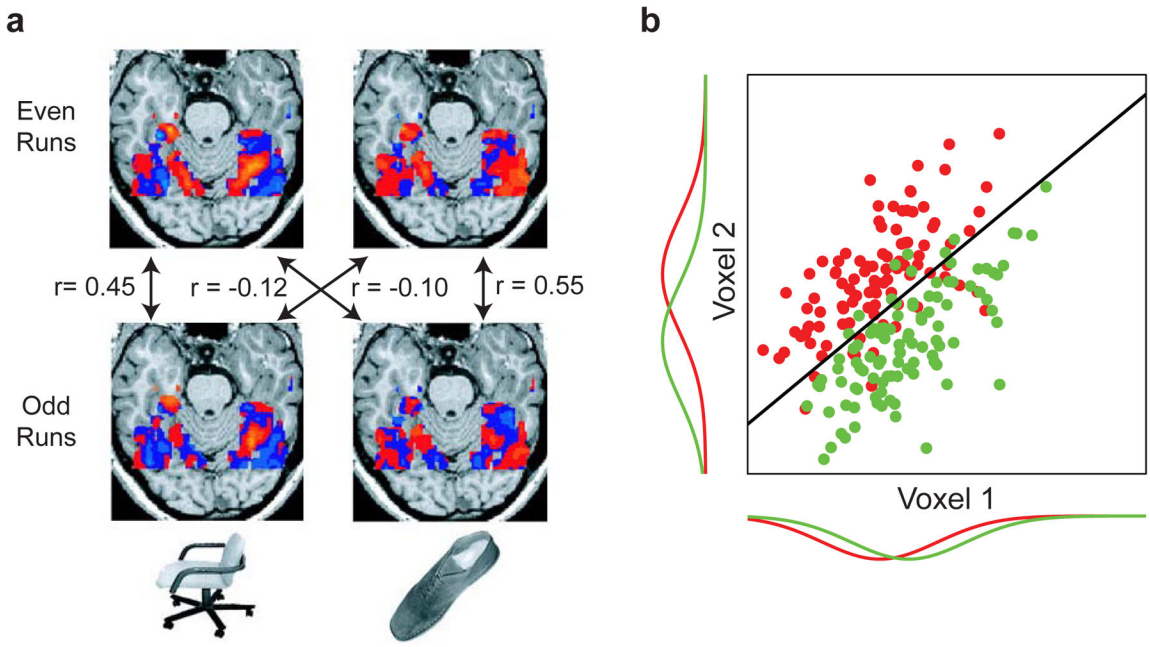

Figure 1.

Correlation and classification approaches to decoding brain activity patterns. **(a)** Average activity patterns for chairs and shoes in the ventral temporal cortex, calculated separately for even and odd runs. Correlations between these spatial patterns of activity were calculated between even and odd runs. Pairwise classifications between any two object categories were considered correct if the correlations were higher within an object category than between the two object categories. Adapted with permission from Haxby et al (2001). **(b)** Hypothetical responses of two voxels to two different experimental conditions, denoted by red and green points. Density plots in the margins indicate the distribution of responses to the two conditions for each voxel considered in isolation. The dividing line between red and green data points shows the classification results from a linear support vector machine applied to these patterns of activity; any points above the line would be classified as red, and those below would be classified as green.

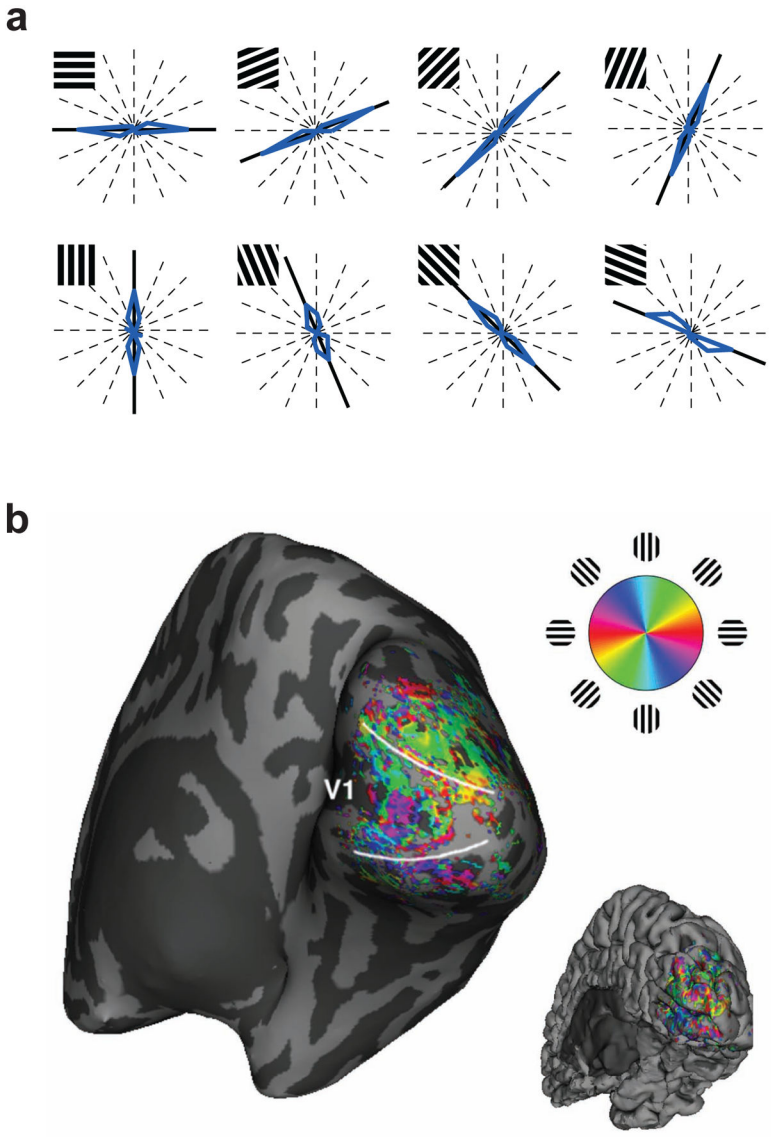

Figure 2.

Decoding the orientation of viewed gratings from activity patterns in the visual cortex. **(a)** Blue curves indicate the distribution of predicted orientations shown on polar plots, with thick black lines indicating the true orientations. Note that common values are plotted at symmetrical directions, because stimulus orientation repeats every 180°. Reproduced with permission from Kamitani & Tong (2005). **(b)** Spatial distribution of weak orientation preferences in the visual cortex, measured using high-resolution functional MRI with 1mm isotropic voxels and plotted on an inflated representation of the cortical surface. Reproduced with permission from Swisher et al (2010).

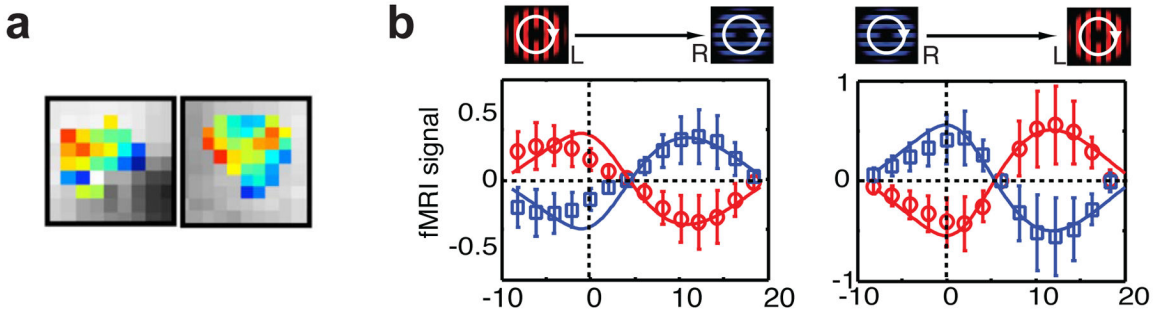

Figure 3.

Eye-specific modulation of activity in the lateral geniculate nucleus (LGN) during binocular rivalry. **(a)** Distribution of weak monocular preferences in the LGN of a representative participant. **(b)** Time course of the decoded eye-specific signal from these LGN activity patterns is correlated with fluctuations in perceptual dominance during rivalry between left-eye and right-eye stimuli. Reproduced with permission from Haynes et al (2005).

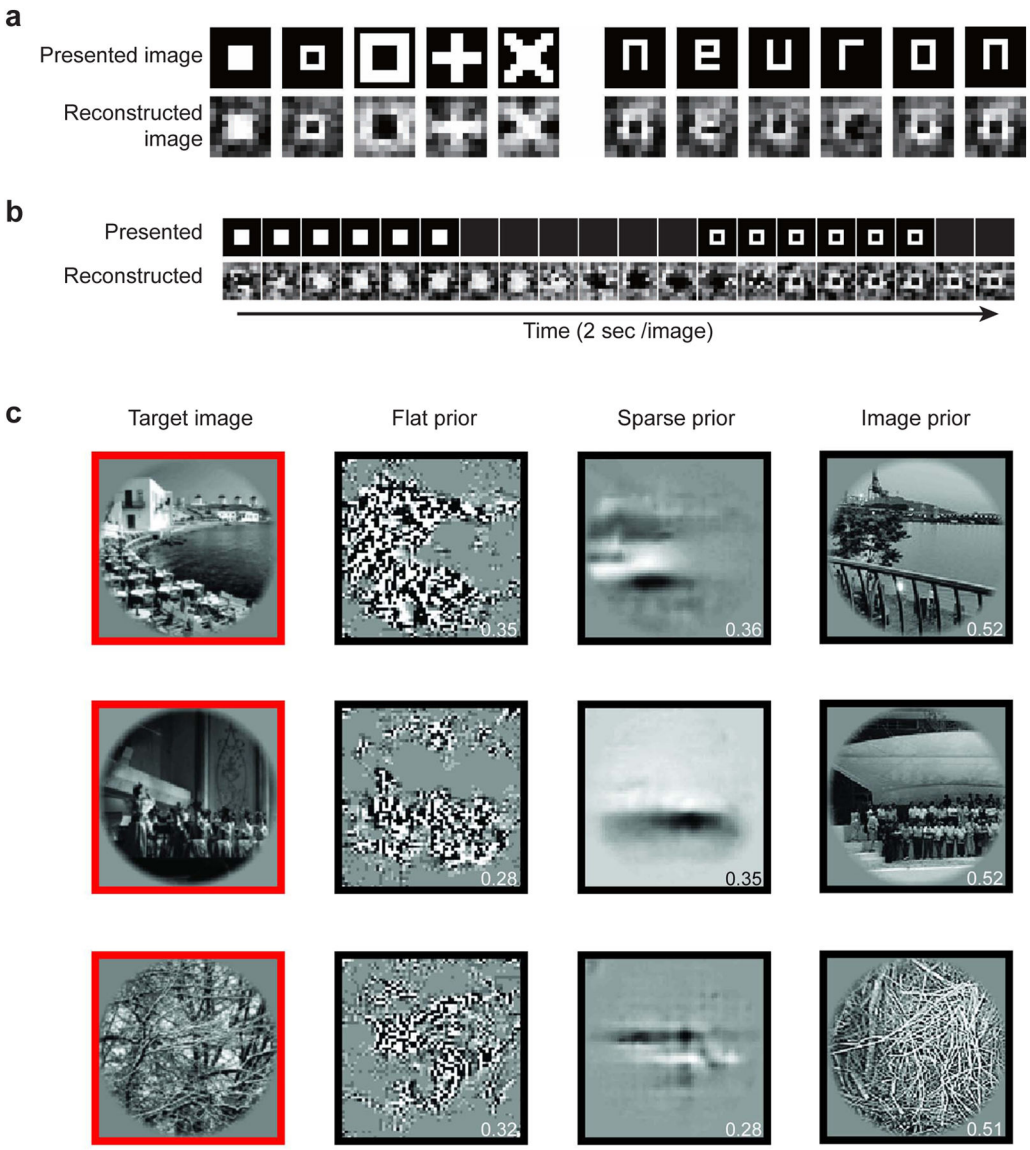

Figure 4.

Reconstruction of viewed images from activity patterns in the visual cortex, based on averaged fMRI activity patterns (**a**) and single fMRI volumes acquired every 2 seconds (**b**). Reproduced with permission from Miyawaki et al (2008). **(c)** Reconstruction of natural scenes from visual cortical activity. Various methods are used to reconstruct the image's high-contrast regions (flat prior) or low spatial frequency components (sparse prior), or to select the most visually and semantically similar image to the target from a database of 6 million predefined images (image prior). Reproduced with permission from Naselaris et al (2009).

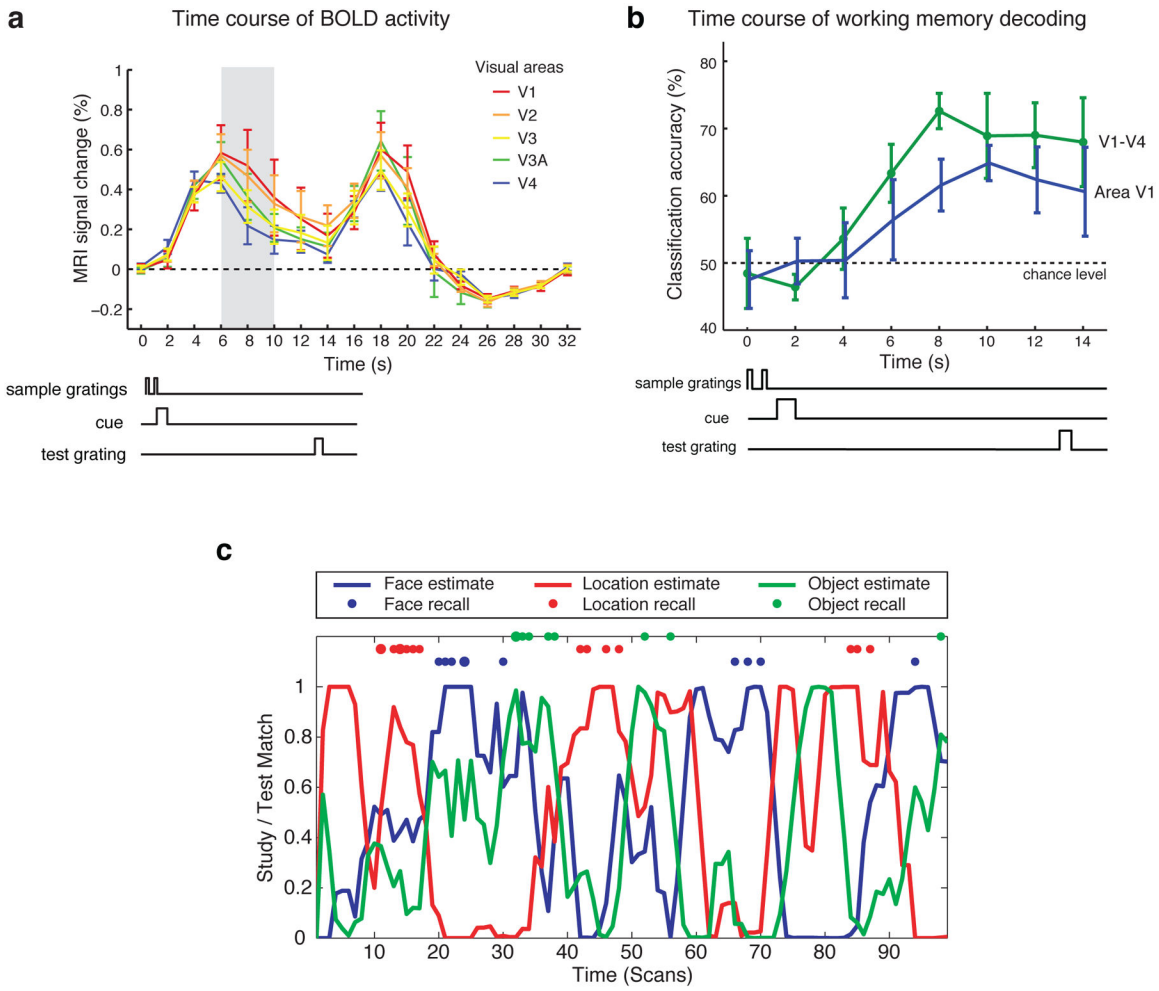

Figure 5.

Decoding item-specific information over time during working memory or free recall from long-term memory. **(a)** Average time course of BOLD activity during a visual working memory task, in which two oriented gratings were briefly shown followed by a postcue indicating which orientation to retain until test. Although the mean BOLD signal steadily declined during the memory retention interval, decoding accuracy for the retained orientation remained elevated throughout the delay period **(b)**. Adapted with permission from Harrison & Tong (2009). **(c)** Classification of the reinstated context during a participant's free recall of famous faces, famous places, and common objects. Dots indicate whenever the participant verbally reported an item from a given category. Curves show estimates of match between fMRI activity patterns at each time point during free recall, using classifiers trained on activity patterns from the prior study period with each of the three categories. Reproduced with permission from Polyn et al (2005).

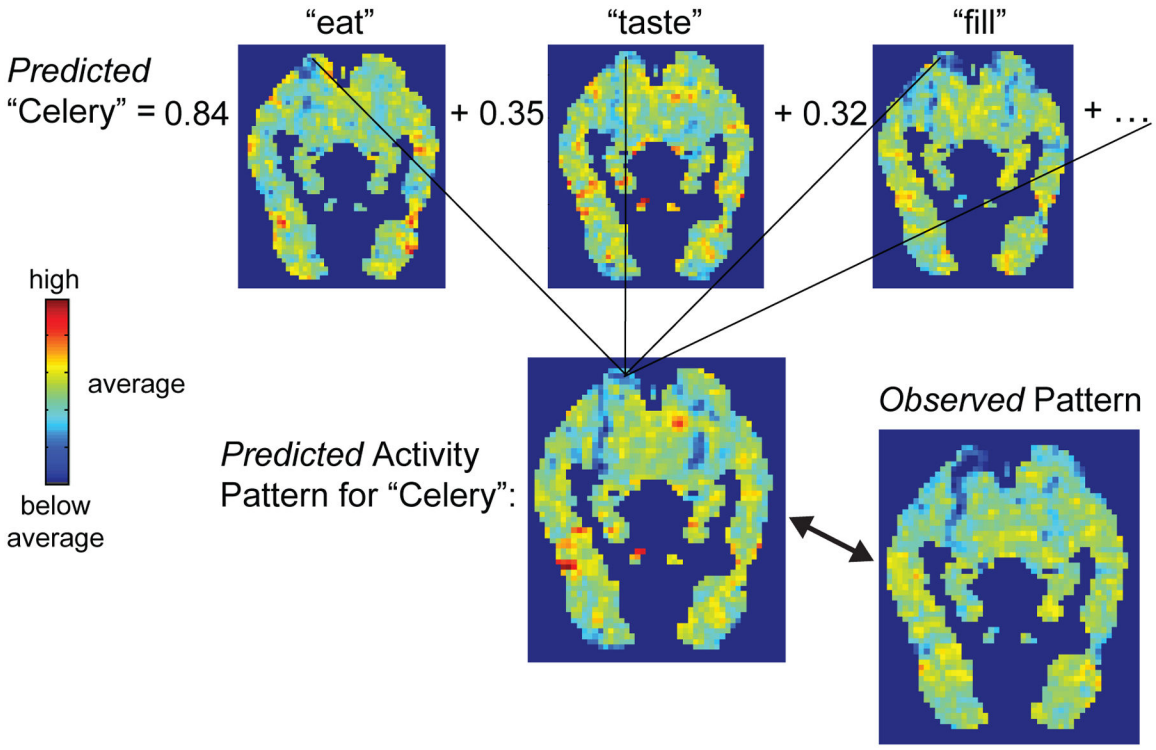

Figure 6.

Semantic encoding model used to predict brain activity patterns to novel nouns. Neural responses to viewed objects and their name, such as "celery", were modeled as the sum of weighted activity patterns to intermediate semantic features consisting of 25 different verbs. Examples of activity patterns for 3 semantic features ("eat", "taste" and "fill") are shown, and the weight of their contribution to the predicted activity pattern reflects their frequency of co-occurrence with the target word. Predicted activity patterns are then compared to the observed activity for celery. Adapted with permission from Mitchell et al (2008).