Reading Hidden Intentions in the Human Brain

John-Dylan Haynes,1,2,3,4,5,6 Katsuyuki Sakai,6
Geraint Rees,4,5 Sam Gilbert,6 Chris Frith,9
and Richard E. Passingham6,7
1 Max Planck Institute for Human Cognitive
and Brain Sciences
04103 Leipzig
Germany
2 Bernstein Center for Computational Neuroscience
10115 Berlin
Germany
3 Charité – Universitätsmedizin
10117 Berlin
Germany
4 Institute of Cognitive Neuroscience
University College London
WC1N 3AR London
United Kingdom
5 Wellcome Department of Imaging Neuroscience
Institute of Neurology
University College London
WC1N 3BG London
United Kingdom
6 Department of Cognitive Neuroscience
Graduate School of Medicine
University of Tokyo
Tokyo 113-0033
Japan
7 Department of Experimental Psychology
University of Oxford
OX1 3UD Oxford
United Kingdom

Summary

When humans are engaged in goal-related processing, activity in prefrontal cortex is increased [1, 2]. However, it has remained unclear whether this prefrontal activity encodes a subject’s current intention [3]. Instead, increased levels of activity could reflect preparation of motor responses [4, 5], holding in mind a set of potential choices [6], tracking the memory of previous responses [7], or general processes related to establishing a new task set. Here we study subjects who freely decided which of two tasks to perform and covertly held onto an intention during a variable delay. Only after this delay did they perform the chosen task and indicate which task they had prepared. We demonstrate that during the delay, it is possible to decode from activity in medial and lateral regions of prefrontal cortex which of two tasks the subjects were covertly intending to perform. This suggests that covert goals can be represented by distributed patterns of activity in the prefrontal cortex, thereby providing a potential neural substrate for prospective memory [8–10].

During task execution, most information could be decoded from a more posterior region of prefrontal cortex, suggesting that different brain regions encode goals during task preparation and task execution. Decoding of intentions was most robust from the medial prefrontal cortex, which is consistent with a specific role of this region when subjects reflect on their own mental states.

Results

We directly addressed whether the current intentions of a subject were encoded in specific regions of prefrontal cortex. This was achieved by assessing whether multivariate pattern recognition [11] could be used to decode that subject’s covert intention from activity patterns in prefrontal cortex. If a cortical region indeed represents a current intention, it must have some way of encoding a set of different potential goals. One possibility is that it uses a spatial code, with different, spatially segregated neural subpopulations encoding different intentions. Unfortunately, because of the limited spatial resolution of human neuroimaging, most researchers have restricted their analyses to activity averaged across extended regions of cortex. This leaves unclear whether there are any regions encoding intentions in a spatially distributed fashion. However, it has recently emerged that functional magnetic resonance imaging can be used to study fine-grained neural representations, even when they are encoded at a finer scale than the resolution of the measurement grid [12, 13]. This technique is powerful enough to reveal distributed representations of visual images in occipital and temporal brain areas [11–14].

In order to investigate whether a subject’s current intentions are reflected in such distributed response patterns in prefrontal cortex, we required subjects to freely select what task they wished to perform. Specifically, they chose either adding or subtracting two numbers (Figure 1). After the subject had freely decided upon one of the two tasks, there was a variable delay of between 2.7 and 10.8 s, after which the task-relevant material (two numbers) was presented. The variable delay rendered the onset of the task-relevant material unpredictable and thus required the subject to maintain a high state of preparation even across long intervals [10]. Shortly after the two numbers, a response screen was presented that contained four numbers: one was the correct answer for addition, one was the correct answer for subtraction, and the two other numbers were similar but incorrect numbers. Subjects only rarely chose one of the two incorrect numbers (average 5%), indicating that they were correctly performing the task and not responding randomly. From the choice of one of the two correct answers, it was possible to infer which task the subject had chosen for the current trial. However, it is important to note that there was no way of telling which task the subject had freely selected prior to the response, because there was no explicit instruction.

*Correspondence: haynes@cbs.mpg.de
Figure 1. Delayed Intention Task
At the beginning of each trial, the word “select” was presented that instructed the subjects to freely and covertly choose one of two possible tasks, addition or subtraction. After a delay during which subjects covertly maintained their intention, two numbers were presented and subjects were then required to perform the selected task (addition or subtraction) on the two numbers. A response screen then appeared showing two correct answers (for either addition or subtraction) and two incorrect answers. Subjects pressed a button to indicate which answer was correct for the task they had performed. From the button press, it was possible to determine the covert intention of the subject during the previous delay period.

Discussion

To summarize, we have demonstrated that regions of both medial and lateral prefrontal cortex contain localizable task-specific representations of freely chosen intentions. In accordance with our findings, activity in several regions of human prefrontal cortex (including the frontopolar, lateral, medial, and prefrontal cortex) is increased during diverse executive processes such as attending to and thinking about intentions [16, 17], task-switching [18–20], set-shifting [21], multitasking [22], storing goals over a delay period [9, 10, 15, 23], branching and processing of subgoals [24, 25], and free task selection [26]. However, these previous studies left unclear whether any region of prefrontal cortex actually encodes signals that are specific for the current task. Increased levels of activity during task preparation might instead reflect unspecific preparatory signals, such as maintaining a representation of the set of all potential choices [6], tracking the memory of previous responses [7], or general preparation. Our new findings resolve this crucial question by showing for the first time that prefrontal cortex encodes information that is specific to a task currently being prepared by a subject, as would be required for regions encoding a subject’s intentions. In accordance with our findings, single cells in monkey lateral prefrontal cortex can prospectively encode expectations about task-relevant information [27]. Furthermore, ensembles of neurons in this region exhibit a gradual increase in the information about simple saccadic movement sequences while animals learn to perform a sequence correctly [28]. Cells have also been reported in the same area that code for specific moves while the monkey is waiting to move a cursor so as to negotiate a maze [29]. Here we show that in humans, a network of brain regions, including not only lateral but also medial prefrontal cortex, contains such task-specific representations.

Although intention-related information was encoded in both lateral and medial regions of prefrontal cortex, decoding accuracy was highest in the medial region.
One possible explanation may be that the current study allowed subjects to freely select which task to perform, whereas in most previous studies the task goal was specified by the experimenter. Medial prefrontal cortex is especially involved in the initiation of willed movements and their protection against interference [30]. Increased levels of medial activation are also found when task sets have to be internally generated as opposed to being fully externally cued [26, 31] and in similar cases of underdetermination [32–34]. For example, a direct comparison between voluntary task selection and externally cued tasks shows increased activation in medial, not lateral, prefrontal regions [31] during the free selection of tasks. Medial frontal cortex is also activated when subjects reflect on their own mental states [35, 36]. Interestingly, we also observed a division of labor between posterior and anterior regions of medial prefrontal cortex, where the anterior regions encode goals during preparation whereas the posterior regions encode goals during task execution (Figure 2). This is consistent with suggestions by previous authors that there is an anterior/posterior gradient on the medial frontal surface [26]. Similarly, the area at which activity is enhanced when subjects attend to their intention to perform a simple finger movement is more anterior than the area at which activity is enhanced when they attend to the finger movement itself [17, 34].

An interesting question for future research is the degree to which the encoding in different prefrontal areas reflects sustained maintenance of intentions across multiple trials, or trial-by-trial switching of intentions. Because of the random trial-by-trial alternation of subjects in our study, we measured predominantly shift-trials (see Figure S1), and therefore we are unable to address the differences between sustained and transient encoding. Previous studies have reported evidence for both sustained [20] and shift-related [37] activity in medial prefrontal cortex, so it would be interesting to apply similar decoding-based techniques to cued paradigms where the number of switch and stay trials can be better balanced.

An important feature of our paradigm is that it ensured that subjects could not covertly prepare for a specific movement prior to the onset of the response-mapping screen. This is in contrast to previous studies on prospective coding that used event-related potentials. These studies have shown that it is possible to decode on a single trial whether subjects are going to choose to move the left or right fingers [5] and that this information is present even prior to the time at which the subjects believe themselves to be making a decision [4]. However, these signals are recorded over motor-related brain regions and thus are likely to reflect the covert preparation of specific motor programs immediately preceding the execution of a movement [4, 5]. In our study, we can rule out the possibility that decoding during the delay was based on motor preparation. Please note that because the task-relevant stimuli and the response screen occurred in tight temporal sequence, we are unable to separate motor preparation from encoding of intentions during the execution period. An interesting question is whether one could decode which

---

Figure 2. Brain Regions Encoding the Subjects’ Specific Intentions during Either Delay or Execution Periods

In order to search in an unbiased fashion for informative voxels, we used a “searchlight” approach [44], which examines the information in the local spatial patterns surrounding each voxel $v_i$. Left: A spherical searchlight centered on one voxel ($v_i$) was used to define a local neighborhood. For each scanning run, the spatial response pattern in this local spherical cluster was extracted during preparation of either subtraction or addition. We then trained a pattern classifier with a subset of the data to recognize the typical response patterns associated with covert preparation of the two mathematical operations (see Experimental Procedures) and measured the local decoding accuracy. Then, the searchlight was shifted to the next spatial location. Middle: Highlighted in green are medial brain regions (superimposed on a sagittal slice of an anatomical template image) where this local classifier was able to decode significantly above chance which intention the subjects were covertly holding in an independent test data set. Highlighted in red are regions where it was possible to decode the intention during the execution of the task. Right: Decoding accuracy in searchlight locations with above-chance decoding during the delay period (MPFCa, anterior medial prefrontal cortex [MNI 3,42,15]; MPFCp, posterior medial prefrontal cortex [MNI 11,32,38]; LLFPC, left lateral frontopolar cortex [MNI –36,54,12]; LIFS, left inferior frontal sulcus [MNI –2,36,12]; RMFG, right middle frontal gyrus [MNI 48,24,45]; LFO, left frontal operculum [MNI –39,9,9]; error bars indicate SEM). In the anterior medial prefrontal cortex (MPFCa), decoding during the delay (green bars) was highest but was at chance level during the task execution (red bars) after onset of the task-relevant stimuli. In contrast, a more posterior and superior brain region on medial prefrontal cortex (MPFCp) encoded the chosen task only once it had entered the stage of execution, but not during the delay period. Several other regions of lateral prefrontal cortex also encoded information during the delay, but not during the execution period.
task our subjects were going to choose before they were aware of choosing it. Unfortunately, because of the rapid pacing of individual trials, we were not able to reliably analyze the period prior to the cued time of selection. However, an important implication of our study is that in future it might be possible to use decoding to reveal which specific brain areas unconsciously determine the intention that a subject is about to choose [38].

Importantly, we found that overall delay-related activity in prefrontal cortex was indistinguishable under both conditions. There was no evidence, therefore, that preparing to perform one task was more difficult than preparing to perform the other. Furthermore, this finding means that the two intentions are encoded, not by some increase in global activity, but by different spatial response patterns. This raises the intriguing question of the precise neural basis of these cortical patterns encoding different intentions, given that there is a strong overlap between cortical responses to different tasks [39]. One possible explanation may be that cells in specific regions of prefrontal cortex have a functional specialization for either of the two tasks, and that there is a fine-grained clustering of cells with similar properties that is smaller than the size of conventional areas. For example, in visual cortex, information encoded in similar fine-grained patterns of visual cortex can be read out by pattern recognition [12, 13]. This is typically explained as a “biased sampling” or “aliasing” of fine-grained feature columns by the individual fMRI voxels [11–13] and is confirmed by simulations based on realistic neural topographies (see Supplemental Data in [12]). This raises the question whether the informative spatial patterns we found might point to the existence of a similar columnar architecture in prefrontal cortex, where cells might be clustered according to similar roles in selective cognitive control. Such a columnar architecture has been highly debated as a general principle of cortical organization [40, 41] and has been claimed for the prefrontal cortex [42]. Alternatively, our sampling patterns might reflect the sampling of a distributed population code for different tasks as has been proposed from the findings of similar studies on object recognition [14]. Future optical imaging studies will be able to extend our findings by studying the local spatial topography of executive signals in prefrontal cortex. An important question for future studies will be whether the medial prefrontal cortex is generally involved in encoding specific tasks during intentional choices or whether encoding in this region is specific for tasks such as the preparation of addition and subtraction.

Taken together, our results extend previous studies on the processing of goals in prefrontal cortex in several important ways. They reveal for the first time that spatial response patterns in medial and lateral prefrontal cortex encode a subject’s covert intentions in a highly specific fashion. They also demonstrate a functional separation in medial prefrontal cortex, where more anterior regions encode the intention prior to its execution and more posterior regions encode the intention during task execution. These findings have important implications not only for the neural models of executive control, but also for technical and clinical applications, such as the further development of brain-computer interfaces, that might now be able to decode intentions that go beyond simple movements and extend to high-level cognitive processes.

Experimental Procedures

Participants and Experimental Design

Three male and five female subjects (age between 21 and 35) gave written informed consent to participate in the experiment, which was approved by the ethics committee at the Institute of Neurology, University College London. All subjects were right-handed and had normal or corrected to normal visual acuity.

At the beginning of each trial, a cue-word (“select”) was presented at fixation that instructed the subjects to rapidly select one of the two possible tasks (see Figure 1). This was followed by a variable delay of between 2.7 and 10.8 s, during which the subject was instructed to prepare for the task. Because of the variable delay, the onset time of the task-relevant stimuli was not predictable, requiring the subjects to maintain a state of continuous preparation across the extended delay [10, 15]. Then, the task-relevant stimuli were presented, which consisted of two 2-digit numbers presented above and below the fixation spot. Subjects were instructed to either add or subtract the two numbers in accordance with the task they had previously covertly chosen. Then after 2 s, a “response-mapping” screen was presented that showed four numbers, one in each visual quadrant on the screen. Two of these numbers were correct responses (one for addition and one for subtraction) and two were incorrect responses. Subjects responded with one of four response buttons operated by the left and right index and middle fingers. The keys corresponded to the positions of the four numbers on the “response mapping” screen. Please note that the decoding analysis (see below) was performed on signals related to brain activity prior to onset of the screen with task stimuli and thus 2 s before the response assignment, so decoding could not have been based on covert motor preparation because the mapping of correct and incorrect responses to keys was randomized from trial to trial. The distribution of phase durations during the main experiment (i.e., sequences of N trials where subjects chose the same task) followed an exponential distribution, as would be assumed if subjects chose randomly on each trial which task to perform (Figure S1). Prior to the experiment, subjects practiced the task for 7 min. During each scanning run, subjects performed 32 trials.

fMRI Acquisition

A Siemens Allegra 3T scanner with standard head coil was used to acquire functional MRI volumes (42 slices, TR = 2730 ms, resolution 3 × 3 × 1.5 mm³). For each subject, 8 runs of functional MRI data were acquired each with 155 images. To avoid susceptibility artefacts, slices were tilted 20° and the resolution in read-out direction was increased to 1.5 mm. The first three images of each run were discarded to allow for magnetic saturation effects.

Data Analysis

The fMRI data were motion corrected, spatially normalized to a standard stereotaxic space (Montreal Neurological Institute EPI template), and resampled to an isotropic spatial resolution of 3 × 3 × 3 mm³ in SPM2 (http://www.fil.ion.ucl.ac.uk/spm). The first analysis was designed to identify brain regions where activity was significantly increased during the delay period while subjects were preparing for the task (Figure S2). This analysis was performed with a general linear model as implemented in SPM2. The model consisted of four boxcar regressors, each convolved with a canonical haemodynamic response function. Each regressor modelled either the delay or execution period of one of the two task types. Prior to the GLM analysis, the data were smoothed with a Gaussian kernel of 6 mm FWHM to account for the anatomical variability across subjects and to satisfy the assumptions of Gaussian random field theory [43].

Pattern Classification

The second analysis was designed to identify regions where spatially distributed fMRI activation patterns carried information about the task the subject was preparing for. First, we estimated a modified general linear model as above but now based on unsmoothed data. This change was made to maximize sensitivity and allow extraction of the full information present in the spatial patterns of prefrontal
cortex, which would have been reduced by the smoothing. Then, in order to search in an unbiased fashion for informative voxels, we used a novel variant of the “searchlight” approach [44], which examines the information in the local spatial patterns surrounding each voxel \( v_i \) (see Figure 2, left). Thus, for each \( v_i \), we investigated whether its local environment contained spatial information that would allow decoding of the current intention.

For a given voxel \( v_i \), we first defined a small spherical cluster of \( N \) voxels \( v_{i-n} \) with radius of three voxels centered on \( v_i \). For each voxel \( v_{i-n} \) in the fixed local cluster, we extracted the unsmoothed parameter estimates for delay-period activity separately for covert preparation of the addition and the selection task. This yielded two \( N \)-dimensional pattern vectors \( x_{1,N} \) and \( y_{1,n} \) for each run \( r \), representing the spatial response patterns in the local cluster during covert preparation for addition and subtraction. Next, we used multivariate pattern recognition to assess how much intention-related information was encoded in the local pattern. To achieve this, we assigned the pattern vectors \( x_{1, n} \) and \( y_{1,n} \) for seven of the eight imaging runs to a “training” data set that was used to train a linear support vector pattern classifier [45] (with fixed regularization parameter \( C = 1 \)) to correctly identify response patterns related to the two different intentions the subject was currently holding. The classification was performed with the LIBSVM implementation (http://www.csie.ntu.edu.tw/~cjlin/libsvm). The amount of intention-related information present within this local cluster could then be assessed by examining how well the intentions during the remaining independent eighth or “test” data set were classified. Good classification implies that the local cluster of voxels spatially encodes information about the specific current intention of the subject. In total, the training and test procedure was repeated eight times, each with a different run assigned as test data set, yielding an average decoding accuracy in the local environment of the central voxel \( v_i \) (8-fold crossvalidation). Then, the procedure was repeated for the next spatial position at voxel \( v_i \). The average decoding accuracy for each voxel was then used to create a 3-dimensional spatial map of decoding accuracy for each position \( v_i \) in prefrontal cortex. Because the subjects’ images had previously been normalized to a common stereotactic template, it was possible to extend previous local decoding approaches [44] and perform a second-level analysis where we computed on a voxel-by-voxel basis how well decoding could be performed on average across all subjects from each position in the brain. This yielded a spatial map of average decoding accuracy that is plotted in green in Figure 2. We also performed a similar pattern classification with the parameter estimates for task execution as opposed to the delay period. This is plotted in Figure 2 in red.

Supplemental Data

Two Supplemental Figures can be found with this article online at http://www.current-biology.com/cgi/content/full/17/4/323/DC1/.

Acknowledgments

This work was supported by the Max Planck Society, the Wellcome Trust, and the Mind-Science Foundation. The authors would like to thank M. Brass and M. Ullsperger for valuable comments.

Received: September 11, 2006
Revised: November 24, 2006
Accepted: November 27, 2006
Published online: February 8, 2007

References


