NEUROIMAGING

Decoding mental states from brain activity in humans

John-Dylan Haynes[*‡§] *and Geraint Rees*[‡§]

Abstract | Recent advances in human neuroimaging have shown that it is possible to accurately decode a person's conscious experience based only on non-invasive measurements of their brain activity. Such 'brain reading' has mostly been studied in the domain of visual perception, where it helps reveal the way in which individual experiences are encoded in the human brain. The same approach can also be extended to other types of mental state, such as covert attitudes and lie detection. Such applications raise important ethical issues concerning the privacy of personal thought.

Multivariate analysis
An analytical technique that considers (or solves) multiple decision variables. In the present context, multivariate analysis takes into account patterns of information that might be present across multiple voxels measured by neuroimaging techniques.

Univariate analysis
Univariate statistical analysis considers only single-decision variables at any one time. Conventional brain imaging data analyses are mass univariate in that they consider how responses vary at very many single voxels, but consider each individual voxel separately.

*Max Planck Institute for Cognitive and Brain Sciences, Stephanstrasse 1a, 04103 Leipzig, Germany. ‡Wellcome Department of Imaging Neuroscience, Institute of Neurology, University College London, 12 Queen Square, London WC1N 3BG, UK. §Institute of Cognitive Neuroscience, University College London, Alexandra House, 17 Queen Square, London WC1N 3AR, UK. Correspondence to J.D.H. e-mail:haynes@cbs.mpg.de
doi:10.1038/nrn1931

Is it possible to tell what someone is currently thinking based only on measurements of their brain activity? At first sight, the answer to this question might seem easy. Many human neuroimaging studies have provided strong evidence for a close link between the mind and the brain, so it should, at least in principle, be possible to decode what an individual is thinking from their brain activity. However, this does not reveal whether such decoding of mental states, or 'brain reading'[1,2], can be practically achieved with current neuroimaging methods. Conventional studies leave the answers to many important questions unclear. For example, how accurately and efficiently can a mental state be inferred? Is the person's compliance required? Is it possible to decode concealed thoughts or even unconscious mental states? What is the maximum temporal resolution? Is it possible to provide a quasi-online estimate of an individual's current cognitive or perceptual state?

Here, we will review new and emerging approaches that directly assess how well a mental state can be reconstructed from non-invasive measurements of brain activity in humans. First we will outline important differences between conventional and decoding-based approaches to human neuroimaging. Then we will review a number of recent studies that have successfully used statistical pattern recognition to decode a person's current thoughts from their brain activity alone. In the final section we will discuss the technical, conceptual and ethical challenges encountered in this emerging field of research.

Multivariate neuroimaging approaches

Conventional neuroimaging approaches typically seek to determine how a particular perceptual or cognitive state is encoded in brain activity, by determining which regions of the brain are involved in a task. This is achieved by measuring activity from many thousands of locations in the brain repeatedly, but then analysing each location separately. This yields a measure of any differences in activity, comparing two or more mental states at each individual sampled location. In theory, if the responses at any brain location differ between two mental states, then it should be possible to use measurements of activity at that brain location to determine which one of those two mental states currently reflects the thinking of the individual. In practice it is often difficult (although not always impossible[3]) to find individual locations where the differences between conditions are sufficiently large to allow for efficient decoding.

In contrast to the strictly location-based conventional analyses, recent work shows that the sensitivity of human neuroimaging can be dramatically increased by taking into account the full spatial pattern of brain activity, measured simultaneously at many locations[2,4–17]. Such pattern-based or multivariate analyses have several advantages over conventional univariate approaches that analyse only one location at a time. First, the weak information available at each location[9] can be accumulated in an efficient way across many spatial locations. Second, even if two single brain regions do not individually carry information about a cognitive state, they might nonetheless do so when jointly analysed[17]. Third, most conventional studies employ processing steps (such as spatial smoothing) that remove fine-grained spatial information that might carry information about perceptual or cognitive states of an individual. This information is discarded in conventional analyses, but can be revealed using methods that simultaneously analyse the pattern of brain activity across multiple locations. Fourth, conventional studies usually probe whether the average activity across all task trials during one condition is significantly different from

Box 1 | **Brain–computer interfaces and invasive recordings**

Important foundations for 'brain reading' in humans have come from research into brain–computer interfaces[79–84] and invasive recordings from human patients[85–87]. For example, humans can be trained to use their brain activity to control artificial devices[79,82–84]. Typically, such brain–computer interfaces are not designed to directly decode cognitive or perceptual states. Instead, individuals are extensively trained to intentionally control certain aspects of recorded brain activity. For example, participants learn by trial and error to voluntarily control scalp-recorded electroencephalograms (EEGs), such as specific EEG frequency bands[79,84] or slow wave deflections of the cortical potential[88]. This can be effective even for single trials of rapidly paced movements[82]. Such voluntarily controlled brain signals can subsequently be used to control artificial devices to allow subjects to spell words[79,88] or move cursors on computer displays in two dimensions[84]. Interestingly, subjects can even learn to regulate signals recorded using functional MRI in real-time[89]. It might be possible to achieve even better decoding when electrodes are directly implanted into the brain, which is possible in monkeys (for a review, see REF. 81) and occasionally also in human patients[87]. Not only motor commands but also perception can, in principle, be decoded from the spiking activity of single neurons in humans[85,86] and animals[39,90]. However, such invasive techniques necessarily involve surgical implantation of electrodes that is not feasible at present for use in healthy human participants.

the average activity across all time points during a second condition. Typically these studies acquire a large number of samples of brain activity to maximize statistical sensitivity. However, by computing the average activity, information about the functional state of the brain at any given time point is lost. By contrast, the increased sensitivity of decoding-based approaches potentially allows even quasi-online estimates of a person's perceptual or cognitive state[10,16].

Taken together, pattern-based techniques allow considerable increases in the amount of information that can be decoded about the current mental state from measurements of brain activity. Therefore, the shift of focus away from conventional location-based analysis strategies towards the decoding of mental states can shed light on the most suitable methods by which information can be extracted from brain activity. Several related fields have laid important foundations for brain reading in studies of animals and patients (BOX 1). Here, we will instead focus on approaches to infer conscious and unconscious perceptual and cognitive mental states from non-invasive measurements of brain activity in humans. These techniques measure neural responses indirectly, through scalp electrical potentials or blood-oxygenation-level-dependent (BOLD) functional MRI (fMRI) signals. It is therefore important to bear in mind that the relationship between the signal used for brain reading and the underlying neural activity can be complex or indirect[18]; although, for most practical purposes the key question is how well brain reading can predict cognitive or perceptual states. Importantly, we will also discuss emerging technical and methodological challenges as well as the ethical implications of such research.

Blood-oxygen-level-dependent (BOLD) signal
Functional MRI measures local changes in the proportion of oxygenated blood in the brain; the BOLD signal. This proportion changes in response to neural activity. Therefore, the BOLD signal, or haemodynamic response, indicates the location and magnitude of neural activity.

Decoding the contents of consciousness

A key factor determining whether the conscious and unconscious perceptual or cognitive states of an individual can be decoded is how well the brain activity corresponding to one particular state can be distinguished or separated from alternate possibilities. An ideal case of such separation is given when different cognitive states are encoded in spatially distinct locations of the brain. As long as these locations can be spatially resolved with the limited resolution of fMRI, this should permit independent measurement of activity associated with these different cognitive or perceptual states. Such an approach has been examined most commonly in the context of the human visual system and in the perception of objects and visual images.

Separable cortical modules. Some regions of the human brain represent particular types of visually presented information in an anatomically segregated way. For example, the fusiform face area (FFA) is a region in the human ventral visual stream that responds more strongly to faces than to any other object category[19,20]. Similarly, the parahippocampal place area (PPA) is an area in the parahippocampal gyrus that responds most to images containing views of houses and visual scenes[21]. As these cortical representations are separated by several centimetres, it is possible to track whether a person is currently thinking of faces or visual scenes by measuring levels of activity in these two brain areas[3]. Activity in FFA measured using fMRI is higher on trials when participants imagine faces (versus buildings), and higher in PPA when they imagine buildings (versus faces). Human observers, given only the activity levels in FFA and PPA from each participant, were able to correctly identify the category of stimulus imagined by the participant in 85% of the individual trials (FIG. 1a). So, individual introspective mental events can be tracked from brain activity at individual locations when the underlying neural representations are well separated. Similar spatially distinct neural representations can also exist as part of macroscopic topographic maps in the brain, such as visual field maps in the early visual cortex[22,23] or the motor or somatosensory homunculi. Activity in separate regions of these macroscopic spatial maps can be used to infer behaviour. For example, when participants move either their right or left thumb, then the difference in activity between left and right primary motor cortex measured with fMRI can be used to accurately predict the movement that they made[24].

Distributed representation patterns. Anatomically distinct 'modular' processing regions in the ventral visual pathway that might permit such powerful decoding from single brain locations have been proposed for many different object categories, such as faces[19,20,25], scenes[21], body parts[26] and, perhaps, letters[27]. However, the number of specialized modules is necessarily limited[28], and the degree to which these areas are indeed specialized for the processing of one class of object alone has been questioned[4,5,29]. Overlapping but distributed representation patterns pose a potential problem for decoding perception from activity in these regions. If a single area responds in many different cognitive states or percepts, then at the relatively low spatial resolution of non-invasive neuroimaging it might be difficult or impossible to use activity in that area to individuate a particular percept or thought. However, the existence

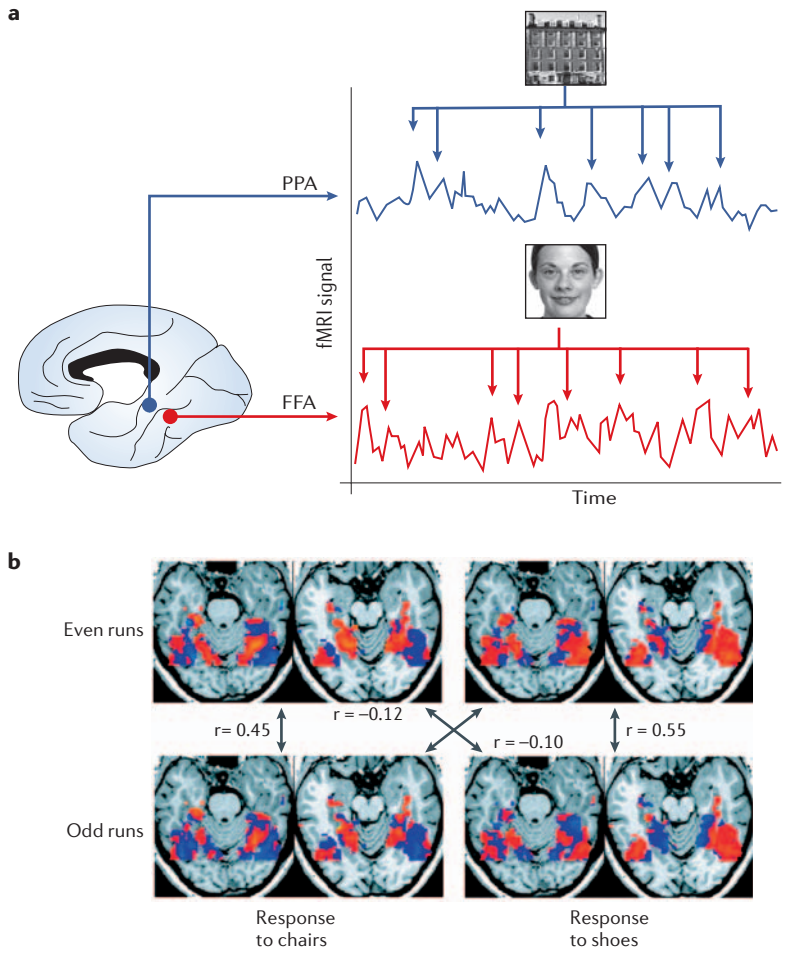


Figure 1 | **Decoding visual object perception from fMRI responses. a** | Decoding the contents of visual imagery from spatially distinct signals in the fusiform face area (FFA, red) and parahippocampal place area (PPA, blue). During periods of face imagery (red arrows), signals are elevated in the FFA whereas during the imagery of buildings (blue arrows), signals are elevated in PPA. A human observer, who was only given signals from the FFA and PPA of each participant, was able to estimate with 85% accuracy which of the two categories the participants were imagining. **b** | Decoding of object perception from response patterns in object-selective regions of the temporal lobe. Viewing of either chairs or shoes evokes a spatially extended response pattern that is slightly different but also partially overlapping for the two categories. To assess how well the perceived object can be decoded from these response patterns, the data are divided into two independent data sets (here odd and even runs). One data set is then used to extract a template response to both chairs and shoes. The response patterns in the remaining data set can then be classified by assigning them to the category with the most 'similar' response template. Here, the similarity is measured using a Pearson correlation coefficient (r), which is higher when comparing same versus different object categories in the two independent data sets. High correlations indicate high similarity between corresponding spatial patterns. Panel **a** modified, with permission, from REF. 3 © (2000) MIT Press. Panel **b** reproduced, with permission, from REF. 4 © (2001) American Association for the Advancement of Science.

of fully independent processing modules for each percept is not necessary for accurate decoding. Instead of separately measuring activity in modular processing regions that are then used to track perception of the corresponding object categories, perceptual decoding can be achieved using an alternative approach that analyses spatially distributed patterns of brain activity[2,4,5,15,30,31]. For example, visual presentation of objects from different categories evokes spatially extended response patterns in the ventral occipitotemporal cortex that partially overlap[4]. Each pattern consists of the spatial distribution of fMRI signals from many locations across the object-selective cortex (FIG. 1b). Perceptual decoding can be achieved by first measuring the representative template (or training) patterns for each of a number of different object categories, and then classifying subsequently acquired test measurements to the category that evoked the most similar response pattern during the training phase. The simplest way to define the similarity between such distributed response patterns is to compute the correlation between a test pattern and each previously acquired template pattern[4]. The test pattern is assigned to the template pattern with which it correlates best. Alternatively, more sophisticated methods for pattern recognition, such as linear discriminant analyses[5] or support vector machines[2] can be used (BOX 2). Using these powerful classification methods, it has proven possible to correctly identify which object a subject is currently viewing, even when several alternative categories are presented[2,4] (for example, diverse objects such as faces, specific animals and man-made objects).

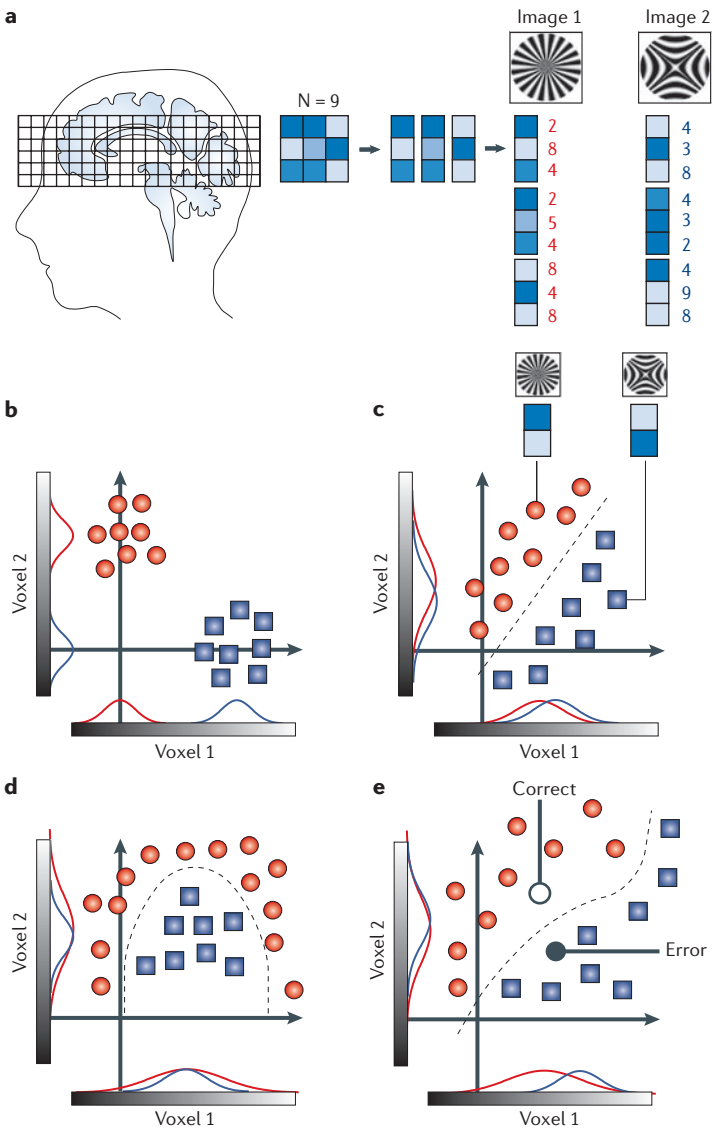
Fine-grained patterns of representation. Many detailed object features are represented at a much finer spatial scale in the cortex than the resolution of fMRI. For example, neurons coding for high-level visual features in the inferior temporal cortex are organized in columnar representations that have a finer spatial scale than the resolution of conventional human neuroimaging techniques[32,33]. Similarly, low-level visual features, such as the orientation of a particular edge, are encoded in the early visual cortex at a spatial scale of a few hundred micrometres[34]. Nevertheless, recent work demonstrates that pattern-based decoding of BOLD contrast fMRI signals acquired at relatively low spatial resolution can successfully predict the perception of such low-level perceptual features (FIG. 2). For example, the orientation[8,9], direction of motion[11] and even perceived colour[10] of a visual stimulus presented to an individual can be predicted by decoding spatially distributed patterns of signals from local regions of the early visual cortex. These spatially distributed response patterns might reflect biased low-resolution sampling by fMRI of slight irregularities in such high resolution feature maps[8,9] (FIG. 2a). Strikingly, despite the relatively low spatial resolution of conventional fMRI, the decoding of image orientation is possible with high accuracy[8] (FIG. 2b, right) and even from brief measurements of primary visual cortex (V1) activity[9]. Prediction accuracy can reach 80%, even when only a single brain volume (collected in under 2 seconds) is classified. By contrast, conventional univariate imaging analyses cannot detect such differences in activation produced by differently oriented stimuli, despite accumulating data over many hundreds of volumes[9]. Evidently, conventional approaches substantially underestimate the amount of information collected in a single fMRI measurement.

Primary visual cortex
Considered to be the first visual cortical area in primates, and receives the majority of its input from the retina via the lateral geniculate nucleus.

Decoding dynamic mental states

The work discussed so far has several important limitations. For example, in all cases the mental state of an individual was decoded during predefined and extended blocks of trials, during which the participants were either instructed to continuously imagine a cued stimulus, or during which a stimulus or class of stimuli were continuously presented under tight experimental control. These situations are not typical of patterns of thought and perception in everyday life, which are characterized by a continually changing 'stream of consciousness'[35]. To decode cognitive states under more natural conditions it would therefore be desirable to know whether the spontaneously changing dynamic stream of thought can be reconstructed from brain activity alone. One promising approach to such a complex question has been to study a simplified model

Box 2 | Statistical pattern recognition



Spatial patterns can be analysed by using the multivariate pattern recognition approach. In panel **a**, functional MRI measures brain activity repeatedly every few seconds in a large number of small volumes, or voxels, each a few millimetres in size (left). The signal measured in each voxel reflects local changes in oxygenated and deoxygenated haemoglobin that are a consequence of neural activity[18]. The joint activity in a subset (N) of these voxels (shown here as a 3x3 grid) constitutes a spatial pattern that can be expressed as a pattern vector (right). Different pattern vectors reflect different mental states; for example, those associated with different images viewed by the subject. Each pattern vector can be interpreted as a point in an N-dimensional space (shown here in panels **b–e** for only the first two dimensions, red and blue indicate the two conditions). Each measurement of brain activity corresponds to a single point. A successful classifier will learn to distinguish between pattern vectors measured under different mental states. In panel **b**, the classifier can operate on single voxels because the response distributions (red and blue Gaussians) are separable within individual voxels. In panel **c**, the two categories cannot be separated in individual voxels because the distributions are largely overlapping. However, the response distributions can be separated by taking into account the combination of responses in both voxels. A linear decision boundary can be used to separate these two-dimensional response distributions. Panel **d** is an example of a case where a linear decision boundary is not sufficient and a curved decision boundary is required (corresponding to a nonlinear classifier). In panel **e**, to test the predictive power of a classifier, data are separated into training and test data sets. Training data (red and blue symbols) are used to train a classifier, which is then applied to a new and independent test data set. The proportion of these independent data that are classified either correctly (open circle, 'correct') or incorrectly (filled circle, 'error') gives a measure of classification performance. Because classification performance deteriorates dramatically if the number of voxels exceeds the number of data points, the dimensionality can be reduced by using, for example, principal component analyses[14], downsampling[44] or voxel selection, according to various criteria[7]. An interesting strategy to avoid the bias that comes with voxel selection is to systematically search through the brain for regions where local clusters of voxels carry information[12]. Panels **b–d** modified, with permission, from REF. 2 © (2003) Academic Press.

Voxel
A voxel is the three-dimensional (3D) equivalent of a pixel; a finite volume within 3D space. This corresponds to the smallest element measured in a 3D anatomical or functional brain image volume.

Pattern vector
A vector is a set of one or more numerical elements. Here, a pattern vector is the set of values that together represent the value of each individual voxel in a particular spatial pattern.

Orientation tuning
Many neurons in the mammalian early visual cortex evoke spikes at a greater rate when the animal is presented with visual stimuli of a particular orientation. The stimulus orientation that evokes the greatest firing rate for a particular cell is known as its preferred orientation, and the orientation tuning curve of a cell describes how that firing rate changes as the orientation of the stimulus is varied away from the preferred orientation.

Spatial anisotropy
An anisotropic property is one where a measurement made in one direction differs from the measurement made in another direction. For example, the orientation tuning preferences of neurons in V1 change in a systematic but anisotropic way across the surface of the cortex.

Electroencephalogram
(EEG). The continuously changing electrical signal recorded from the scalp in humans that reflects the summated postsynaptic potentials of cortical neurons in response to changing cognitive or perceptual states. The EEG can be measured with extremely high temporal resolution.

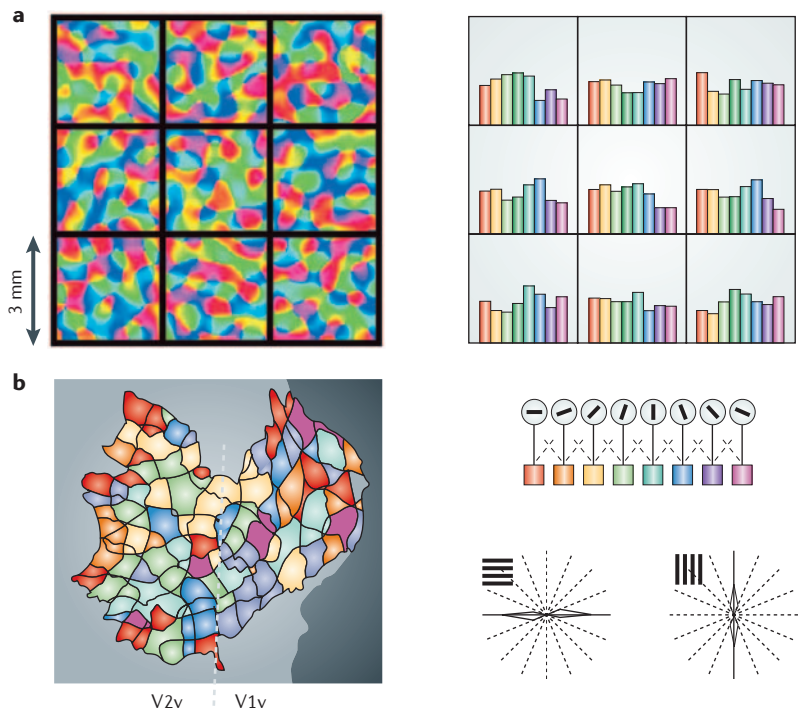


Figure 2 | **Decoding perceived orientation from sampling patterns in the early visual cortex. a** | In the primary visual cortex (V1) of primates, neurons with different orientation preferences are systematically mapped across the cortical surface, with regions containing neurons with similar orientation tuning separated by approximately 500 μm[33] (schematically shown in the left panel, where the different colours correspond to different orientations[106]). The cortical representation of different orientation preferences should therefore be too closely spaced to be resolved by functional MRI at conventional resolutions of 1.5–3 mm (indicated by the measurement grid in the left panel). Nonetheless, simulations[9] reveal that slight irregularities in these maps cause each voxel to sample a slightly different proportion of cells with different tuning properties (right panel), leading to potential biases in the orientation preference of each voxel. **b** | When subjects view images consisting of bars with different orientations, each orientation causes a subtly different response pattern in the early visual cortex[8,9]. The map shows the spatial pattern of preferences for different orientations in V1 and V2 plotted on the flattened cortical surface[8] (v, ventral). Although the preference of each small measurement area is small, the perceived orientation can be reliably decoded with high accuracy when the full information in the entire spatial response pattern is taken into account. The right panel shows the accuracy with which two different orthogonal orientation stimuli can be decoded when a linear support vector classifier is trained to classify the responses to different orientations[8]. It is important to note that, in principle, such sampling bias patterns should be observable for any fine-grained cortical microarchitecture such as object feature columns[32,33] and ocular dominance columns[10,37]. Although such sampling biases provide a potentially exciting bridge between macroscopic techniques such as fMRI and single unit recording, it is important to note that it will not be possible to directly infer the orientation tuning width of individual neurons in the visual cortex from the tuning functions of these voxel ensembles. The tuning width of the voxel ensemble will depend on many parameters[107] other than the tuning of single neurons, such as the spatial anisotropy in the distribution of cells with different orientation preferences, the voxel size, the signal-to-noise level of the fMRI measurements, and also the pattern and spatial scale of neurovascular coupling giving rise to the blood-oxygen-level-development signal[108,109]. However, such sampling bias patterns are important because they provide a means of non-invasively revealing the presence of feature-specific information in human subjects.

system where dynamic changes in conscious awareness are limited to a small number of possibilities.

Binocular rivalry is a popular experimental paradigm for studying spontaneous and dynamic changes in conscious perception[36]. When dissimilar images are presented to the two eyes, they compete for perceptual dominance so that each image is visible in turn for a few seconds while the other is suppressed. Because perceptual transitions between each monocular view occur spontaneously without any change in physical stimulation, neural responses associated with conscious perception can be distinguished from those due to sensory processing. This provides an opportunity to study brain reading of conscious percepts rather than of stimuli. Signals from the visual cortex can distinguish different dominant percepts during rivalry (for a review, see REF. 37). However, these studies relied on averaging brain activity across many individual measurements to improve signal quality, and therefore cannot address the question of whether perception can be decoded on a second-by-second basis. Signals from the scalp electroencephalogram (EEG) can dynamically reflect the fluctuating perceptual dominance of each eye during binocular rivalry[38], and therefore can be used to track the time course of conscious perception. However, the limited spatial precision of EEG leaves the precise cortical sources of these signals unclear.

We recently demonstrated that perceptual fluctuations during binocular rivalry can be dynamically decoded from fMRI signals in highly specific regions of the early visual cortex[10]. This was achieved by training a pattern classifier to distinguish between the distributed fMRI response patterns associated with the dominance of each monocular percept. The classifier was then applied to an independent test dataset to attempt dynamic prediction of any perceptual fluctuations. Dynamic prediction of the currently dominant percept during rivalry was achieved with high temporal precision (FIG. 3a). This also revealed important differences between brain regions in the type of information on which successful decoding depends. Decoding from V1 is based on signals that reflect which eye was dominant at the time. By contrast, decoding from the extrastriate visual cortex is based on signals that reflect the dominant percept[10]. Similar time-resolved decoding approaches can also reveal dissociations between the relative timing of the neural representation of information and its subsequent availability for report by the participant. For example, patterns of cortical activity associated with different types of picture reappear in the visual cortex several seconds before the participant verbally reports recall of the particular picture[16]. Therefore, decoding can provide important insights into the way in which information is encoded in different brain areas and into the dynamics of its access[39].

Natural scenes pose an even harder challenge to the decoding of perception. They are both dynamic and have added complexities compared to the simplified and highly controlled stimuli used in most experiments[40,41]. For example, natural visual scenes typically contain not just one but many objects that can appear, move and disappear independently. Under natural viewing conditions, individuals typically do not fixate a central fixation spot but freely move their eyes to scan specific paths[42]. This creates a particular problem for decoding spatially organized patterns from activity in retinotopic maps, as eye movements will create dynamic spatial shifts in such activity[43]. Natural scenes therefore provide a much greater challenge to the

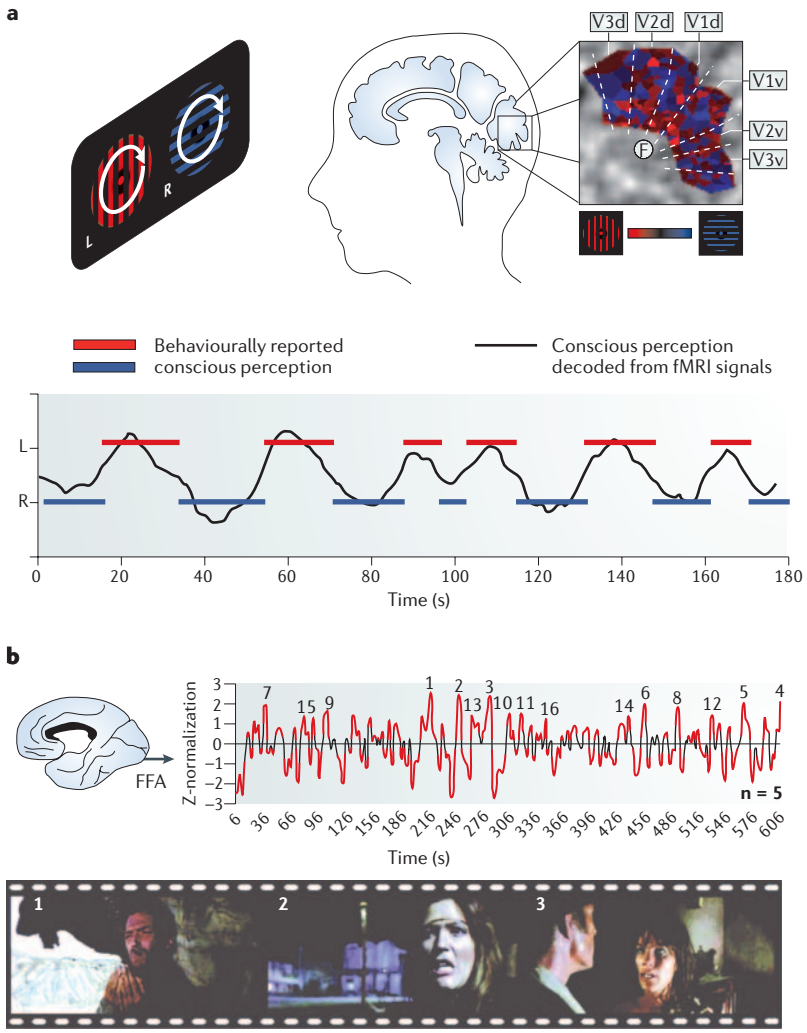


Figure 3 | **Tracking dynamic mental processes. a** | Decoding spontaneously fluctuating changes in conscious visual perception[10]. When conflicting stimuli are presented to each eye separately, they cannot be perceptually fused. Instead perception alternates spontaneously between each monocular image. Here, a rotating red grating was presented to the left eye and an orthogonal rotating blue grating was presented to the right eye. By pressing one of two buttons, subjects indicated which of the two grating images they were currently seeing. Distributed fMRI response patterns recorded concurrently showed some regions with higher signals during perception of the red grating and other regions with higher signals during perception of the blue grating (v, ventral; d, dorsal; F, fovea). A pattern classifier was trained to identify phases of red versus blue dominance based only on these distributed brain response patterns. By applying this classifier blindly to an independent test data set (lower panel) it was possible to decode with high accuracy which grating the subject was currently consciously aware of (red and blue lines, conscious perception; black line, conscious perception decoded from pattern signals from the early visual cortex, corrected for the haemodynamic latency). Note that this classifier decodes purely subjective changes in conscious perception despite there having been no corresponding changes in the visual input, which remains constant. **b** | Tracking perception under quasi-natural and dynamic viewing conditions using movies[40]. Subjects viewed a movie while their brain activity was simultaneously recorded. Activity in face-selective regions of the temporal lobes (the face fusiform area, FFA) was elevated when faces were the dominant content of the visual scene. The timecourse shows the average activity in FFA across several subjects. The peaks of this timecourse are indicated with numbers in descending order. The red sections of the plot indicate when the signal is significantly different from baseline. The bottom shows a snapshot of the visual scene for the first three peaks of FFA activity, revealing that when FFA activity was high the scenes were dominated by views of human faces[40]. Panel **a** modified, with permission, from REF. 10 © (2005) Elsevier Science, and from REF. 57 © (2004) American Association for the Advancement of Science.

decoding of perception. One initial approach is to study the pattern of correlated activity between the brains of different individuals as they freely view movies, therefore approximating viewing of a natural scene[40,41]. Under these dynamic conditions, signals in functionally specialized areas of the brain seem to reflect perception of fundamental object categories such as faces and buildings (FIG. 3b), and even action observation.

Decoding unconscious or covert mental states

Decoding approaches can be successfully applied to predict purely covert and subjective changes in perception when sensory input is unchanged. For example, it is possible to identify which one of two superimposed oriented stimuli a person is currently attending to without necessarily requiring them to explicitly report where their attention is directed[8]. This opens up the possibility of decoding covert information in the brain that is deliberately concealed by the individual. Indeed, in the domain of lie detection, limited progress has been made in decoding such covert states[44–46] (BOX 3). Although hypothetical at present, the possibility of decoding concealed states that are potentially or deliberately concealed by the individual raises important ethical and privacy concerns that will be discussed later.

Unconscious mental states constitute a special case of covert information that is even concealed from the individuals themselves. Various unconscious states have been experimentally demonstrated, such as the perceptual representation of invisible stimuli[47,48], or even unconscious motor preparation[49]. Decoding-based approaches seem to be particularly promising, both for predicting such unconscious mental states and for characterizing their temporal dynamics. Spatially distributed patterns of fMRI signals in human V1 can be used to decode the orientation of an oriented stimulus, even when it is rendered invisible to the observer by masking[9] (FIG. 4a). This has important implications for theoretical models of human consciousness, as it shows that some types of informational representation in the brain have limited availability for conscious access. The ability to decode the orientation of an invisible stimulus from activity in V1 indicates that under conditions of masking, subjects are unable to consciously access information that is present in this brain area. Furthermore, such findings have important implications for characterizing the neural correlates of consciousness[50], by showing that feature-specific activity in human V1 is insufficient for awareness. So, not only can decoding-based approaches address specific biological questions, but they also provide a general approach to study how informational representation might differ in conscious and unconscious states.

Similar approaches could, in principle, also be applied to higher cortical areas or to more complex cognitive states. For example, neural representations of unconscious racial biases can be identified in groups of human participants[51]. A decoding approach might be able to predict specific biases on an individual basis. Similarly, brain processes might contain information that reflects an unconscious motor intention immediately preceding a voluntary action[52,53] (FIG. 4b). Although

Box 3 | **Lie detection**

Some progress has been made in decoding a specific type of covert mental state; the detection of deception. The identification of lies is of immense importance, not only for everyday social interactions, but also for criminal investigations and national security. However, even trained experts are very poor at detecting deception[91]. Therefore, it was proposed almost a century ago that physiological measures of emotional reactions during deception might, in principle, be more useful in distinguishing innocent and guilty suspects[92,93]. Several physiological indicators of emotional reactions have been used for lie detection such as blood pressure[92], respiration[94], electrodermal activity[95] and even voice stress analysis[96] or thermal images of the body[97]. These approaches have variable success rates and the use of 'polygraphic' tests, where several physiological responses are simultaneously acquired, has been heavily debated. Criticism has been raised about the lack of research on their reliability in real-world situations[98], the adequacy of interrogation strategies[99] and whether test results can be deliberately influenced by covertly manipulating level of arousal[100].

Many of these problems might be due to the reliance of polygraphy on measuring deception indirectly, by the emotional arousal caused in the PNS. These indicators are only indirectly linked to the cognitive and emotional processing in the brain during deception. To overcome this limitation, direct measures of brain activity (such as evoked electroencephalogram responses[101] and functional MRI[44–46,102–105]) have been explored as a basis for lie detection, with the aim of directly measuring the neural mechanisms involved in deception. This could potentially allow the identification of intentional distortions of test results, therefore increasing the sensitivity of lie detection. Several recent studies have now investigated the feasibility of using fMRI responses to detect individual lies in individual subjects. Information in several individual brain areas (particularly the parietal and prefrontal cortex) can be used to detect deception[45,46], and this can be improved by combining information across different regions[45], especially when directly analysing patterns of distributed spatial brain response[44]. These brain measures of deception might be influenced less by strategic countermeasures[100], as would be expected when directly measuring the cognitive and emotional processes involved in lying. An important challenge for future research in this field is now to assess whether these techniques can be usefully and reliably applied to lie detection in real-world settings.

these studies have not addressed the question of how precisely and accurately an emerging intention can be decoded on a trial-by-trial basis, they reveal the presence of neural information about intentions prior to their entering awareness. This raises the intriguing question of whether decoding-based approaches will in future be able to reveal unconscious determinants of human behaviour.

Technical and methodological challenges

The work reviewed above suggests that it is possible to use non-invasive neuroimaging signals to decode some aspects of the mental states of an individual with high accuracy and reasonable temporal resolution. However, these encouraging first steps should not obscure the fact that a 'general brain reading device' is still science fiction[54]. Technical limitations of current neuroimaging technology restrict spatial and temporal resolution. Caution is required when interpreting the results of fMRI decoding because the neural basis of the BOLD signal is not yet fully understood. Therefore, any information that can be decoded from fMRI signals might not reflect the information present in the spiking activity of neural populations[18]. Also, the high cost and limited transportability of fMRI and magnetoencephalography scanners impose severe restrictions on potential real-world applications. Of current technologies, only the recording of electrical or optical signals with EEG or near infrared spectroscopy[55] over the scalp might be considered portable and reasonably affordable. But even if highly sensitive and portable recording technology with similar resolution to microelectrode recordings were available for normal human subjects, there are a number of crucial methodological obstacles that stand in the way of a general and possibly even practical brain reading device.

Magnetoencephalography
A non-invasive technique that allows the detection of the changing magnetic fields that are associated with brain activity on the timescale of milliseconds.

Generalization and invariance. In almost all human decoding studies, the decoding algorithm was trained for each participant individually, for a fixed set of mental states and based on data measured in a single recording session. This is a highly simplified situation compared with what would be required for practical applications. An important and unresolved question is the extent to which classification-based decoding strategies might generalize over time, across subjects and to new situations. When training and test sets are recorded on different days, classification does not completely break down[2,8,10]. If adequate spatial resampling algorithms are used, then even fine-grained sampling bias patterns can be partly reproduced between different recording sessions. This accords with predictions from computational simulations[10].

More challenging than generalization across time is generalization across different instances of the same mental state. Typically, any mental state can occur in many different situations, but with sometimes subtle contextual variations. Successful decoding therefore requires accurate detection of the invariant properties of a particular mental state, to avoid the impossible task of training on the full set of possible exemplars. This, in turn, requires a certain flexibility in any classification algorithm so that it ignores irrelevant differences between different instances of the same mental state. Pattern classifiers trained on a subset of exemplars from a category can generalize to other features[8], new stimulation conditions[10] and even new exemplars[2]. This suggests that classification can indeed be based on the invariant properties of an object category rather than solely on low-level features. However, the ability to identify a unique brain pattern corresponding to an invariant property of several exemplars will strongly depend on the grouping of different mental states as belonging to one 'type'. If a category is chosen to include a very

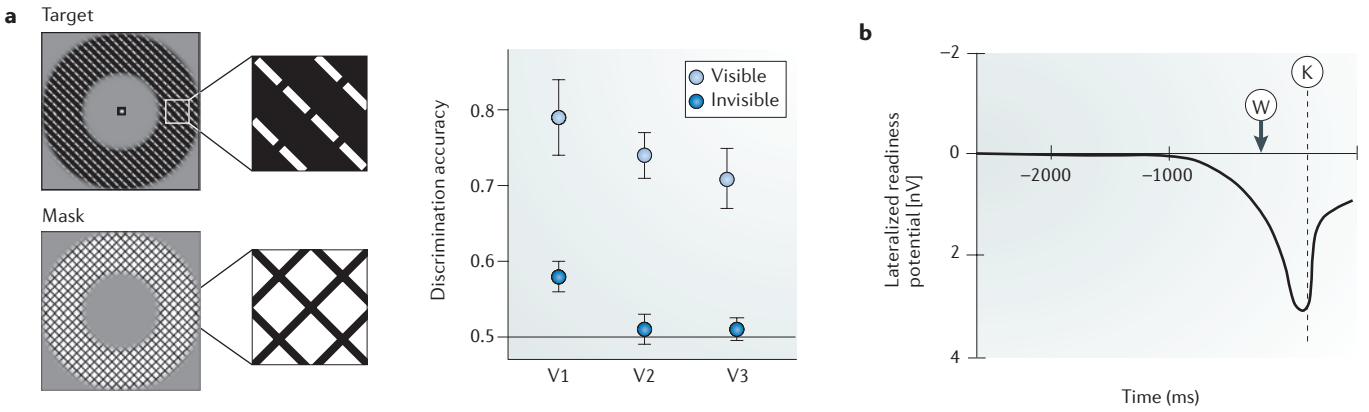


Figure 4 | **Decoding unconscious processing. a** | Decoding the orientation of invisible images[9]. Left panel, when oriented target stimuli are alternated rapidly with a mask, subjects subjectively report that the target appears invisible and are unable to objectively discern its orientation. Right panel, although such orientation information is inaccessible to the subject, it is possible to decode the orientation of the invisible target from activity in their primary visual cortex. However, target orientation cannot be decoded from V2 or V3, even though the orientation of fully visible stimuli can be decoded with high accuracy from these areas[9]. This indicates the presence of feature-selective information in V1 that is not consciously accessible to the human observer. **b** | Predicting the onset of a 'voluntary' intention prior to its subjective awareness. This study is a variant of the classical task by Benjamin Libet[52]. Here, subjects were asked to freely choose a timepoint and a response hand and then to press a response key (K). They were also asked to memorize, using a rotating clock, the time when their free decision occurred. Almost half a second prior to the subjective occurrence of the intention (W), a deflection was visible in the electroencephalogram (known as the lateralized readiness potential; solid line) that selectively indicated which hand was about to be freely chosen. Therefore, a brain signal can be recorded that predicts which of several options a subject is going to freely choose even before the subjects themselves become aware of their choice. The study does not resolve the degree to which it is possible to predict the outcome of a subjective decision on single trials. However, the application of decoding methods to such tasks might in future render it possible to reliably predict a subject's choice earlier than they themselves are able to do so. Panel **b** modified, with permission, from REF. 53 © (1999) Springer.

heterogeneous collection of mental states then it might not be possible to map them to a single neural pattern. Therefore, a careful categorization of mental states is required. It is also important to note that generalization is often achieved at the cost of decreased discrimination of individual exemplars. Any decoding algorithm must not only detect invariant properties of a mental state but also permit sufficient discrimination between individual exemplars. As a result, a successful classifier must carefully balance invariance with discriminability.

Finally, any decoding approach should ideally generalize not only across different recording sessions and different mental states but also across different individuals. For example, the ideal brain reading device for real-world applications such as lie detection would be one that is trained once on a fixed set of representative subjects, and subsequently requires little or even no calibration for new individuals. Although cases of such generalization have been reported[7,44,56], common application of decoding approaches will be strongly dependent on whether it is possible to identify functionally matching brain regions in different subjects associated with the mental state in question. Algorithms for spatially aligning and warping individual structural brain images to stereotactic templates are well established[57]. However, at the macroscopic spatial scale there is not always precise spatial correspondence between homologous functional locations in different individual brains, even when sophisticated alignment procedures are used[58]. Especially at a finer spatial scale, there are many situations where spatial matching is unlikely to be successful. For example, the pattern of orientation columns in V1 is strongly dependent on the early visual experience of an individual. Such intersubject variability will necessarily obscure any between-subject generalization of orientation-specific patterns[8,9]. Therefore, the extent to which cross-subject generalization is possible for specific mental states remains an open and intriguing question.

Measuring concurrent cognitive or perceptual states. To date, it is not clear whether it is possible to independently detect several simultaneously occurring mental states. For example, the current behavioural goal of an individual commonly coexists with simultaneous changes in their current focus of attention. Decoding two or more such mental states simultaneously requires some method to address superposition. A problem arises with such a decoding task because the spatial patterns indicating different mental states might spatially overlap. A method must be found that permits their separation if independent decoding is to be achieved. Previous research on 'virtual sensors' for mental states has demonstrated that a degree of separation and therefore independent measurement of mental states can be achieved using the simplified assumption that the patterns linearly superimpose[7]. However, it is currently unclear how to deal with cases where different mental states are encoded in the same neuronal population. Success in measuring

concurrent perceptual or cognitive states will be closely linked to the further improvement of statistical pattern recognition algorithms.

Extrapolation to novel perceptual or cognitive states. Perhaps the greatest challenge to brain reading is that the number of possible perceptual or cognitive states is infinite, whereas the number of training categories is necessarily limited. As long as this problem remains unsolved, brain reading will be restricted to simple cases with a fixed number of alternatives, for all of which training data are available. For example, it would be useful if a decoder could be trained on the brain activity evoked by a small number of sentences, but could then generalize to new sentences not in the training set. To generalize from a sparsely sampled set of measured categories to completely new categories, some form of extrapolation is required. This is possible if the underlying representational space can be determined, in which different mental states (or categories of mental state) are encoded. If the brain activation patterns associated with particular cognitive or perceptual states are indeed arranged in some systematic parametric space, this would allow brain responses to novel cognitive or perceptual states to be extrapolated. For some types of mental content, this appears to be possible. Multidimensional scaling indicates that the perceived similarity of relationships between objects is reflected in the corresponding similarity between distributed patterns of cortical responses to those objects in the ventral occipitotemporal cortex[15,30]. It may therefore be possible to classify response patterns to new object categories according to their relative location in an abstract shape space that is spanned by responses to measured training categories. Extrapolation is especially required to decode not just mental processes but also specific contents of cognitive or perceptual states, or even sentence-like semantic propositions where there are an infinite number of alternatives. For example, decoding of sentences has so far required training on each individual example sentence[56]. Generalizing such an approach to new sentences will ultimately depend on the ability to measure the neural structure of the underlying semantic representations.

Examining how a decoder (BOX 2) generalizes to new stimuli also provides knowledge about the invariance of the neural code in a given brain region (or at least how that neural activity is encoded in the fMRI signal). In this respect such techniques complement existing indirect methods, such as repetition priming or fMRI adaptation paradigms[59]. The latter technique relies on a repetition-associated change in neural activity (usually expressed as a decrease in the BOLD signal) across consecutive presentations of the same (or similar) stimuli. Such a change in signal is assumed to reflect neural adaptation or some other change in the stimulus representation, but the underlying mechanisms remain unclear. Importantly, the decoding approach reviewed here does not make any such assumptions, and so it might provide an important complement to the fMRI adaptation method in the future.

Scope. Despite the qualified successes of the decoding approach reviewed here, it remains unclear whether there are any limitations on which types of mental state could be decoded. Decoding depends intimately on the way in which different perceptual or cognitive states are encoded in the brain. In some areas of cognition, such as visual processing, a relatively large amount is known about the local organization of individual cortical areas. We and others have proposed that success in decoding the orientation of a visual stimulus from visual cortex activity depends on the topographic cortical organization of neurons with different selectivity for that feature. Topographic organization of neuronal selectivities is clearly a systematic feature of sensory and motor processing, but the extent to which it might also be associated with higher cognitive processes and different cortical areas remains unknown. The clustering of neurons with similar functional roles into cortical columns might be a ubiquitous feature of brain organization[60], even if it is currently unclear to what degree it is a principally necessary feature[61]. However, this ultimately remains an empirical issue, and it will be of crucial importance in future studies to identify any organizational principles for cortical areas associated with high-level cognitive processes such as goal-directed behaviour. Such enquiries will be helped by decoding-based research strategies that attempt to extract the information relevant to some mental state from local feature maps.

Finally, as with any other non-invasive method it is important to appreciate that decoding is essentially based on inverse inference. Even if a specific neural response pattern co-occurs with a mental state under a specific laboratory context, the mental state and pattern might not be necessarily or causally connected; if such a response pattern is found under a different context (such as a real-world situation), this might not be indicative of the mental state. Such inverse inference has been criticized in a number of domains of neuroimaging[62], and such cautions equally apply here.

Ethical considerations

Both structural and functional neural correlates have been identified for a number of mental states and traits that could potentially be used to reveal sensitive personal information without a person's knowledge, or even against their will. This includes neural correlates of conscious and unconscious racial attitudes[51,63,64], emotional states and attempts at their self-regulation[65,66], personality traits[67], psychiatric diseases[68,69], criminal tendencies[70], drug abuse[71], product preferences[72] and even decisions[73]. The existence of these neural correlates does not in itself reveal whether they can be used to decode the mental states or traits of a specific individual. This is an empirical question that has not yet been specifically addressed for real-world applications. In most cases it is unclear whether current decoding methods would be sensitive enough to reliably reveal such personal information for individual subjects. However, it is important to realize that the recent methodological advances reviewed here are likely to also enter these new areas, and lead to new applications.

Many potential applications are highly controversial, and as in many fields of biomedical research the benefits have to be weighed against potential abuse. Benefits

include the large number of important potential clinical applications, such as the ability to reveal cognitive activity in fully paralysed 'locked-in' patients[74,75], the development of brain–computer interfaces for control of artificial limbs or computers (BOX 1), or even the reliable detection of deception (BOX 3). But there are also potentially controversial applications. It might prove possible to decode covert mental states without an individual's consent or potentially even against their will. While this would be entirely prohibited by current ethical frameworks associated with experimental neuroscience, such established frameworks do not vitiate the need to consider such potential abuses of technology. Whereas the usual forms of communication, such as speech and body language, are to some extent under the voluntary control of an individual (for example, see BOX 3), brain reading potentially allows these channels to be bypassed. For example, brain reading techniques could, in theory, be used to detect concealed or undesirable attitudes during job interviews or to decode mental states in individuals suspected of criminal activity.

Such an ability to reveal covert mental states using neuroimaging techniques could potentially lead to serious violations of 'mental privacy'[76]. Further development of this area highlights even further the importance of ethical guidelines regarding the acquisition and storage of brain scanning results outside medical and scientific settings. In such settings, ethical and data protection guidelines generally consider it essential to treat the results of brain scanning experiments with similar caution and privacy as with the results of any other medical test. Maintaining the privacy of brain scanning results is especially important, because it might be possible to extract collateral information besides the medical, scientific or commercial use originally consented to by the subject. For example, information might be contained within structural or functional brain images acquired from an individual for a particular use that is also informative about completely different and unrelated behavioural traits.

Although the ability to decode mental states as reviewed here is still severely limited in practice, the effectiveness of brain reading has nevertheless been considered sufficient to justify the founding of several commercial enterprises offering services such as neuromarketing[77] and the detection of covert knowledge[78]. Because little information on the reliability of these commercial services is available in peer-reviewed journals it is difficult to assess how reliable such commercial services are. However, it has been noted that there is a tendency in the media and general public to overestimate the conclusions that can be drawn from neuroimaging findings[76]. Careful and considered public engagement by the neuroimaging community is therefore a prerequisite for informed dialogue and discussion about these important issues.

Conclusions and future perspectives

Decoding-based approaches show great promise in providing new empirical methods for predicting cognitive or perceptual states from brain activity. Specifically, by explicitly taking into account the spatial pattern of responses across the cortical surface, such approaches seem to be capable of revealing new details about the way in which cognitive and perceptual states are encoded in patterns of brain activity. Conventional univariate approaches do not take such spatially distributed information into account, and so these new methods provide a complementary approach to determining the relationship between brain activity and mental states. We believe that decoding-based approaches will find increasing application in the analysis of non-invasive neuroimaging data, in the service of understanding how perceptual and cognitive states are encoded in the human brain. However, the success of this approach will also depend on addressing important technical and methodological questions, especially regarding invariance, superposition and extrapolation. The currently emerging applications that allow the practical prediction of behaviour from neuroimaging data also raise ethical concerns. They can potentially lead to serious violations of mental privacy, and therefore highlight even further the importance of ethical guidelines for the acquisition and storage of human neuroimaging data.

1. Farah, M. J. Emerging ethical issues in neuroscience. *Nature Neurosci.* **5**, 1123–1129 (2002).
2. Cox, D. D. & Savoy, R. L. Functional magnetic resonance imaging (fMRI) 'brain reading': detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* **19**, 261–270 (2003).
This study compares various classification techniques and outlines important principles of decoding-based fMRI research.
3. O'Craven, K. M. & Kanwisher, N. Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *J. Cogn. Neurosci.* **12**, 1013–1023 (2000).
4. Haxby, J. V. *et al.* Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).
One of the first studies using pattern-based analysis to investigate the nature of object representations in the human ventral visual cortex.
5. Carlson, T. A., Schrater, P. & He, S. Patterns of activity in the categorical representation of objects. *J. Cogn. Neurosci.* **15**, 704–717 (2003).
6. Mitchell, T. M. *et al.* Classifying instantaneous cognitive states from FMRI data. *AMIA Annu. Symp. Proc.* 465–469 (2003).
7. Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F. & Wang, X. Learning to decode cognitive states from brain images. *Machine Learning* **57**, 145–175 (2004).
8. Kamitani, Y. & Tong, F. Decoding the visual and subjective contents of the human brain. *Nature Neurosci.* **8**, 679–685 (2005).
The first application of multivariate classification to reveal processing of features in the primary visual cortex represented below the resolution of fMRI.
9. Haynes, J. D. & Rees, G. Predicting the orientation of invisible stimuli from activity in primary visual cortex. *Nature Neurosci.* **8**, 686–691 (2005).
This study directly compares perceptual performance with the performance of a decoder trained on fMRI-signals from the early visual cortex.
10. Haynes, J. D. & Rees, G. Predicting the stream of consciousness from activity in human visual cortex. *Curr. Biol.* **15**, 1301–1307 (2005).
This work reveals the potential power of multivariate decoding to track perception quasi-online on a second-to-second basis.
11. Kamitani, Y. & Tong, F. Decoding motion direction from activity in human visual cortex. *J. Vision* **5**, 152a (2005).
12. Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl Acad. Sci. USA* **103**, 3863–3868 (2006).
This study introduces the 'searchlight' approach that searches across the entire brain for specific local patterns that encode information about a cognitive or perceptual state.
13. LaConte, S., Strother, S., Cherkassky, V., Anderson, J. & Hu, X. Support vector machines for temporal classification of block design fMRI data. *Neuroimage* **26**, 317–329 (2005).
14. Mourao-Miranda, J., Bokde, A. L., Born, C., Hampel, H. & Stetter, M. Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. *Neuroimage* **28**, 980–995 (2005).

15. O'Toole, A., Jiang, F., Abdi, H. & Haxby, J. V. Partially distributed representation of objects and faces in ventral temporal cortex. *J. Cogn. Neurosci.* **17**, 580–590 (2005).
16. Polyn, S. M., Natu, V. S., Cohen, J. D. & Norman, K. A. Category-specific cortical activity precedes retrieval during memory search. *Science* **310**, 1963–1966 (2005).
17. Sidtis, J. J., Strother, S. C. & Rottenberg, D. A. Predicting performance from functional imaging data: methods matter. *Neuroimage* **20**, 615–624 (2003).
18. Logothetis, N. K. & Pfeuffer, J. On the nature of the BOLD fMRI contrast mechanism. *Magn. Reson. Imaging* **22**, 1517–1531 (2004).
19. Allison, T. *et al.* Face recognition in human extrastriate cortex. *J. Neurophysiol.* **71**, 821–825 (1994).
20. Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
21. Epstein, R. & Kanwisher, N. A cortical representation of the local visual environment. *Nature* **392**, 598–601 (1999).
22. Engel, S. A. *et al.* FMRI of human visual cortex. *Nature* **369**, 525 (1994).
23. Sereno, M. I. *et al.* Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science* **268**, 889–893 (1995).
24. Dehaene, S. *et al.* Inferring behavior from functional brain images. *Nature Neurosci.* **1**, 549–550 (1998).
25. Tsao, D. Y., Freiwald, W. A., Tootell, R. B. & Livingstone, M. S. A cortical region consisting entirely of face-selective cells. *Science* **311**, 670–674 (2006).
26. Downing, P. E., Jiang, Y., Shuman, M. & Kanwisher, N. A cortical area selective for visual processing of the human body. *Science* **293**, 2470–2473 (2001).
27. Cohen, L. *et al.* The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients, *Brain* **123**, 291–307 (2000).
28. Downing, P. E., Chan, A. W. Y., Peelen, M. V., Dodds, C. M. & Kanwisher, N. Domain specificity in visual cortex. *Cereb. Cortex* Dec 7 2005 (doi: 10.1093/cercor/bhj086).
29. Ishai, A., Schmidt, C. F. & Boesiger, P. Face perception is mediated by a distributed cortical network. *Brain Res. Bull.* **67**, 87–93 (2005).
30. Edelman, S., Grill-Spector, K., Kushnir, T. & Malach, R. Towards direct visualization of the internal shape space by fMRI. *Psychobiology* **26**, 309–321 (1998).
This early work is the first application of multivariate techniques to the study of object representation. Of special interest is the demonstration that shape space is reflected in the similarity space between evoked cortical responses.
31. Spiridon, M. & Kanwisher, N. How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron* **35**, 1157–1165 (2002).
32. Tanaka, K. Mechanisms of visual object recognition: monkey and human studies. *Curr. Opin. Neurobiol.* **7**, 523–529 (1997).
33. Wang, G., Tanaka, K. & Tanifuji, M. Optical imaging of functional organization in the monkey inferotemporal cortex. *Science* **272**, 1665–1668 (1996).
34. Obermayer, K. & Blasdel, G. G. Geometry of orientation and ocular dominance columns in monkey striate cortex. *J. Neurosci.* **13**, 4114–4129 (1993).
35. James, W. *The Principles of Psychology* (Henry Holt, New York, 1890).
36. Blake, R. & Logothetis, N. K. Visual competition. *Nature Rev. Neurosci.* **3**, 13–21 (2002).
37. Haynes, J. D., Deichmann, R. & Rees, G. Eye-specific effects of binocular rivalry in the human lateral geniculate nucleus. *Nature* **438**, 496–499 (2005).
38. Brown, R. J. & Norcia, A. M. A method for investigating binocular rivalry in real-time with the steady-state VEP. *Vision Res.* **37**, 2401–2408 (1997).
39. Hung, C. P., Kreiman, G., Poggio, T. & DiCarlo, J. J. Fast readout of object identity from macaque inferior temporal cortex. *Science* **310**, 863–866 (2005).
40. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G. & Malach, R. Intersubject synchronization of cortical activity during natural vision. *Science* **303**, 1634–1640 (2004).
41. Bartels, A. & Zeki, S. Functional brain mapping during free viewing of natural scenes. *Hum. Brain Mapp.* **21**, 75–85 (2004),
42. Yarbus, A. L. *Eye Movements and Vision* (Plenum, New York, 1967).
43. Duhamel, J. R., Colby, C. L. & Goldberg, M. E. The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* **255**, 90–92 (1992).
44. Davatzikos, C. *et al.* Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage* **28**, 663–668 (2005).
This study is the first demonstration that multivariate decoding can be applied to lie detection.
45. Langleben, D. D. *et al.* Telling truth from lie in individual subjects with fast event-related fMRI. *Hum. Brain Mapp.* **26**, 262–272 (2005).
46. Kozel, F. A. *et al.* Detecting deception using functional magnetic resonance imaging. *Biol. Psychiatry* **58**, 605–613 (2005).
47. Marcel, A. J. Conscious and unconscious perception: experiments on visual masking and word recognition. *Cognit. Psychol.* **15**, 197–237 (1983).
48. Reingold, E. M. & Merikle, P. M. Using direct and indirect measures to study perception without awareness. *Percept. Psychophys.* **44**, 563–575 (1988).
49. Dehaene, S. *et al.* Imaging unconscious semantic priming. *Nature* **395**, 597–600 (1998)
50. Crick, F., & Koch, C. Are we aware of neural activity in primary visual cortex? *Nature* **375**, 121–123 (1995)
51. Phelps, E. A. *et al.* Performance on indirect measures of race evaluation predicts amygdala activation. *J. Cogn. Neurosci.* **12**, 729–738 (2000).
52. Libet, B., Gleason, C. A., Wright, E. W., Pearl, D. K. Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain* **106**, 623–642 (1983).
53. Haggard, P. & Eimer, M. On the relation between brain potentials and the awareness of voluntary movements. *Exp. Brain. Res.* **126**, 128–133 (1999).
54. Tesla, N. Tremendous New Power to be Unleashed. *Kansas City Journal-Post* (10 Sep 1933).
55. Obrig, H. & Villringer, A. Beyond the visible — imaging the human brain with light. *J. Cereb. Blood Flow Metab.* **23**, 1–18 (2003).
56. Suppes, P., Han, B., Epelboim, J. & Lu, Z. L. Invariance between subjects of brain wave representations of language. *Proc. Natl Acad. Sci. USA* **96**, 12953–12958 (1999).
57. Friston, K. J. *et al.* Spatial registration and normalisation of images. *Hum. Brain Mapp.* **2**, 165–189 (1995).
58. Fischl, B., Sereno, M. I., Tootell, R. B. & Dale, A. M. High resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Brain Mapp.* **8**, 272–284 (1999).
59. Grill-Spector, K., Henson, R. & Martin, A. Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn. Sci.* **10**, 14–23 (2006).
60. Mountcastle, V. B. The columnar organization of the neocortex. *Brain* **120**, 701–722 (1997).
61. Horton, J. C. & Adams, D. L. The cortical column: a structure without a function. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **360**, 837–862 (2005).
62. Poldrack, R. A. Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* **10**, 59–63 (2006).
63. Cunningham, W. A. *et al.* Separable neural components in the processing of black and white faces. *Psychol. Sci.* **15**, 806–813 (2004).
64. Richeson, J. A. *et al.* An fMRI investigation of the impact of interracial contact on executive function. *Nature Neurosci.* **6**, 1323–1328 (2003).
65. Beauregard, M., Levesque, J. & Bourgouin, P. Neural correlates of conscious self-regulation of emotion. *J. Neurosci.* **21**, RC165 (2001).
66. Phan, K. L., Wager, T., Taylor, S. F. & Liberzon, I. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage* **16**, 331–348 (2002).
67. Canli, T. & Amin, Z. Neuroimaging of emotion and personality: scientific evidence and ethical considerations. *Brain Cogn.* **50**, 414–431 (2002).
68. McCloskey, M. S., Phan, K. L. & Coccaro, E. F. Neuroimaging and personality disorders. *Curr. Psychiatry Rep.* **7**, 65–72 (2005).
69. Pridmore, S., Chambers, A. & McArthur, M. Neuroimaging in psychopathy. *Aust. N. Z. J. Psychiatry* **39**, 856–865 (2005).
70. Raine, A. *et al.* Reduced prefrontal and increased subcortical brain functioning assessed using positron emission tomography in predatory and affective murderers. *Behav. Sci. Law* **16**, 319–332 (1998).
71. Childress, A. R. *et al.* Limbic activation during cue-induced cocaine craving. *Am. J. Psychiatry* **156**, 11–18 (1999).
72. McClure, S. M. *et al.* Neural correlates of behavioral preference for culturally familiar drinks. *Neuron* **44**, 379–387 (2004).
73. Heekeren, H. R., Marrett, S., Bandettini, P. A. & Ungerleider, L. G. A general mechanism for perceptual decision-making in the human brain. *Nature* **431**, 859–862 (2004).
74. Levy, D. E. *et al.* Differences in cerebral blood flow and glucose utilization in vegetative versus locked-in patients. *Ann. Neurol.* **22**, 673–682 (1987).
75. Laureys, S., Perrin, F., Schnakers, C., Boly, M. & Majerus, S. Residual cognitive function in comatose, vegetative and minimally conscious states. *Curr. Opin. Neurol.* **18**, 726–733 (2005).
76. Farah, M. J. Neuroethics: the practical and the philosophical. *Trends Cogn. Sci.* **9**, 34–40 (2005).
77. Brammer, M. Brain scam? *Nature Neurosci.* **7**, 1015 (2004).
78. Dickson, K. & McMahon, M. Will the law come running? The potential role of 'brain fingerprinting' in crime investigation and adjudication in Australia. *J. Law Med.* **13**, 204–222 (2005).
79. Dewan, E. M. Occipital alpha rhythm, eye position and lens accommodation. *Nature* **214**, 975–977 (1967).
80. Nicolelis, M. A. Actions from thoughts. *Nature* **409**, 403–407 (2001).
81. Andersen, R. A., Burdick, J. W., Musallam, S., Pesaran, B. & Cham, J. G. Cognitive neural prosthetics. *Trends Cogn. Sci.* **8**, 486–493 (2004).
82. Blankertz, B. *et al.* Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis. *IEEE Trans. Neural Syst. Rehabil. Eng.* **11**, 127–131 (2003).
83. Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G. & Vaughan, T. M. Brain–computer interfaces for communication and control. *Clin. Neurophysiol.* **113**, 767–791 (2002).
84. Wolpaw, J. R. & McFarland, D. J. Control of a two-dimensional movement signal by a noninvasive brain–computer interface in humans. *Proc. Natl Acad. Sci. USA* **101**, 17849–17854 (2004).
85. Kreiman, G., Koch, C. & Fried, I. Category-specific visual responses of single neurons in the human median temporal lobe. *Nature Neurosci.* **3**, 946–953 (2000).
86. Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C. & Fried, I. Invariant visual representation by single neurons in the human brain. *Nature* **435**, 1102–1107 (2005).
Here, multivariate decoding is applied to simultaneous recordings of spike trains from the human medial temporal lobe.
87. Kennedy, P. R. & Bakay, R. A. Restoration of neural output from a paralyzed patient by a direct brain connection. *Neuroreport* **9**, 1707–1711 (1998).
88. Birbaumer, N. *et al.* A spelling device for the paralysed. *Nature* **398**, 297–298 (1999).
89. Weiskopf, N. *et al.* Self-regulation of local brain activity using real-time functional magnetic resonance imaging (fMRI). *J. Physiol. (Paris)* **98**, 357–373 (2004).
90. Stanley, G. B., Li, F. F. & Dan, Y. Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *J. Neurosci.* **19**, 8036–8042 (1999).
91. Ekman, P. & O'Sullivan, M. Who can catch a liar? *Am. Psychol.* **46**, 913–920 (1991).
92. Marston, W. M. The systolic blood pressure symptoms of deception. *J. Exp. Psy.* **2**, 117–163 (1917).
93. Geddes, L. A. History of the polygraph, an instrument for the detection of deception. *Biomed. Eng.* **8**, 154–156 (1973).
94. Burtt, H. E. The inspiration–expiration ratio during truth and falsehood. *J. Exp. Psy.* **4**, 1–23 (1921).
95. Thackeray, R. J. & Orne, M. T. A comparison of physiological indices in detection of deception. *Psychophysiol.* **4**, 329–339 (1968).
96. Horvath, F. Detecting deception: the promise and the reality of voice stress analysis. *J. Forensic Sci.* **27**, 340–351 (1982).
97. Pavlidis, I., Eberhardt, N. L. & Levine, J. A. Seeing through the face of deception. *Nature* **415**, 35 (2002).

98. Pollina, D. A., Dollins, A. B., Senter, S. M., Krapohl, D. J. & Ryan, A. H. Comparison of polygraph data obtained from individuals involved in mock crimes and actual crime investigations. *J. Appl. Psychol.* **89**, 1099–1105 (2004).
99. Lykken, D. T. *Tremor in the Blood: Uses and Abuses of the Lie Detector* (McGraw-Hill, New York, 1981).
100. Honts, C. R., Raskin, D. C. & Kircher, J. C. Mental and physical countermeasures reduce the accuracy of polygraph tests. *J. App. Psy.* **79**, 252–259 (1994).
101. Farwell, L. A. & Smith, S. S. Using brain MERMER testing to detect knowledge despite efforts to conceal. *J. Forensic Sci.* **46**, 135–143 (2001).
102. Spence, S. A. *et al.* Behavioral and functional anatomical correlates of deception in humans. *Neuroreport* **12**, 2849–2853 (2001).
103. Phan, K. L. *et al.* Neural correlates of telling lies: a functional magnetic resonance imaging study at 4 Tesla. *Acad. Radiol.* **12**, 164–172 (2005).
104. Ganis, G., Kosslyn, S. M., Stose, S., Thompson, W. L. & Yurgelun, T. Neural correlates of different types of deception: an fMRI investigation. *Cereb. Cortex* **13**, 830–836 (2003).
105. Lee, T. M. *et al.* Lie detection by functional magnetic resonance imaging. *Hum. Brain Mapp.* **15**, 157–164 (2002).
106. Boynton, G. Imaging orientation selectivity: decoding conscious perception in V1. *Nature Neurosci.* **8**, 541–542 (2005).
107. Nevado, Y., Young, M. P. & Panzeri, S. Functional imaging and neural information coding. *Neuroimage* **21**, 1083–1095 (2004).
108. Turner, R. How much cortex can a vein drain? Downstream dilution of activation-related cerebral blood oxygenation changes. *Neuroimage* **16**, 1062–1067 (2002).
109. Duvernoy, H. *The Human Brain* (Springer, New York, 1999).

Acknowledgements
This work was supported by the Wellcome Trust and the Mind–Science Foundation. We thank V. Lamme for bringing the reference to Nikola Tesla to our attention, and thank J. Driver, C. Frith and K.-E. Stephan for helpful comments on the manuscript.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Haynes's homepage: http://www.cbs.mpg.de/~haynes
Rees's laboratory: http://www.fil.ion.ucl.ac.uk/~grees
Access to this links box is available online.