Title: Flexible decision under ambiguous cues reverses neural signaling from occipitotemporal to prefrontal cortex

Authors: Kaho Tsumura,[1] Keita Kosugi,[1] Ryuta Aoki,[2] Masaki Takeda,[2] Junichi Chikazoe,[3] Kiyoshi Nakahara,[2] Koji Jimura[1,2]

Affiliations: [1] Department of Biosciences and Informatics, Keio University, Yokohama, Japan
[2] Research Center for Brain Communication, Kochi University of Technology, Kami, Japan
[3] Supportive Center for Brain Research, National Institute for Physiological Sciences, Okazaki, Japan

Correspondence should be addressed to:
Koji Jimura, Ph.D.
Department of Biosciences and Informatics,
Keio University,
3-14-1 Hiyoshi Kohoku-ku,
Yokohama, 223-0061, Japan
Email: jimura@bio.keio.ac.jp
Phone: +81-45-566-1625

Abstract

Adaptation to changing environments involves the appropriate extraction of environmental information to achieve a behavioral goal. It remains unclear how behavioral flexibility is guided under situations where the relevant behavior is ambiguous. Using functional brain mapping of machine-learning decoders and directional functional connectivity, we show that brain-wide reversible neural signaling underpins behavioral flexibility in ambiguously changing environments. When relevant behavior is cued ambiguously during behavioral shifting, neural coding is attenuated in distributed cortical regions, but top-down signals from the prefrontal cortex complement the coding. When behavioral shifting is cued more explicitly, modality-specialized occipitotemporal regions implement distinct neural coding about relevant behavior, and bottom-up signals from the occipitotemporal region to the prefrontal cortex supplement the behavioral shift. These results suggest that our adaptation to an ever-changing world is orchestrated by the alternation of top-down and bottom-up signaling in the fronto-occipitotemporal circuit depending on the availability of environmental information.

Introduction

Executive control guides flexible adaptation to changing environments, and is most developed in humans throughout evolution (1, 2). Shifting between different types of behavior is one of the core executive control functions (3, 4), and switching paradigms have been often used to investigate behavioral flexibility and its underlying neural mechanisms. Previous neuropsychological and neuroimaging studies of human and non-human animals suggest a critical role of the prefrontal cortex in switching tasks and rules (5-15) (Fig. 1A).

Importantly, executive control depends on perceived information of external environments, and relevant information is appropriately extracted from the external environment to achieve a behavioral goal. Perception of sensory information from the external environment guides the course of action, which is referred to as perceptual decision-making (16, 17). It involves the extraction of goal-relevant information, which is integrated to form a relevant decision. Studies of perceptual decision-making have used behavioral tasks that demand discrimination of sensory stimuli involving perceptual uncertainty (18-20).

By manipulating perceptual uncertainty, neurophysiological and neuroimaging studies have examined cortical mechanisms of the perceptual decision-making. For example, the middle temporal (MT) region is known to play an important role in the perception of moving stimuli (18, 21-24). It has been also suggested that the fusiform face area (FFA) and parahippocampal place area (PPA) are associated with the perception of face (25-29) and place (27, 28, 30) stimuli, respectively. These collective results suggest that temporal and occipital regions play important roles in perceptual decision-making, and are functionally segmented depending on the modality of the stimulus (Fig. 1A).

In our daily life, goal-relevant information in external environments is not always evident. Such situation can be well illustrated by an incorporation of behavioral shifting and perceptual decision-making, in which the behavioral goal was indicated by a cue involving perceptual uncertainty. Notably, the task switching paradigm is composed of hierarchically structured configuration of a set of task rules, called task sets (7, 31-33), allowing the manipulation of the uncertainty of task cue at the multi-task (most abstract) tier (Fig. 1A). As such, the uncertainty of multi-task level information

may provide novel opportunities to examine the relationships between executive control and perceptual decision-making.

One potential approach to elucidate underlying neural mechanisms is to identify the signal contents of responsible brain regions; this has recently been demonstrated for perception by neural decoding techniques (34-38). In particular, prior neuroimaging studies have shown that mental and behavioral states can be decoded from neural coding by machine learning techniques that classify distributed patterns of brain activity (34-39). One commonly used technique is the support vector machine (SVM) (34, 40-43), which enables categorical discrimination by dividing multidimensional space composed of brain activation pattern using a linear hyperplane. Convolutional neural network (CNN) classifier (44, 45) is one of the deep neural network (DNN) classifiers consisting of multiple feature-aggregating layers, which allows more robust classification, especially for 2D-images. Importantly, recent technical advancements of CNN allow mapping that highlights image locations characterizing a classified image (46). The mapping technique may provide novel information about neural coding and functional localization of the brain.

The current study aimed to elucidate relationships between behavioral flexibility and perceptual decision-making under cue uncertainty, and to explore the underlying neural mechanisms (Fig. 1A). Functional MRI was administered while human participants performed a task-switching paradigm with a cue involving perceptual uncertainty. Standard univariate analysis identified brain regions associated with task switching, motion strength, and task modality. In order to elucidate causal network dynamics during task switching under cue ambiguity, we examined effective connectivity among the task-related brain regions. Finally, whole brain exploratory analyses based on machine learning techniques, CNN and SVM, were performed to identify brain regions that coded relevant task information.

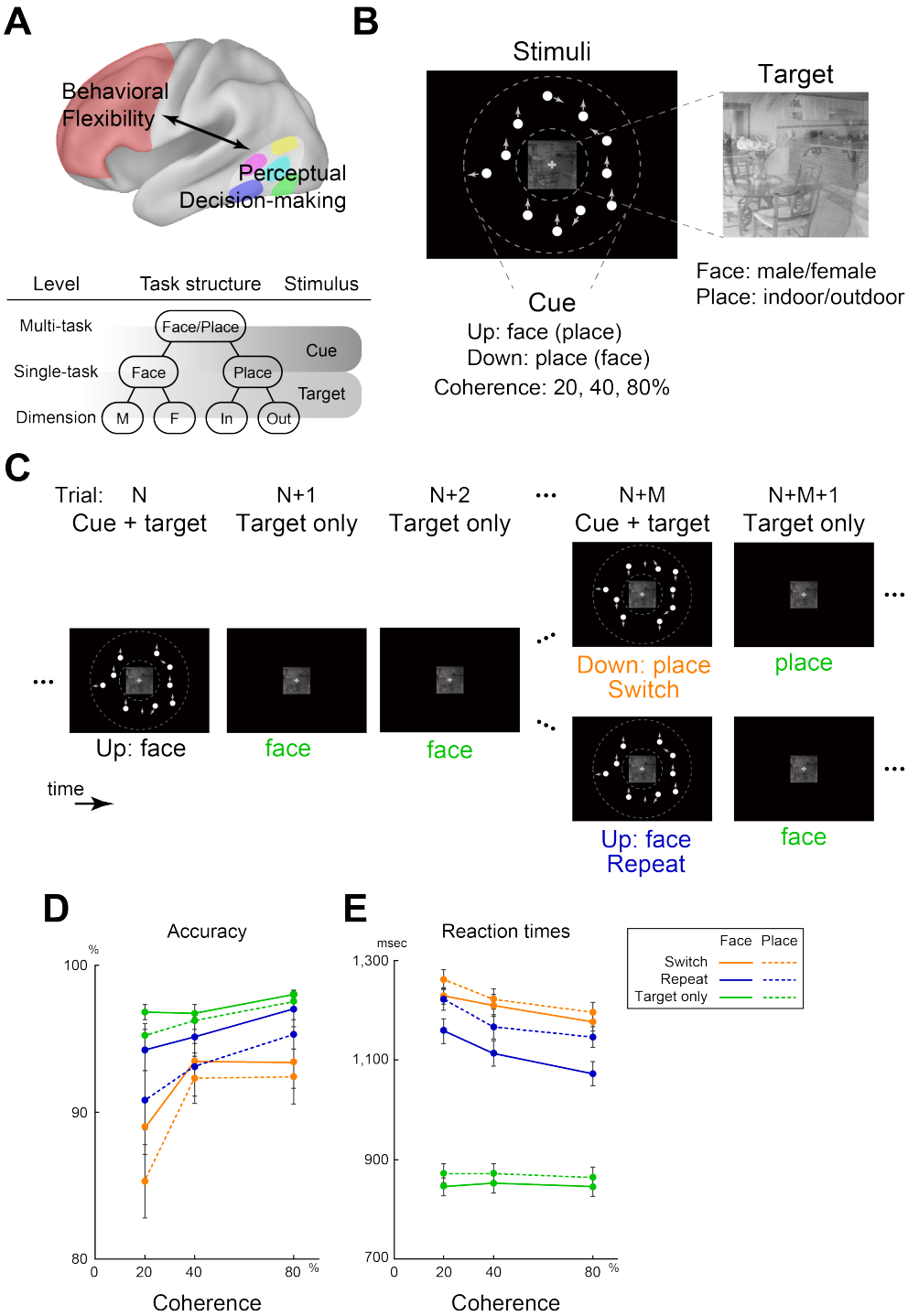


Fig 1. Experimental design and behavioral results.
(A) Schematic illustration of a brain-wide model of interaction of behavioral flexibility and perceptual decision-making. Behavioral flexibility associated with prefrontal areas is subject to appropriate perception of the external world implemented in stimulus-modality-dependent occipitotemporal areas (top), and is guided by cue indicating relevant behavior at the top layer in the hierarchical structure of the behavioral set (bottom). (B/C) Behavioral task. (B) Stimuli. Cue stimulus indicates the task to be

performed (face or place task) and is composed of a set of white moving dots presented within a donut-shaped display circle (indicated by dotted lines). The arrow indicates the motion direction of each dot, and overall motion was either upward or downward. Upward and downward motion indicate face and place task, respectively. Motion strength of the cue stimulus was manipulated by coherence of dot motion. The target image was superimposed picture of face and place, which was presented at the center of screen. Participants judged whether the face picture was male or female, or the place is indoor or outdoor, depending on the task to be performed. (C) Task procedure. Trials with simultaneous presentation of a dot cue and target (switch and repeat; N and N+M th trials) were followed by target-only trials (N+1, N+2 and N+M+1 th trials). Switch and repeat trials were dependent on the performed task prior to the cue presentation. (D/E) Behavioral results. Accuracy (D) and reaction times (E) as a function of motion coherence. Solid and dotted lines indicate face and place task, respectively, and orange, blue, and green lines indicate switch, repeat, and target-only trials, respectively. Error bars indicate standard error of the mean across participants.

Results

Behavioral results.

Human participants performed a task-switching paradigm (7, 31-33), in which they alternated discrimination tasks for face and place stimuli (Figs. 1B-C). The relevant task was indicated by a cue stimulus involving perceptual uncertainty, which was manipulated by the motion strength of randomly moving dots.

Accuracy was lower in low-coherence (i.e., more uncertain) trials compared to high-coherence trials [$F(1, 28) = 12.7$; $P < .005$; Fig. 1D], and became lower in switch trials than in repeat trials [$F(1, 28) = 11.8$; $P < .005$]. Likewise, reaction times (RTs) were longer in low-coherence trials compared to high-coherence trials [$F(1, 28) = 69.9$; $P < .001$; Fig. 1E], and were longer in switch trials than in repeat trials [$F(1, 28) = 43.2$; $P < .001$]. These behavioral results suggest that the current behavioral task successfully manipulated task switching (3-5, 7, 8, 33) and perceptual decision-making (17, 18, 21). The interaction effect of trial type (switch/repeat) and coherence levels did not reach statistical significance [$F(1,28) = 2.1$; $P = 0.155$].

In trials without cue presentation, occurring after the switch and repeat trials until the next cue trials (target only trial; Fig. 1C), accuracy was lower after lower coherence cue trials [$F(1,28) = 16.3$; $P <. 001$; Fig 1D], but RTs remained unchanged after lower coherence cue trials [$F(1,28) = 1.0$; $P = 0.317$], suggesting that switching might not complete immediately after low coherence cue trials. RTs were longer in place than face tasks [$F(1, 28) = 5.0$; $P <. 05$; Fig. 1E].

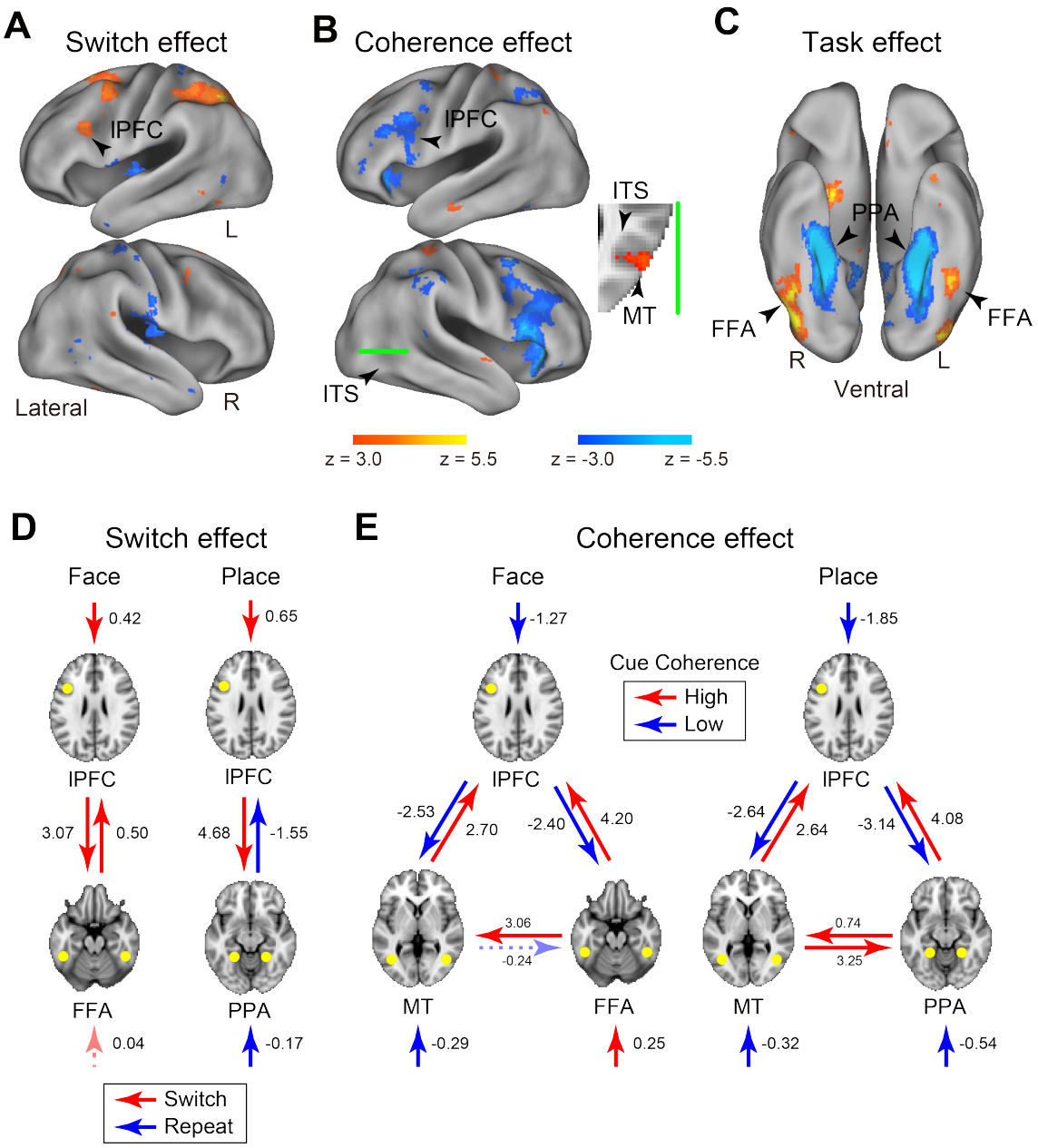


Fig 2. Whole-brain exploration activation and functional connectivity analyses
(A-C) Statistical activation map of univariate analysis and functional connectivity analysis. Maps are overlaid onto the 3D surface of the brain. Hot and cool colors indicate positive and negative effects, respectively. (A) Switch effect (switch minus repeat trials) (B) Motion coherence effect (high- vs. low-coherence trials). Green solid line on the surface indicates the axial section on the right. (C) Task effect (face minus place tasks in target only trials). lPFC: lateral prefrontal cortex; ITS: inferior temporal sulcus; MT: middle temporal; FFA: fusiform face area; PPA: parahippocampal place area. (D/E) Effective connectivity analysis. (D) Effective connectivity and extrinsic inputs modulated by the contrast switch vs. repeat trials during face (left) and place (right) tasks. Red and blue arrows indicate connectivity enhancements in switch and repeat trials, respectively. The values next to the arrows indicate the magnitude of connectivity enhancements (positive: switch > repeat; negative: repeat > positive). (E) Effective connectivity and extrinsic inputs modulated by motion coherence during face (left) and place

(right) tasks. Red and blue arrows indicate enhancements in high- and low-coherence trials, respectively. The values next to the arrows indicate the magnitude of connectivity (positive: high coherence > low coherence; negative: low coherence > high coherence). Arrows with solid line indicate statistically significant connectivity (P < .05, uncorrected).

Exploration of switch-related and stimulus-modality-dependent brain regions.

We first explored brain regions associated with task switching, motion coherence, and the perception of face and place based on univariate general linear model (GLM) analysis. Figure 2A shows brain regions showing significant increases and decreases in univariate brain activity during switch relative to repeat trials (P < .05 corrected with cluster-wise family-wise error rate based on non-parametric permutation tests; see Materials and Methods). Robust activation increases were observed in the left frontal regions including the inferior frontal cortex (IFC), dorsolateral prefrontal cortex (DLPFC), inferior frontal junction (IFJ), and pre-supplementary motor area (pre-SMA), and in left parietal regions including posterior parietal cortex (PPC), consistent with prior studies (5-11, 33). A full list of brain regions is shown in Table S1.

We then explored brain regions associated with motion coherence. Figure 2B shows brain regions showing significant modulation of brain activity in relation to motion coherence during the cue (i.e. switch and repeat) trials. In low-coherence trials, activation was increased in multiple fronto-parietal regions including IFC, DLPFC, IFJ, pre-SMA, and PPC (Fig. 2B and Table S2), consistent with a prior study (24). In contrast, activation was greater in high-coherence trials in the MT region (Fig 2B *right*; Table S2), which is also consistent with prior studies of perceptual decision-making for motion (18, 21-24).

We next explored brain regions associated with face and place tasks (Fig. 2C and Table S3). Consistent with prior studies of perception of face (25-29) and place (27, 28, 30), in the face task, activity was greater in the FFA, whereas in the place task, increased activity was observed in the PPA.

These collective univariate activation results suggest that the current univariate activation analysis successfully identified brain regions associated with task switching in addition to the perception of face, place and motion stimulus, and that those regions were cooperatively engaged in task switching when the cue stimulus involved perceptual uncertainty.

195 Reversal of functional connectivity depending on cue uncertainty.

196 The whole brain exploratory analyses of univariate activation identified three
197 types of brain regions: 1) left PFC associated with task switching and perceptual
198 uncertainty (Figs. 2A-B); 2) MT region associated with motion coherence of task cue
199 (Fig. 2B); and 3) FFA and PPA associated with discrimination of face and place stimuli,
200 respectively (Fig. 2C). One possible mechanism to explain these regions playing
201 differential roles in task switching with cue uncertainty is that lPFC, MT, and FFA/PPA
202 mutually received or sent task-related signals during switching, which was modulated
203 by cue coherence, such that the task-related signal complemented the engagement of
204 these regions depending on cue uncertainty and the task to be performed.

205 In order to test this hypothesis, we performed an interregional effective
206 connectivity analysis based on dynamic causal modeling (DCM) that allows the
207 examination of directionality of task-related functional connectivity based on the
208 state-space model (see Materials and Methods).

209 We first examined the task-related effective connectivity during task
210 switching, and found that the connectivity was enhanced from the lPFC toward the FFA
211 during switch relative to repeat trials of the face task (i.e. switch-to-face versus
212 repeat-face trials; Fig. 2D *left*), and also enhanced from the lPFC toward the PPA in
213 switch relative to repeat trials of the place task (Fig. 2D *right*). These results are in line
214 with the well-known role of the left lPFC in behavioral flexibility (4-7), and suggest
215 top-down signaling from the prefrontal cortex to stimulus-modality-dependent
216 occipitotemporal regions during task switching.

217 We then asked a critical question whether the top-down signaling from the
218 lPFC to the stimulus-modality-dependent regions is modulated depending on the
219 uncertainty of the relevant task (i.e., multi-task level) that was manipulated by the
220 motion coherence of task cue. During face tasks with high-coherence cue, the
221 task-related effective connectivity was enhanced from the MT and FFA regions to the
222 lPFC, and the directionality of connectivity between these regions was reversed in
223 low-coherence trials (Fig. 2E *left*). Likewise, the effective connectivity was enhanced
224 from the MT and PPA regions to the lPFC during the place task with the high coherent
225 cue, and the directionality of the connectivity was also reversed in low-coherence trials
226 (Fig. 2E *right*).

In order to test the reliability and robustness of the results above, we estimated the effective connectivity by 1) using an alternative estimation method (Figs. S1A-B), 2) changing the number of the regions of interest (ROIs) in the models (Figs. S1C-D), and 3) changing the definition of the ROIs (Figs. S1E-F) (see Materials and Methods). Overall results were maintained, confirming that the connectivity results were robust against estimation procedures, model structures, and model parameters. Task-unrelated intrinsic connectivity is shown in Figs S1G-H.

These collective results of effective connectivity suggest that, when relevant task was indicated ambiguously, top-down signal from prefrontal regions become stronger toward stimulus-modality-dependent occipitotemporal regions (i.e. MT and FFA/PPA for face/place tasks). On the other hand, when task cue information is more evident, bottom-up signals from the occipitotemporal regions to the prefrontal regions become stronger.

Whole-brain decoding by a CNN classifier.

Given the brain regions associated with tasks, switching tasks, and perceptual decision-making identified by univariate activation analysis and their effective connectivity mechanisms, we explored brain regions that code relevant task information using a convolutional neural network (CNN) classifier (45). More specifically, we examined whether brain activity patterns involve discriminable information about face and place tasks during task switching with cue uncertainty, and then identified brain regions involving critical information about the relevant task.

The current analysis used VGG16 (47) that was trained to classify a concrete object image dataset provided by ImageNet (44) (http://www.image-net.org/; Fig. S2A). We re-trained the VGG16-ImageNet model using flat whole cortical activation maps (Figs. 3A and S2B) such that it classified face and place tasks based on fine-tuning [48] (Fig. S2C). The retraining was performed based on cortical maps that were independent of the tested maps. We retrained the model using flat maps during a working memory (WM) task for face and place stimuli obtained from the Human Connectome Project (HCP) (Fig. S2C *top*; see also Materials and Methods), followed by additional retraining based on flat activation maps during target-only trials of the current task in which only the face/place target stimulus was presented without the dot cue stimulus (Figs. 1C and S2C *bottom*).

Classification accuracy for target-only trials was 82.1 ± 5.0% (mean ± SD with 10-fold cross validation), which was significantly greater than chance level ($P < .001$) (Fig S3A; see Materials and Methods). Interestingly, direct retraining of VGG16-ImageNet model to classify the current target-only trial maps showed little increase in accuracy (Fig. S3A). Training of a randomly initialized VGG model to classify HCP WM maps was also not successful (Fig. S3B). Thus, these results demonstrate that the current two-step retraining of the VGG16-ImageNet model was sufficient for the CNN model to learn from small sample data sets of brain images.

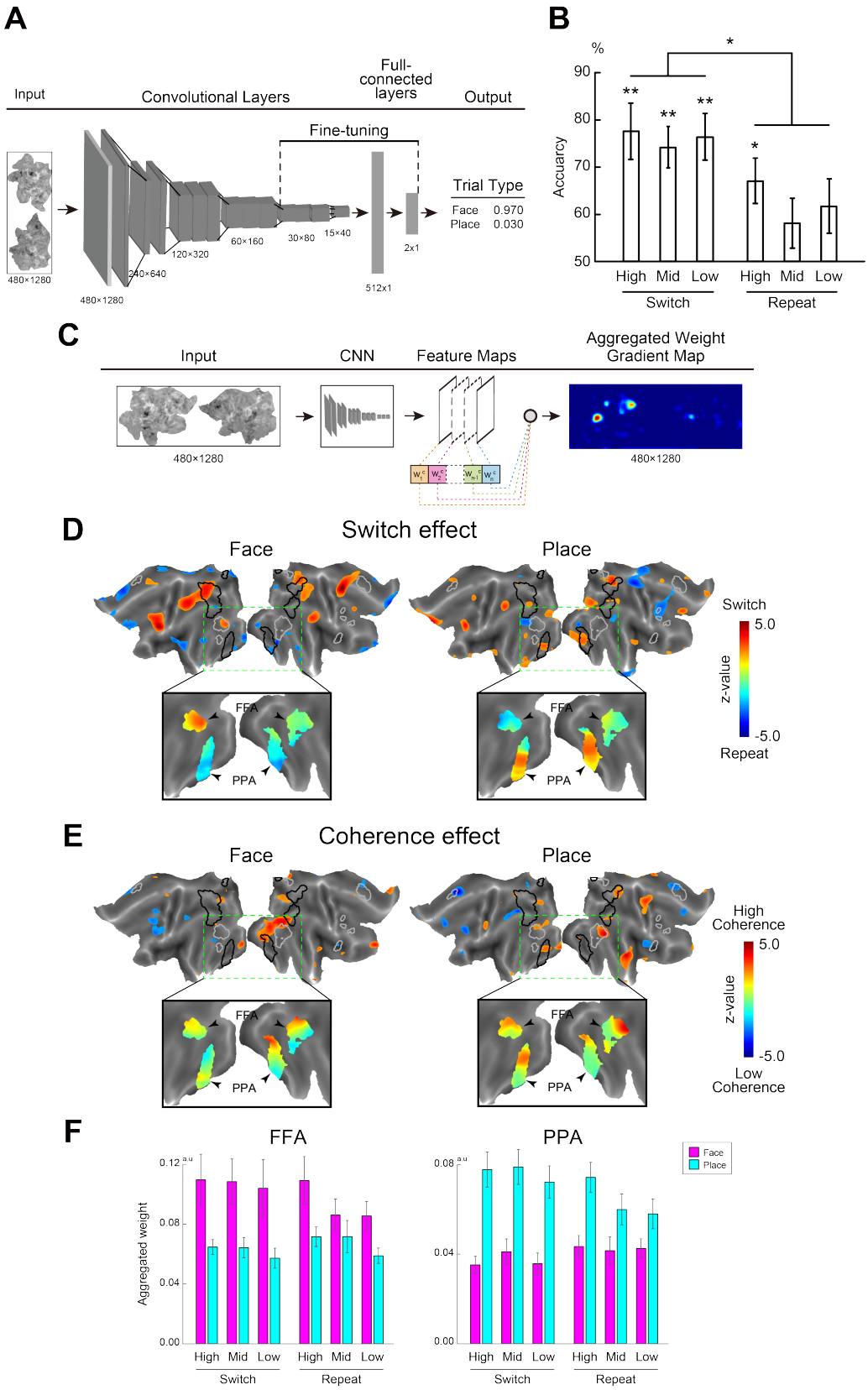


Fig 3. Convolutional neural network (CNN) classifier mapping

(A) CNN model was based on VGG16 pre-trained by ImageNet data. The model was retrained to classify performed task (face or place). Retraining of brain maps was based on fine tuning from the 5th convolution layer to the full-connected layers. (B) Classification accuracy for each task condition. *: P < .005; **: P < .001. High: high-coherence (80%) trials; Mid: middle-coherence (40%) trials; Low:

low-coherence (20%) trials. (C) Weight gradients between convolution layers was aggregated and then visualized to identify cortical regions involving more information to classify tasks. (D) Visualization of aggregated weight contrasts for switch vs. repeat trials in face task (left) and place task (right). Statistical maps are overlaid onto flat cortical anatomical images with a statistical threshold of $|z| > 2.0$ (top). Hot and cool colors indicate statistical z-values of positive (switch > repeat) and negative (repeat > switch) effect, respectively. Gray and black closed lines overlaid on flat map indicate brain regions significantly activated during face and place tasks in univariate analysis, respectively (Fig 2C and S6 Fig). Occipitotemporal regions in rectangular boxes with green broken lines were expanded below. Maps were overlaid onto flat maps masked by the univariate activation contrast face vs. place tasks for target only trials (gray and black closed lines in the top panels). The fusiform face area (FFA) and parahippocampal place area (PPA) are indicated by arrow heads. (E) Visualization of weight contrast for coherence effect in face task (left) and place task (right). Hot and cool colors indicate greater statistical values of weights in high- and low-coherence trials, respectively. The formats are similar to those in panel (D). (F) Regions of interests (ROIs) analysis. PPA and FFA ROIs were defined based on target-only trials, and weight magnitudes were collected for each cue trial conditions and ROIs. All error bars in the figure indicate standard errors of the means across participants.

Given that the CNN model successfully classified face and place tasks during the target only trials with high accuracy, we then examined the classification accuracy for cue trials (Fig. S2D). Accuracy was higher than chance level in switch trials at all coherence levels [80% switch: $t(28) = 4.7$, $P < .001$; 40% switch: $t(28) = 5.5$, $P < .001$; 20% switch: $t(28) = 5.4$, $P < .001$] and 80%-repeat trials [$t(28) = 3.6$, $P < .005$] (Fig. 3B), which ensured those maps contained information about performed tasks. More importantly, classification accuracy was higher in switch than in repeat trials [$F(1, 28) = 10.9$; $P <. 005$], suggesting that cortical activation patterns involves more information about task dimension in switch than in repeat trials, although the coherence effect was absent [$F(1, 28) = 0.3$; $P = 0.6$].

We explored brain regions involving critical information to classify face and place tasks by visualizing weights of convolution layers of the CNN model. We used Grad-CAM (46) that aggregates weights across convolution layers, and highlights image locations with greater weights when important information to classify the image are involved (Fig. 3C). The weight maps were created for each of the tested images (i.e. flat activation maps for cue trials), and then collected for each of the tasks (face/place), switching conditions (switch/repeat), and coherence levels (high/mid/low) (see Methods).

We first contrasted weight maps between switch and repeat trials, and calculated pixel-wise group-level z-statistics with participants treated as a random effect (Fig. 3D). Prefrontal, parietal and occipitotemporal areas including FFA and PPA (see Figs. 2C and S4 for references) showed greater weights in switch trials than repeat trials

(Fig. 3D). In particular, weights became greater in trials switching to face task but not to place task in the FFA (Fig. 3D *left*). On the other hand, greater weights were observed in trials switching to the place task but not to face task in the PPA (Fig. 3D *right*). These results suggest that modality-dependent FFA and PPA encode task-relevant information to a greater degree during the switch to the task that demands stimulus discrimination of optimal modality. Next we examined the coherence effect during cue trials by calculating the weighted sum of weights for each pixel (Fig. 3E). The FFA and PPA also showed greater weights in high-coherence trials than low-coherence trials in both of the face and place tasks.

In order to statistically test dissociated weight patterns in the FFA and PPA during cue trials, ROI analysis was performed (Fig. 3F). ROIs were defined based on target only trials, independently of cue trials (see Materials and Methods). In the FFA, weights were greater during the face task than the place task [$F(1, 28) = 15.2$, $P < 01$], and in the PPA, weights were greater during the place task than the face task [$F(1, 28) = 66.5$, $P < .001$]. For weight magnitudes with optimal task-region relation (i.e., face task in FFA and place task in PPA), weights became greater in switch relative to repeat trials [$F(1, 28) = 6.4$, $P < .05$], and in high relative to low-coherence trials [$F(1, 28) = 13.4$, $P < .01$ with linear contrast]; however, there was no without switching-by-coherence interaction [$F(1, 28) = 0.1$, $P = .75$]. In a separate analysis, we used a leave-one-subject-out procedure when retraining the classifier to classify the activation maps of the target only trials, and then tested the remaining subject (Fig. S5). Overall results are consistent, suggesting that subject-specific noises are not dominant in our results.

These results suggest that the FFA and PPA involve more modality specialized task-related pattern information in high-coherence trials. Thus, modality-dependent occipitotemporal regions may encode relevant task information (i.e. FFA for face task and PPA for place task), which is enhanced in switch trials with a high-coherence task cue.

<u>Whole-brain decoding by SVM.</u>

In order to complement decoding and mapping by the CNN classifier, we performed another decoding analysis using an SVM classifier. Similar to the CNN classifier analysis above, the classifier was trained based on HCP WM task such that the classifier

discriminates the dimension of the tasks (face or place) (see Materials and Methods). We then tested the experimental data to examine classification performance for face and place tasks of the cue trials.

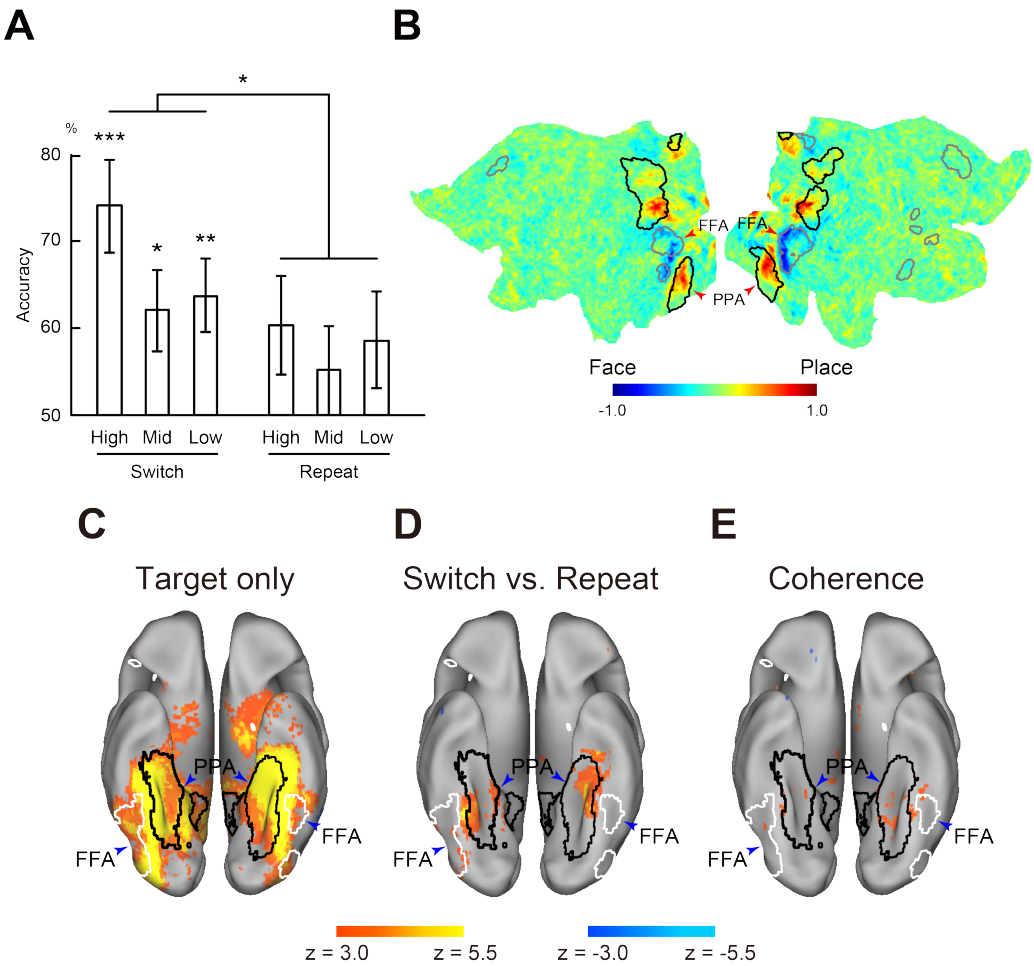


Fig 4. Multi-variate pattern analysis (MVPA) with support vector machine (SVM). (A) Classification accuracy for each task condition with SVM classification in the whole brain. The model was trained for classification of face and place tasks with the working memory task in the Human Connectome Project (HCP) and tested with each condition of cue trials in the current study. Error bars indicate standard error of the mean across participants. *: P < .05; **: P < .005; ***: P < .001. (B) Visualization of weight assigned to the pixels for classification of face and place tasks in the working memory task in the HCP. Formats are similar to those in Figs. 3D/E. (C-E) Statistical significance maps for searchlight MVPA. Classifiers were trained to classify the performed task. Maps are overlaid onto a 3D surface of the brain and displayed from a ventral view. White and black closed lines overlaid onto the 3D surface of the brain indicate significant clusters for contrast face vs. place tasks in univariate analysis, respectively (Fig. 2c). The fusiform face area (FFA) and parahippocampal place area (PPA) are indicated by blue arrow heads. (C) Target-only trial. Hot and cool colors indicate statistical level for classification accuracy relative to chance level. (D) Accuracy difference between switch and repeat trials. Hot and cool colors indicate higher accuracy in switch and repeat trials, respectively. (E) Differential classification accuracy depending on the coherence effect. Hot and cool colors indicate higher accuracy in high- and low-coherence trials, respectively.

We found that accuracy was higher than chance level in switch trials at all coherence levels [80% switch: $t(28) = 4.5$, $P < .001$; 40% switch: $t(28) = 2.5$, $P < .05$; 20% switch: $t(28) = 3.3$, $P < .005$; Fig. 4A], and higher in switch than repeat trials [$F(1, 28) = 5.0$; $P <. 05$], which is consistent with the CNN classifier results.

Weights of the SVM classifier were mapped onto 2D cortical surface of the brain in order to identify brain regions with greater weights to classify the face and place tasks. Occipitotemporal regions including the FFA and PPA showed prominent reverse directed weights (Fig. 4B), indicating that these regions involve important information to classify the two tasks. Notably, these maps are consistent with the CNN-based mapping, especially in the FFA and PPA (Figs. 3D/E). We also trained another SVM based on target-only trials in the current task, and found that the classification accuracy of cue trials (Fig. S6A) and weight maps (Fig. S6B) were consistent to those with the CNN classifier (Figs. 3B/D-E) and whole-brain SVM (Figs. 4A-B). We note that SVM weight maps (Figs. 4B and S6B) reflect a hyperplane calculated by training data (i.e. maps for HCP WM or current target-only trials), whereas CNN weight maps reflect degree of contribution to classify tested image (i.e. current cue trial maps) (see Discussion).

<u>Decoding mapping by searchlight SVM.</u>

The above CNN and SVM classifiers were based on pattern information of whole brain cortical regions. Another SVM analysis was also performed using the searchlight procedure (see Materials and Methods). By exploring across the whole brain, searchlight was used to identify brain regions where local image voxels involved pattern information about the performed task to classify face and place tasks. Again, the classifier in each searchlight was trained using HCP datasets, and thus the training and testing datasets were independent.

For target-only trials, classification accuracy was significantly higher in the FFA and PPA regions (Fig. 4C and Table S4), suggesting that modality-dependent occipitotemporal regions involve relevant task information; this result is consistent with the univariate analysis (Fig. 2C).

Voxel-wise classification accuracy maps were contrasted between switch vs. repeat trials for each participant, and group-level statistical tests were performed in

order to identify brain regions where discriminable pattern information is greater in switch trials than in repeat trials. Occipitotemporal regions showed a significant effect of switching (Fig. 4D and Table S5), indicating that, in these regions, classification accuracy is higher in switch relative to repeat trials. Interestingly, these regions were spatially located in-between the FFA and PPA, where pattern information of searchlight classifiers may modestly involve both of face-related and place-related information, possibly in a balanced manner. These regions also showed a coherence effect with higher accuracy in high-coherence trials (Fig. 4E and Table S6). Notably, these results were consistent with weight mapping of the CNN classifier (Figs. 3D-E).

It is important that these regions showed significantly higher classification accuracy than chance-level in switch and repeat trials (Figs. 7A-B and Tables S7-8), and cue trials at each coherence level (Figs. S7C-E and Tables S9-11). This assures that the differential accuracies between switch and repeat trials, and across coherence levels were attributable to accuracy enhancement in switch trials with the high-coherence cue.

These collective results suggest that occipitotemporal regions adjacent to stimulus-modality-dependent FFA/PPA areas involves information about ongoing task, and that the information amount is increased during task switching with more coherent cue presentation. These results are also consistent with those of the classification performance and mapping based on whole-brain CNN and SVM classifiers. Such differential classification accuracy was not observed in fronto-parietal regions well known to be involved in executive control (Figs. S7F-G and Tables S5-6), even when the classifier was trained by target-only trials in the current experiment (Figs. S7H-I).

Discussion

The current study examined neural mechanisms during task switching under situation where task cue involved uncertainty. Task-related neural coding in FFA/PPA became more evident during task switching and also when the relevant task was cued more explicitly. When task-cue was distinct, the lPFC received task-related signals from the MT region and PPA/FFA, and the direction of the signal was reversed when the task cue involved more ambiguity. These results suggest a distributed cortical network of fronto-occipitotemporal regions for behavioral flexibility where task-related signal among these regions help to implement task representation depending on the ambiguities of external cue (Fig. 5).

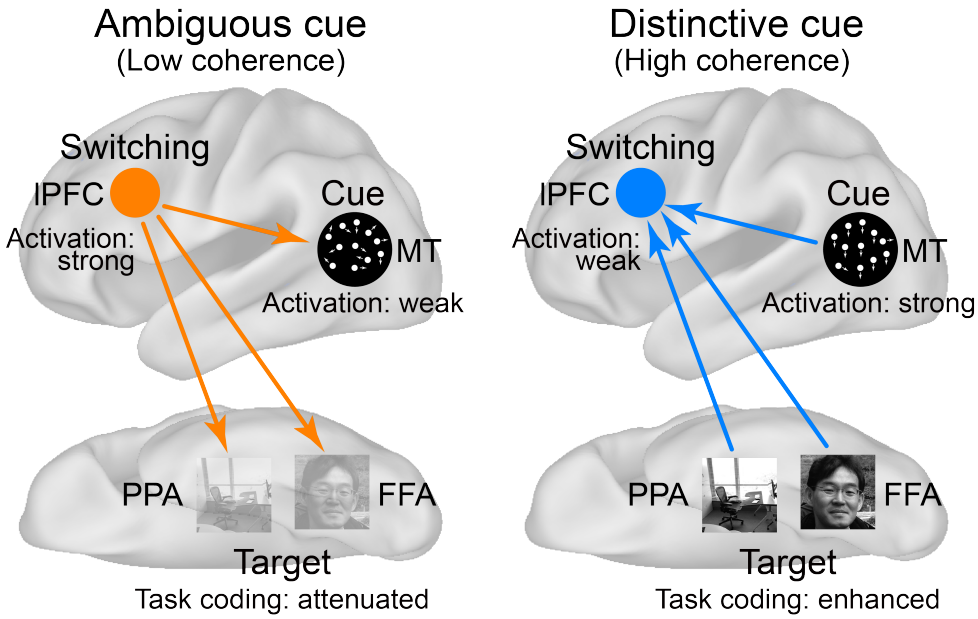


Fig 5. Putative model of behavioral flexibility under perceptual uncertainty
Schematic diagrams for functional mechanism among the middle temporal (MT) region, fusiform face area (FFA), parahippocampal place area (PPA), and lateral prefrontal cortex (lPFC) during task switching with cue uncertainty. The arrows indicate signal directions.

<u>Neural mechanism for task switching and perceptual decision making.</u>

Prior work of task switching has examined switch-related neural mechanisms under situations where perceptual uncertainties were applied to target stimulus (24, 49-51). In contrast, the current study manipulated uncertainty of the task cue that involves multi-task information (Figs. 1A-C and 5). Notably, the task cue indicates relevant task dimension at the top tier of the hierarchical task set structure, and thus the current study

allowed us to elucidate higher-level cognitive functions governing task switching and perceptual decision-making.

For putative mechanisms to achieve task switching under cue uncertainty, three hypotheses are possible: 1) a unitary mechanism implements task switching under perceptual uncertainty in a task cue; 2) distinct mechanisms for perceptual decision-making and task switching interactively guides successful task switching; and 3) a hub-like region links the two distinct mechanisms. Behaviorally, an interaction between switching and coherence levels was absent (Figs. 1D/E; see also Results); this is consistent with the univariate imaging analysis and CNN classifier mapping showing no interaction effect. The absence of an interaction effect, together with distinct brain regions associated with cue/target perception and task switching (Figs. 2A-C and 3D-F), suggests distributed mechanisms for task switching under cue ambiguities, which supports the second hypothesis. Thus, by modulating effective connectivity and interregional signaling depending on cue ambiguity, these regions cooperatively guide behavioral flexibility (Fig. 5). This interpretation is also compatible with the central role of frontal regions for flexible task control (52).

<u>Comparisons of classification and mapping among machine learning techniques.</u>
In the current study, whole brain exploration of task-related neural representation was performed by three approaches based on three machine learning techniques, 1) CNN classifier, 2) whole-brain cortical SVM, and 3) searchlight SVM.

CNN classified activation maps along task dimensions based on all pixels across whole cortical regions, and the classification accuracy was higher in switch trials than repeat trials. One novel signature of the current CNN classifier approach is that brain regions involving critical information to classify cue trials were mapped by aggregated weight gradients across convolution layers based on the Grad-CAM technique. It is notable that this CNN weight mapping is available on image-by-image basis for testing data, which is not the case for the SVM mapping using whole cortical images. Then, the CNN weight mapping revealed that task representation in the FFA during the face task and in the PPA during the place task was enhanced in switch and more coherent trials than in repeat and less coherent trials. Increased task-related activation (Figs. 2C and S4) may be associated with higher classification accuracy and enhanced task representation.

Standard voxel-wise univariate GLM analysis identifies brain regions where the MRI signal is differentiated between task conditions, but does not necessarily indicate that identified brain regions are critical for task performance; this makes it hard to identify brain regions playing an important role in cognitive functions (i.e., reverse inference). In contrast, the current CNN classifier demonstrated that the visualization of convolution layers of the classifier for brain activation is useful in identifying brain regions that characterize task performance.

The analysis based on the CNN classifier was complemented by a standard SVM analysis for whole brain cortical maps that tested identical map images. The classification accuracy of cue trials was highly consistent between the two classifiers. Additionally, SVM weight maps also showed differential weights in the FFA and PPA, which is also consistent with the CNN classifier. One notable technical limitation of the SVM is that weight mapping for testing classification for cue trials was unavailable, unlike Grad-CAM of CNN classifier. Thus the SVM weight map indicates that the FFA and PPA is critical to classify tasks for the HCP N-back working memory task or target only trials in the current task (i.e. training data), but not necessarily for cue trials of the current task (i.e. testing data). Nonetheless, together with differential classification accuracy among cue conditions, the SVM weight maps suggest that pattern information in the FFA and PPA is distinct (i.e. distant from separating hyperplane) in switch trials than in repeat trials.

Another approach to identify brain regions that characterize task performance is the searchlight SVM, which also allows whole-brain exploration of activation patterns, but individual classifications were restricted in local brain regions (Figs 4C-E and S7A-E). This is in contrast to the CNN classifier and whole-brain cortical SVM that are trained and classify based on a whole brain image. Nonetheless, results were complementary to those whole-brain-based classifiers in that 1) occipitotemporal regions adjacent to the FFA and PPA were capable of task classification, and 2) the classification performance in these regions became higher in switch and more coherent trials.

<u>Classifier training using independent open resource data.</u>

One notable analysis procedure in the current machine-learning-based functional brain mapping is that classifier was trained using an open resource dataset that was

independently collected from the current experiment. This procedure ensured independence between the training and testing data.

Task switch and working memory may involve distinct cognitive control demands, with the former related to behavioral flexibility (3, 4) and the latter related to active maintenance and updates of goal-relevant information (53). However, the two tasks used common visual stimulus categories (face and place); thus, perceptual demands may involve some degree of commonality. Recognition demands for the presented stimuli were also distinct: the current task-switching paradigm used male-female and indoor-outdoor discriminations during face and place tasks, respectively, but HCP working memory tasks used discrimination of identicalness to past stimuli. Additionally, the current task used face-place superimposed stimuli, and thus the identical stimulus set was used during face and place tasks, whereas HCP working memory task used distinct visual stimulus sets during face and place blocks. Thus task representation examined in the current study may reflect visual perception or attention rather than low level visual features.

For the CNN classifier, training involved two steps: retraining of HCP working memory maps, and additional retraining of target-only trials. CNN classifies maps based on whole cortical areas including fronto-parietal regions in which differential sub-regions are recruited during task switching (5-15, 31, 33) and working memory (1, 53, 54). Thus additional retraining of the CNN model based on identical recognition demands (i.e. target-only trials) was effective in classifying tasks to optimize the CNN model for classification using whole cortical images. Distinct weight differences in the parietal cortex (Figs. 3D-E) may partially be attributable to higher performance with the additional retraining.

Because incremental training of HCP working memory trials and the target only trials in the current study is irrelevant to SVM, these two datasets were separately trained for whole brain cortical SVM, and weight maps were consistent especially in occipitotemporal regions (Figs. 4B and S6B). Importantly, classification accuracy for the cue conditions were consistent in SVMs with the two training datasets, and also with the CNN classifier. The sample size was much smaller for the current target-only trials than HCP dataset, but classification performance was comparable between those two classifiers (Figs. 4A and S6A). Thus, SVM may thus not require a larger sample

size like the HCP data for training, while CNN training needed incremental training even with large sets of image data.

The searchlight SVM using HCP working memory maps as training data identified occipitotemporal regions spatially closed to the FFA/PPA showing higher classification accuracy for target-only trials. Moreover, classification accuracy was higher in high-coherence switch trials. Interestingly, these classification results were absent in fronto-parietal regions, well known to be involved in executive control (Figs. S7F/G), even when the searchlight classifier was trained by the target-only trials of the current task (Figs. S7H/I). One possibility for this discrepancy is that control and recognition demands are incompatible while perceptual modality is compatible in HCP working memory trials, target only trials, and switch/repeat trials. Then the distinct control and recognition demands might be reflected in classification incompatibility in the fronto-parietal regions.

Fronto-occipitotemporal network mechanisms.

DCM analyses revealed top-down signal from the lPFC to the FFA or PPA, depending on the task to be switched, and this top-down signaling is enhanced when task cue involved ambiguity during switching (Fig. 5). The top-down mechanisms may reflect supplemental attention to visual stimulus required to collect task cue information about the tasks to be performed. The supplemental attention involving frontal engagement may complement stimulus-modality-dependent activation in occipitotemporal regions (55-58).

In contrast, bottom-up signaling with increased cue information may reflect the conversion of sensory information to behavioral information through an information stream from the visual sensory area to executive control areas (55, 56). Thus, when the task cue was apparent, the bottom-up signal was strengthened because cue-related information is more available, which may help to enhance task switching performance.

Materials and Methods

Participants.

Written informed consent was obtained from 30 healthy right-handed subjects (age range: 18-22; 11 females). Experimental procedures were approved by the institutional review board of Keio University and Kochi University of Technology. Participants received 2000 yen for each of the training and scanning sessions. One participant was excluded from analyses due to low behavioral performance; accuracy was lower than 30% in one of the experimental conditions.

Behavioral procedures.

The experiment consisted of two sessions administered on separate days. The first day was a training session, in which participants practiced discrimination tasks (random dot motion) and switching between two tasks (face and place tasks; see below for behavioral procedures). On the second day, while fMRI scanning was administered, the participants performed the switching paradigm identical to those practiced in the training sessions.

Stimuli.

All stimuli were generated in Matlab version 2012a, using the Psychophysics Toolbox (59) extension version 3.0.10, and were visually presented on a computer screen. The current task cue stimuli were randomly moving dot stimuli similar to those used in a previous study of perceptual decision-making (55). Each motion stimulus involved 60 dots moving inside a donut-shaped display patch with a white cross in the center of the patch on a black background (Fig. 1B). The display patch and cross were centered on the screen and extended from 6 to 12° degrees of visual angle (dva). Within the display patch, every dot moved at the speed of 10 dva per second. Some dots moved coherently toward one direction (upward or downward) while the others moved randomly. The percentage of coherently moving dots determined the "motion coherence", which was set to three levels (20, 40, and 80%).

Dot presentation was controlled to remove local motion signals following a standard method for generating motion stimuli (18, 60-62). Namely, upon stimulus onset, the dots were presented at new random locations on each of the first three frames. They were relocated after two subsequent frames, such that the dots in frame 1 were

606 repositioned in frame 4, and the dots in frame 2 were repositioned in frame 5, and so on.
607 When repositioned, each dot was either randomly presented at the new location or
608 aligned with the pre-determined motion direction, depending on the pre-determined
609 motion strength on that trial. Each stimulus was composed of 18 video frames with a 60
610 Hz refresh rates (i.e. 300-msec presentation).
611 Within the center circle mask of the donut-shape motion stimulus, a
612 face/place superimposed stimulus was presented simultaneously (Fig. 1B). The face
613 image set consisted of an image of picture of 4 male and 4 female unfamiliar Japanese
614 faces, and the place image set consisted of 4 indoor and 4 outdoor unfamiliar places;
615 this resulted in 64 overlaid images.
616
617 Task procedure.
618 At the beginning of the task, a dot patch and face/place stimulus were simultaneously
619 presented. The direction of the dot patch (up or down) indicated the task to be
620 performed (discrimination of face or place). Depending on the motion direction,
621 participants were required to judge whether the face was male or female, or whether the
622 place was indoor or outdoor (Fig. 1B), and pressed the corresponding button with their
623 right thumb. Both of accuracy and speed were stressed. Stimulus was presented for 1.8
624 sec, followed by a 0.7 sec intertrial interval. If participants made an incorrect response,
625 or did not respond within 1.8 sec from the stimulus onset, feedback stimulus indicating
626 an error (X) was presented for 1.0 sec, followed by high-coherence (80%) cue trials for
627 the same task dimension. The cue trials immediately after the error were discarded from
628 analyses. Stimulus-response and cue-task associations for the two tasks were identical
629 on days 1 and 2, and counterbalanced across participants.
630 The trial with simultaneous cue/target presentation was followed by trials
631 with presentation of the face/place target stimulus without the dots cue stimulus (target
632 only trials). The target-only trials were repeated for 3-5 times (Fig. 1C). In the
633 target-only trials, participants were required to discriminate the center image stimulus
634 along the same dimension until the next task cue (moving dots) was presented. One task
635 block lasted for approximately for 90 secs, and 20-sec fixation blocks were inserted
636 between task blocks. Each functional run involved 3 task blocks and lasted for 305 secs.
637 The first trial at the beginning of each task block presented the dot cue with highest
638 coherence (80%), and was discarded from analysis.

Practice procedure.

On the first day, participants practiced the tasks outside of the scanner. They first practiced a discrimination task for the moving dot stimulus, and were required to judge the direction of overall motion (up or down) and to press the correct corresponding button as quickly as possible. The response window was 1050 msec. Each practice run involved 70 trials, and 5 runs were administered for each participant. The first 5 trials and last 5 trials in each run used the highest coherence level (80%). Thus the middle 60 trials were composed of 20 trials for each of coherence levels (20, 40, or 80%).

The participants then practiced discrimination tasks for the face and place stimulus (see above). Across task switching practice runs, the switching frequency and coherence levels of the moving dot cue were manipulated such that the cue trials gradually became more difficult (i.e. more switch trials with low-coherence cue). Participants performed 8 practice runs for task switching.

Behavioral procedure in scanning session.

On the second day, after practicing task switching for one run, the participants performed 9 runs of task switching with identical procedure as for day 1 (see above) while functional MRI was administered. The frequency of switch and repeat trials and coherence level were approximately equivalent across runs.

Imaging procedure.

MRI scanning was administered by a 3T MRI scanner (Siemens Verio, Germany) with a 32ch head coil. Functional images were acquired using a multi-band acceleration echo-planar imaging sequence (63) [repetition time (TR): 0.8 sec; echo time (TE): 30 msec; Flip angle (FA): 45 deg; 80 slices; slice thickness: 2 mm; in-plane resolution: 3 x 3 mm; multiband factor: 8]. Each functional run involved 385 volume acquisitions. The first 10 volumes were discarded from analysis to take into account the equilibrium of the longitudinal magnetization. High-resolution anatomical images were acquired using an MP-RAGE T1-weighted sequence [TR: 2500 msec; TE = 4.32 msec; FA: 8 deg; 192 slices; slice thickness: 1 mm; in-plane resolution: 0.9 x 0.9 mm^2].

Behavioral analysis.

672 Trials were classified into three types in terms of switching: 1) task switch trials
673 presenting the random dot cue and target stimuli simultaneously, where the task to be
674 performed alternated (i.e., face to place or place to face); 2) task repeat trials presenting
675 the random dot cue and target stimuli simultaneously, where the same task was
676 repeated; and 3) target-only trials presented after the switch and repeat trials, and their
677 subsequent trials without random dot cue stimuli presentation (Fig. 1C). These trial
678 types were analyzed separately. Trials were also classified by task dimension (face or
679 place), and switch and repeat trials were examined at each coherence level (20, 40, or
680 80%). Accuracy and reaction times (RTs) were calculated for each trial condition, and
681 then compared. Statistical testing was performed based on repeated measures ANOVAs
682 implemented in SPSS Statistics 24 (IBM Corporation, NY USA).
683
684 Image preprocessing.
685 MRI data were analyzed using SPM12 software (http://fil.ion.ac.uk/spm/). All
686 functional images were initially temporally realigned across volumes and runs, and the
687 anatomical image was coregistered to a mean image of the functional images. The
688 functional images were subsequently spatially normalized to a standard MNI template
689 with normalization parameters estimated based on the anatomical scans. The images
690 were resampled into 2-mm isotropic voxels, and spatially smoothed with a 6-mm
691 full-width at half-maximum (FWHM) Gaussian kernel.
692
693 Imaging analysis: general linear model (GLM).
694 Single level analysis.
695 A GLM approach (64) was used to estimate parameter values for task events. The
696 events of interest were correct switch, repeat, and target-only trials. For switch and
697 repeat trials, the normalized (z-scored) coherence level of the dot stimuli was also added
698 as a parametrical effect of interest. Error trials in all conditions were separately coded in
699 GLM as nuisance effects. Those task events were time-locked to the onset of target
700 images and then convolved with canonical hemodynamic response function (HRF)
701 implemented in SPM. Additionally, six-axis head movement parameters, white-matter
702 signals, lateral-ventricle signals, and parametrical effect of reaction times normalized
703 across trials were also included in GLM as nuisance effects. The parameters were then
704 estimated for each voxel across the whole brain.

705

706 Group-level analysis.

707 Maps of parameter estimates were first contrasted within individual participants.
708 Contrast maps were collected from all participants, and subjected to a group-level
709 paired t-test. For the coherence effect, the contrast maps were subjected to a one-sample
710 group-mean test, with maps weighted and summed based on normalized coherence
711 levels. Voxel clusters were identified using an uncorrected threshold of $P < .001$ based
712 on voxel-wise t-statistics. The voxel clusters were tested for a significance with a
713 threshold of $P < .05$ corrected by family-wise error (FWE) rate based on permutation
714 methods (65) (5000 permutations) implemented in *randomise* in FSL suite
715 (http://fmrib.ox.ac.uk/fsl/). This group analysis procedure was validated to appropriately
716 control false positive rates in a prior study (66). Peaks of significant clusters were then
717 identified and listed on tables. If multiple peaks were identified within 12 mm in one
718 cluster, the most significant peak was retained. When exploring brain regions associated
719 with motion coherence, exploration was restricted within a mask obtained from
720 Neurosynth (67) (http://neurosynth.org/) for the search word 'motion' ($z > 3.0$, for
721 uniformity test), in order to ensure the extraction of motion-related regions, because the
722 current cue trials simultaneously presented face/place stimuli in addition to cues
723 indicating task switching tasks between face and place discrimination.

724

725 Effective connectivity analysis.

726 The current analysis was designed to test the hypothesis that functional connectivity
727 among brain regions associated with task switching, motion perception, face perception,
728 and place perception identified in univariate analysis (Figs. 2A-C) is modulated by task
729 manipulations and brain signals. Dynamic causal modeling (68) (DCM) analysis was
730 performed in order to examine functional connectivity mechanisms associated with task
731 switching under cue uncertainty (Figs. 2D-E). DCM allows us to explore the effective
732 connectivity among brain regions under the premise that the brain is a deterministic
733 dynamic system that is subject to environmental inputs and produces outputs based on
734 the space-state model. The model constructs a nonlinear system involving intrinsic
735 connectivity, task-induced connectivity, and extrinsic inputs. Parameters of the
736 nonlinear system are estimated based on fMRI signals (system states) and task events.

Four regions of interest were first defined based on univariate analysis and prior studies of task switching and perceptual decision-making: 1) task switching [left lateral prefrontal cortex (lPFC) (5-11, 33, 69); Fig. 2A]; 2) motion perception [middle temporal (MT) (16-18, 21-24); Fig. 2B]; 3) face perception [fusiform face area (FFA) (23-27); Fig. 2C]; 4) place perception [parahippocampal place area (PPA) (27, 28, 30); Fig. 2C]. More specifically, meta-analysis maps were obtained from Neurosynth[64] (http://neurosynth.org/) using a keyword search for "switching", "mt", "ffa", "place" to obtain the meta-analysis maps for lPFC, MT, FFA, and PPA regions of interest (ROIs), respectively. ROI images were then created with 6-mm radius spheres centered in the peak coordinates in the meta-analysis activation maps thresholded above $z > 10$ (uniformity test).

Given these ROIs, we first tested whether the switch-related prefrontal region sends (or receives) a task-related signal toward (or from) the stimulus-modality dependent occipitotemporal region of the target (i.e. FFA/PPA) during task switching (Fig. 2D). If this were the case, then we tested whether signaling between prefrontal and occipitotemporal regions changed depending on the uncertainty of cue stimuli (Fig. 2E).

Signal time courses of four ROIs and regressors in events of interest were extracted from first-level GLMs. The events of interest were cue (switch and repeat) trials and target-only trials of each task. For switch and repeat trials, the contrast of the two trials and normalized coherence level of the dot stimuli were added as a parametrical effects of interest. Nuisance effects of head-motion, white matter signal, ventricle signal, functional run, and contrast were subtracted out from the ROI timecourses. The input matrix was U mean-centered.

For each trial effect, causal models were defined as those that differed in external inputs and modulatory effects among ROIs. Because the current analysis involved 2 or 3 ROIs (Figs. 2D-E), the tested models included 16 or 512 types (i.e. 2^2 inputs and 2^2 connection effects or 2^3 inputs and 2^6 connection effects). Connectivity matrices reflecting 1) first-order connectivity, 2) effective change in coupling induced by the inputs, and 3) extrinsic of inputs on MRI signal in ROIs were estimated for each of 1the 6 or 512 models based on DCM analysis implemented in SPM12. Parametric regressor (switch vs repeat / coherence) was used as an extrinsic effect for effective connectivity between ROIs and ROIs inputs.

In order to estimate strength of effective connectivity, a Bayesian model reduction method (70) was used. The reduction method allows the calculation of posterior densities for all possible reduced models, which were then inverted to a fully connected model. The reduced models were then supplemented by second-level parametric empirical Bayes (70) (PEB) to apply empirical priors that remove subjects' variability from each model.

Next, the parameters of these models were estimated based on Bayesian model averaging (68) (BMA) to estimate group-level statistics. Because the current analysis aimed to identify effective connectivity observed as an average across participants, we used a fixed effect (FFX) estimation assuming that every participant uses the same model. This is in contrast to using a random effect (RFX) estimation assuming different participants use different models, which is often used to test group differences in effective connectivity (71). The significance of connectivity was then tested by thresholding at a posterior probability at the 95% confidence interval. We used the uncorrected threshold, because the current analysis aimed to test if connectivity between two specific brain regions was enhanced depending on task manipulation and brain activity, not to explore one model involving connectivity among multiple brain regions that best fits to the imaging and behavioral data.

Additionally, in order to test the robustness of the functional connectivity, we performed supplemental analyses. We estimated model parameters 1) without an empirical prior (Figs. S1C-D); 2) changing the number of the regions of interest (ROIs) in the models (Figs. S1E-F), and 3) changing the definition of the ROIs (Figs. S1G-H). When changing the ROI definition, we used a leave-one-out procedure; the centers of ROIs of one participant were determined based on group-level univariate activation maps of corresponding contrasts (i.e. lPFC: switch vs. repeat trials; MT: high vs. low coherence trials; FFA/PPA: face vs. place of target-only trial) without the participant in order to circumvent circular analysis. The ROIs were created as spheres with 6-mm radius for individual participants. Complete results can be provided upon request.

Convolutional neural network (CNN) classifier.

In order to explore brain regions involving task-related neural representation, a convolutional neural network classifier (CNN) (44, 45) was used. The current CNN model was based on VGG16 (47), with five convolution layers for extracting image features and two fully connected layers for binary classification. Initial parameters of convolution layers were set to parameters pre-trained with concrete object images provided from ImageNet (72) (Fig. S2A; http://www.image-net.org/).

The VGG16/ImageNet model is capable of classifying concrete object images into 1,000 item categories. Importantly, it has been demonstrated that the pre-trained model can learn novel image sets more efficiently than the non-trained model by tuning convolution and fully connected layers and fully connected layers only (48, 73) (Figs. 3A and S2C). Thus, the current analysis retrained the pre-trained VGG16-ImageNet model to classify brain activation maps.

Training data were single-subject 2nd level z-maps during the N-back working memory task from the S1200 release of the Human Connectome Project (74, 75) (N = 992; HCP; http://www.humanconnectomeproject.org/). From each participant, statistical z-maps for activation contrasts for face vs. fixation and place vs. fixation (2-back and 0-back corrupted) were collected. The pre-trained VGG16 model was retrained by the activation maps during the N-back working memory task such that the model classifies face and place trials. We used gray-scaled flat 2D cortical maps (76) provided from HCP (992 images; face: 496, place: 496; Fig. S2B) for dimensional compatibility of images between VGG16-ImageNet and activation maps. The training data set was divided into 10 subsets, and 9 subsets were used for retraining and the remaining 1 set was used for validation, enabling a 10-fold cross-validation test. Model training and testing was implemented using Keras (https://keras.io/) under Tensorflow backend (https://www.tensorflow.org/) [input image size: 480 x 1280 pixels; batch size: 10; epoch: 50; learning rate: 0.0001; optimizer: Stochastic Gradient Descent (SGD); Figs. 3A and S2C *top*].

After retraining of the HCP working memory maps, the model with highest classification accuracy was further retrained to classify activity maps for face and place tasks during target-only trials of the current dataset (Fig. S2C *bottom*). For each functional run of each participant, a single-level GLM estimation was performed with regressors identical to those in the univariate analysis as described above. The GLM

831 estimations were performed within standard MNI space. Activation z-maps for the
832 contrasts for face vs. fixation and place vs. fixation during correct target-only trials were
833 collected from each functional run. Activity maps for the contrast for face vs. fixation
834 and place vs. fixation were then gray-scaled and flattened such that these maps were
835 anatomically and geometrically identical to those from the HCP working memory task
836 using Connectome Workbench
837 (https://www.humanconnectome.org/software/connectome-workbench/). The training
838 dataset consisting of 522 images (261 face and 261 place maps from each of 9 runs of
839 29 participants) was divided into 10 subsets, enabling a 10-fold cross-validation test.
840 Given the limited number of images available from the current experiment,
841 this two-step retraining of the model was found to be effective when classifying current
842 tasks, because 1) training randomly initialized models failed in classifying the current
843 target-only trials (Fig. S3A) and in classifying the HCP working memory conditions
844 (Fig. S3B); 2) retraining VGG16-ImageNet was successful in classifying the HCP
845 working memory conditions (Fig. S3B); 3) retraining VGG16-ImageNet model failed in
846 classifying the current target-only trials (Fig. S3A); and 4) retraining
847 VGG16-ImageNet/HCP was successful when classifying the current face/place tasks
848 (Fig. S3A).
849 After learning of the target-only trials from the current dataset, the retrained
850 10 models were tested to classify activation maps during task switching and repeat trials
851 where the dot cue stimulus was presented to indicate the task to be performed. Testing
852 data was created based on a GLM analysis where switch and repeat trials at differential
853 coherence levels were coded separately. For each functional run of each participant, a
854 single-level GLM estimation was performed, and activation contrast z-maps for face vs.
855 fixation and place vs. fixation were collected during those correct switch and repeat
856 trials. Gray-scale 2D activation maps were created for the contrasts for face vs. fixation
857 and place vs. fixation during 6 types of cue presentation trials (switch/repeat x
858 high/middle/low coherence). Importantly, the testing data were independent of the 2 sets
859 of retraining data (HCP working memory and current target-only trials). The maps were
860 tested, and accuracy was averaged across cross-validation models within each
861 participant. A statistical test of classification accuracy was performed based on repeated
862 measures ANOVAs implemented in SPSS Statistics 24 (IBM Corporation, NY USA).

In a separate supplemental analysis, in order to examine whether the current results were biased by subject-specific characteristics of image data, we used a leave-one-subject-out procedure to retrain the CNN classifier to classify activation maps of the target only trials, and then tested the remaining subject (Fig. S5).

Visualization of convolution layer weights.

In order to identify brain regions involving discriminative information to classify performed tasks, Grad-CAM (46) was used. Grad-CAM visualizes the aggregation of weight gradients between convolution layers, and highlights image locations critical for classification on a pixel-by-pixel basis (Fig. 3C).

Aggregated weights were visualized onto 2D brain surface maps for each tested map. For each participant, the weight maps were averaged across models and maps within each of the 6 cue trial conditions (switch/repeat x high/middle/low coherence cue). The averaged maps were contrasted between switch and repeat trials to explore brain regions showing differential weights gradients between switch and repeat trials within participants. For coherence the effect, three maps for coherence level trials were weighted and summed based on behavioral accuracy estimated by sigmoid fitting within participants. For each contrast, maps were collected from all participants, and pixel-wise z-values were calculated treating participants as a random effect, and the z-values were mapped on the 2D surfaces of the cortical areas.

In order to statistically test dissociable weight patterns during cue trials in the FFA and PPA, ROI analyses were performed (Fig. 3F). ROIs were defined as occipitotemporal regions showing greater activation in the contrast of face minus place tasks or place minus face tasks (Figs. 2C and S4; Table S3), independently of the cue trials. From each ROI, weight magnitudes were collected for each of the trial conditions: task (face/place), switching (switch/repeat), and cue coherence [high/middle/low: 80/40/20%], and then averaged within ROIs for each participant. Statistical tests were then performed based on repeated measures ANOVA.

Support vector machine (SVM) analysis for whole brain cortical regions.

In order to supplement the CNN classifier analysis, multi-variate pattern analyses (MVPA) based on SVM were performed. The SVM classifier was trained to perform bivariate classification for face and place tasks. Training and testing were implemented

using scikit-learn package (https://scikit-learn.org/stable/) with a Tensorflow backend (https://www.tensorflow.org/). We used a linear kernel, and adjusted C parameters (C = 0.1, 1.0, 10.0). As the overall results were maintained with the C adjustment, then we reported the results with the default parameter (C = 1.0).

For classifier training, we used the image set of single-subject 2nd level z-maps during the N-back working memory task (face vs. fixation and place vs. fixation) obtained from the S1200 HCP (N = 992), which was identical to those used in the CNN classifier training. The classifier was trained by the activation images, such that it classified face and place tasks during the HCP N-back working memory task. Weights of the trained classifier were mapped on the 2D cortical surface.

The testing dataset was also identical to those used in the CNN classifier analysis: 2-D z-maps for activation contrast of switch and repeat trials (switch vs. fixation and repeat vs. fixation) at each coherence level (20, 40, 80%) of the current experiment (Fig. S2B). Testing activation images were subject to the trained classifier for each cue trial condition of each participant, and classification accuracy was averaged for each trial condition across participants (Fig. 4A). In a separate analysis, the SVM classifier was trained based on single level z-maps during target only trials from the current dataset (N = 29), and the identical image set was tested (Fig. S6).

Searchlight SVM.

In order to explore brain regions in which local activity patterns involves information about a performed task (face/place), searchlight MVPA (77) was conducted. Bivariate classification based on SVM was used to decode the performed task (face or place). A searchlight procedure with a 5-voxel radius was used to provide a measure of decoding accuracy in the neighborhood of each voxel. Training and testing were performed based on the Decoding Toolbox (TDT; version 3.95; https://sites.google.com/site/tdtdecodingtoolbox/). Again, training and testing data were independent, based on different behavioral tasks and data sets (training: HCP working memory; testing: current task switching with male/female or indoor/outdoor judgements).

Training data were 3D single subject 2nd level z-maps during N-back working memory task (N = 1,000) from HCP S1200 release (74, 75). Similar to the CNN analysis, we used activation contrasts for face task vs. fixation and for place task vs.

929 fixation (2-back and 0-back corrupted). These activation contrasts were collected from
930 each participant, and the whole dataset was divided into 10 subsets (N = 100 each). For
931 each subset of the training data, a classifier in each searchlight was trained based on
932 these z-maps such that it classified face and place conditions during the HCP N-back
933 working memory task.

934 Test data were single-subject z-maps during switch, repeat and target only
935 trials of the current experiment. For each functional run of each participant, single-level
936 GLM estimation was performed with regressors identical to those in the univariate
937 analysis as described above. Activation maps for face vs. fixation and place vs. fixation
938 during switch and repeat correct trials were collected from each functional run. These
939 GLM analyses were performed within standard MNI space.

940 Another set of testing data was created based on a separate GLM analysis
941 with correct cue (switch and repeat) trials separately coded at each coherence level
942 (20/40/80%). For each functional run of each participant, a single-level GLM estimation
943 was performed, and activation maps for face vs. fixation and place vs. fixation were
944 collected for each coherence level.

945 For each training subset, the classifier was tested on whether it correctly
946 classified the performed task (face or place task) for each trial condition (i.e.
947 switch/repeat/target only and coherence level). Classification performance was then
948 collected from all functional runs and averaged within participants for each searchlight.
949 The performance of classification was calculated as the accuracy minus chance level for
950 bivariate classification. Accuracy maps were then averaged across testing data sets
951 within participants.

952 Accuracy maps for switch, repeat and target only trials were first averaged
953 across training subset models within participants, and averaged accuracy maps were
954 collected from all participants. Voxel-wise one-sample group-mean test was performed
955 for each trial condition, with a procedure similar to that in the univariate analysis as
956 stated above. In order to explore brain regions showing differential classification
957 accuracy between switch and repeat trials, voxel-wise group-level paired test was
958 performed, and significance was tested similarly.

959 Accuracy maps for each coherence level trials were also averaged across
960 training subset models within participants, and the averaged accuracy maps were
961 collected from all participants. Voxel-wise one-sample group-mean test was performed

for each coherence level, and significance was tested similarly. In order to examine coherence effect, the maps for three coherence levels were weighted and summed based on behavioral accuracy estimated by sigmoid fitting. Voxel-wise one-sample group-mean test was performed, and significance was tested similarly.

In separate analyses, all group-level tests above were also performed for each of the 10 training subset models, and we confirmed that overall results were consistent.

Acknowledgements

We thank Drs. Akira Funahashi, Shori Nishimoto, and Teppei Matsui for scientific comments on the study and manuscript. We thank Ms. Maoko Yamanaka for administrative assistance.

Funding

This study was supported by Kakenhi (Japan Society for the Promotion of Science) 19H04914, 17K01989, 17H05957, 17H00891, 26350986, and 26120711 to KJ; Kakenhi 18H04953 and 18H05140 to MT; 18H05017 to JC; 17H00891 to KN. This study was also supported by a grant from Uehara Memorial Foundation, a grant from Takeda Science Foundation, Keio Gijuku Academic Development Funds, and grant from Keio Leading-edge Laboratory of Science and Technology to KJ.

Author contributions

K.T. and K.J. designed the experiment and study. K.T., R.A., K.N. and K.J. collected the data. K.T., K.K. and K.J. analyzed the data. M.T. and J.C. contributed to the analysis design of imaging data and to the development of machine learning classifier. K.T., M.T., J.C., K.N. and K.J. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Data and code availability

The datasets and codes supporting the current study are available from the corresponding author (Koji Jimura, jimura@bio.keio.ac.jp) on request.

References

1. E.K. Miller, J.D. Cohen, An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167-202 (2001).
2. G. Stoet, L.H. Snyder, Neural correlates of executive control functions in the monkey. *Trends Cogn. Sci.* **13**, 228-234 (2009).
3. D. A. Allport, E.A. Styles, S. Hsieh, *Attention and performance XV*: Conscious and nonconscious information processing (The MIT Press, Cambridge, 1994).
4. R.D. Rogers, S. Monsell, Cost of a predictable switch between simple cognitive tasks. *J. Exp. Psychol. Gen.* **124**, 207–231 (1995).
5. A. Dove, S. Pollmann, T. Schubert, C.J. Wiggins, D.Y. von Cramon, Prefrontal cortex activation in task switching: an event-related fMRI study. *Cognitive Brain Res.* **9**, 103-109 (2000).
6. M.F. Rushworth, R.E. Passingham, A.C. Nobre, Components of switching intentional set. *J. Cogn. Neurosci.* **14**, 1139–1150 (2002).
7. S.A. Bunge, J.D. Wallis, A. Parker, M. Brass, E.A. Crone, E. Hoshi, K. Sakai, Neural circuitry underlying rule use in human and nonnuman primates. *J. Neurosci.* **25**, 10347-10350 (2005).
8. E.A. Crone, C. Wendelken, S.E. Donohue, S.A. Bunge, Neural evidence for dissociable components of task-switching. *Cereb. Cortex* **16**, 475-486 (2006).
9. J. Derrfuss, M. Brass, J. Neumann, D.Y. von Cramon, Involvement of the inferior frontal junction in cognitive control: Meta-analyses of switching and Stroop studies. *Hum. Brain Mapp.* **25**, 22-34 (2005).
10. N. Yeung, L.E. Nystrom, J.A. Aronson, J.D. Cohen, Between-task competition and cognitive control in task switching. *J. Neurosci.* **26**, 1429–1438 (2006).
11. C. Kim, S.E. Cilles, N.F. Johnson, B.T. Gold, Domain general and domain preferential brain regions associated with different types of task switching: a meta-analysis. *Hum. Brain Mapp.* **33**, 130–142 (2012).
12. H. Malagon-Vina, S. Ciocchi, J. Passecker, G. Dorffner, T. Klausberger, Fluid network dynamics in the prefrontal cortex during multiple strategy switching. *Nat. Commun.* **9**, 309 (2018).
13. G. B. Bissonette, M. R. Roesch, Neurophysiology of rule switching in the corticostriatal circuit. *Neuroscience* **345**, 64-76 (2017).
14. E.F. Fouragnan, B.K.H. Chau, D. Folloni, N. Kolling, L. Verhagen, M.

Klein-Flügge, L. Tankelevitch, G.K. Papageorgiou, J-.F. Aubry, J. Sallet, M.F.S. Rushworth. The macaque anterior cingulate cortex translates counterfactual choice value into actual behavioral change. *Nat. Neurosci.* **22**, 797-808 (2019).

15. D.E. Nee, M., D'Esposito. The hierarchical organization of the lateral prefrontal cortex. *eLife* **5**, e12112 (2016).
16. J.I. Gold, M.N. Shadlen, The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535-574 (2007).
17. T.D. Hanks, C. Summerfield, Perceptual decision making in rodents, monkeys, and humans. *Neuron* **93**, 15-31 (2017).
18. W.T. Newsome, E.B.A. Pare, A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *J. Neurosci.* **8**, 2201-2211 (1988).
19. M. Corbetta, F.M. Miezin, S. Dobmeyer, G.L. Shulman, S.E. Petersen, Selective and divided attention during visual discriminations of shape, color, and speed: functional-anatomy by positron emission tomography. *J. Neurosci.* **11**, 2383-2402 (1991).
20. T. Kurikawa, T. Haga, T. Handa, R. Harukuni, T. Fuai, Neuronal stability in medial frontal cortex set individual variability in decision-making *Nat. Neurosci.* **21**, 1764-1773 (2018).
21. M.N. Shadlen, K.H. Britten, W.T. Newsome, J.A Movshon, A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J. Neurosci.* **16**, 1486-1510 (1996).
22. M.S. Beauchamp, R.W. Cox, E.A. Deyoe, Graded effects of spatial and featural attention on human area MT and associated motion processing areas. *J. Neurophysiol.* **78**, 516-520 (1997).
23. A.C. Huk, R.F. Dougherty, D.J. Heeger, Retinotopy and functional subdivision of human areas MT and MST. *J. Neurosci.* **22**, 7195-7205 (2002).
24. A.S. Kayser, D.T. Erickson, B.R. Bushsdaum, M. D'Esposito, Neural representations of relevant and irrelevant features in perceptual decision making. *J. Neurosci.* **30**, 15778-15789 (2010).
25. N. Kanwisher, J. McDermott, M.N. Chun, The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).

26. G. McCarthy, A. Puce, J.C. Gore, T. Allison, Face-specific processing in the human fusiform gyrus. *J. Cogn. Neurosci.* **9**, 605-610 (1997).
27. A. Ishai, L.G. Ungerleider, A. Martin, J.L. Schouten, J.V. Haxby, Distributed representation of objects in the human ventral visual pathway. *Proc. Natl Acad. Sci. USA* **96**, 9379–9384 (1999).
28. A. Gazzaley, J.W. Cooney, K. McEvoy, R.T. Knight, M. D'Esposito, Top-down enhancement and suppression of the magnitude and speed of neural activity. *J. Cogn. Neurosci.* **17**, 507-517 (2005).
29. W.A. Freiwald, D.Y. Tsao, Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* **330**, 845-851 (2010).
30. R. Epstein, N. Kanwisher, A cortical representation of the local visual environment. *Nature* **392**, 598–601 (1998).
31. E. Koechlin, C. Ody, F. Kouneiher, The architecture of cognitive control in the human prefrontal cortex. *Science* **302**, 1181-1185 (2003).
32. D. Badre, Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends Cogn. Sci.* **12**, 193-200 (2008).
33. K. Jimura, T.S. Braver, Age-related shifts in brain activity dynamics during task switching. *Cereb. Cortex* **20**, 1420–1431 (2010).
34. Y. Kamitani, F. Tong, Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.* **8**, 679–685 (2005).
35. J.D. Haynes, G. Rees, Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* **7**, 523-534 (2006).
36. K.A. Norman, S.M. Polyn, G.J. Detre, J.V. Haxby, Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* **10**, 424-430 (2006).
37. S. Nishimoto, A.T. Vu, T. Naselaris, Y. Benjamini, B. Yu , J.L. Gallant, Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* **21**, 1641-1646 (2011).
38. J. Chikazoe, D. Lee, N. Kriegeskorte, A.K. Anderson, Distinct representation of basic taste qualities in human gustatory cortex. *Nat. Comm.* **10**, 1048 (2019).
39. X. Wang, X. Liang, Z, Jiang, B. A. Nguchu, Y. Zhou, Y. Wang, et al. Decoding and mapping task states of the human brain via deep learning. *Hum. Brain. Mapp.* **41**, 1505-1519 (2020).

40. V.N. Vapnik, Statistical Learning Theory (Wiley, New York, 1998).
41. M. Misaki, Y. Kim, P.A. Bandettini, N. Kriegeskorte, Comparison of multivariate classifiers and response normalization for pattern-information fMRI. *NeuroImage* **53**, 103-118 (2010).
42. K. Jimura, R.A. Poldrack, Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia* **50**, 544-552 (2012).
43. K. Nakahara et al., Associative-memory representations emerge as shared spatial patterns of theta activity spanning the primate temporal cortex. *Nat. Comm.* **7**, 11827 (2016).
44. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural network. *Adv. Neural Inf. Proc. Sys.* **25**, 1097-1105 (2012).
45. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436-444 (2015).
46. R.R. Selvaraju, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV), Venice*, 618-626 (2017).
47. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. Preprint at https://arxiv.org/pdf/1409.1556.pdf (2015).
48. J. Donahue et al., Decaf: A deep convolutional activation feature for generic visual recognition. *Proceedings of the 31st International Conference on Machine Learning, PMLR* **32**, 647-655 (2014).
49. V. Mante, D. Sussillo, K.V. Shenoy, W.T. Newsome, Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78-84 (2013).
50. J. Zhang, N. Kriegeskorte, J.D. Carlin, J.B. Rowe, Choosing the rules: Distinct and overlapping frontoparietal representations of task rules for perceptual decisions. *J. Neurosci.* **33**, 11852-11862 (2013).
51. H. Kumano, Y. Suda, T. Uka, Context-dependent accumulation of sensory evidence in the parietal cortex underlies flexible task switching. *J. Neurosci.* **36**, 12192–12202 (2016).
52. M.W. Cole et al., Multi-task connectivity reveals flexible hubs for adaptive task control. *Nat. Neurosci.* **16**, 1348–1355 (2013).
53. M. D'Esposito, B.R. Postle, The cognitive neuroscience of working memory. *Annu. Rev. Psychol.* **66**, 115-142 (2015).

54. S.M. Courtney, L.G. Ungerleider, K. Keil, J.V. Haxby, Transient and sustained activity in a distributed neural system for human working memory. *Nature* **386**, 608-611 (1997).
55. R. Desimone, J. Duncan, Neural mechanisms of selective visual-attention. *Annu. Rev. Neurosci.* **18**, 193-222 (1995).
56. S. Kastner, L.G. Ungerleider, Mechanisms of visual attention in the human cortex. *Annu. Rev. Neurosci.* **23**, 315-341 (2000).
57. M. Corbetta, G.L. Shulman, Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* **3**, 201-215 (2002).
58. J.A. Lewis-Peacock, B.R. Postle, Temporary activation of long-term memory supports working memory. *J. Neurosci.* **28**, 8765-8771 (2008).
59. D.H. Brainard, The psychophysics toolbox. *Spatial Vision* **10**, 433–436 (1997).
60. M.Y. Chen, K. Jimura, C.N. White, W.T. Maddox, R.A. Poldrack, Multiple brain networks contribute to the acquisition of bias in perceptual decision-making. *Front. Neurosci.* **9**, 63 (2015).
61. K.H. Britten, M.N. Shadlen, W.T. Newsome, J.A. Movshon, Responses of neurons in macaque MT to stochastic motion signals. *Visual Neurosci.* **10**, 1157-1169 (1993).
62. J. Palmer, A.C. Huk, M.N. Shadlen, The effect of stimulus strength on the speed and accuracy of a perceptual decision. *J. Vision* **5**, 376-404 (2005).
63. S. Moeller, E. Yacoub, C.A. Olman, E. Auerbach, J. Strupp, N. Harel, K. Uğurbil, Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magn. Reson. Med.* **63**, 1144-1153 (2010).
64. D. B. Worsley, K. J. Friston, Analysis of fMRI time-series revisited again. *NeuroImage* **2**, 173-181 (1995).
65. T.E. Nichols, A.P. Holmes, Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* **15**, 1-25 (2001).
66. A. Eklund, T.E. Nichols, H. Knutsson, Cluster failure: Inflated false positives for fMRI. *Proc. Natl Acad. Sci. USA* **113**, 7900-7905 (2016).
67. T. Yarkoni, R.A. Poldrack, T.E. Nichols, D.C. van Essen, T.D. Wager, Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665-670 (2011).

68. K.J. Friston, L. Harrison, W. Penny, Dynamic causal modelling. *NeuroImage* **19**, 1273-1302 (2003).
69. S. Konishi et al., Hemispheric asymmetry in human lateral prefrontal cortex during cognitive set shifting. *Proc. Natl Acad. Sci. USA* **99**, 7803-7808 (2002).
70. K.J. Friston et al., Bayesian model reduction and empirical Bayes for group (DCM) studies. *NeuroImage* **128**, 413-431 (2016).
71. W.D. Penny et al., Comparing families of dynamic causal models. *PLoS Comput. Biol.* **6**, e1,000709 (2010).
72. J. Deng et al,. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 248-255 (2009).
73. S.J. Pan, Q.A. Yang, A Survey on Transfer Learning. *IEEE T. Knowl. Data En.* **22**, 1345-1359 (2010).
74. M.F. Glasser et al., The human connectome project's neuroimaging approach. *Nat. Neurosci.* **19**, 1175-1178 (2016).
75. D.M. Barch et al., Function in the human connectome: task-fMRI and individual differences in behavior. *NeuroImage* **80**, 169-189 (2013).
76. M.F. Glasser et al., A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171-178 (2016).
77. N. Kriegeskorte, R. Goebel, P. Bandettini, Information-based functional brain mapping. *Proc. Natl Acad. Sci. USA* **103**, 3863–3868 (2006).

Supplementary materials

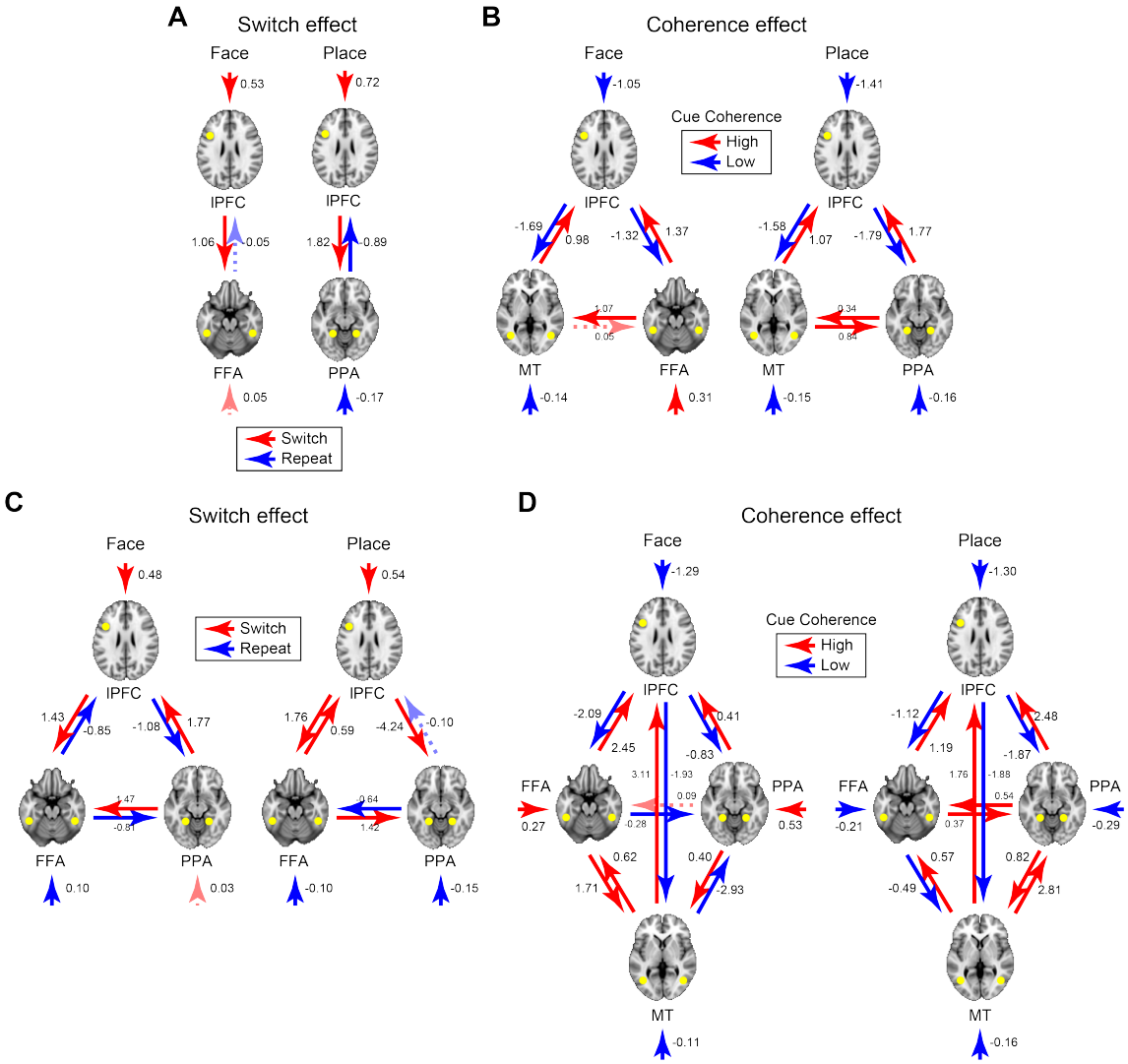
Title: Flexible decision under ambiguous cues reverses neural signaling from occipitotemporal to prefrontal cortex

Authors: Kaho Tsumura,[1] Keita Kosugi,[1] Ryuta Aoki,[2] Masaki Takeda,[2] Junichi Chikazoe,[3] Kiyoshi Nakahara,[2] Koji Jimura[1,2]

Affiliations: [1] Department of Biosciences and Informatics, Keio University, Yokohama, Japan
[2] Research Center for Brain Communication, Kochi University of Technology, Kami, Japan
[3] Supportive Center for Brain Research, National Institute for Physiological Sciences, Okazaki, Japan

Correspondence should be addressed to:
Koji Jimura, Ph.D.
Department of Biosciences and Informatics,
Keio University,
3-14-1 Hiyoshi Kohoku-ku,
Yokohama, 223-0061, Japan
Email: jimura@bio.keio.ac.jp
Phone: +81-45-566-1625

Supplementary figures



cont.

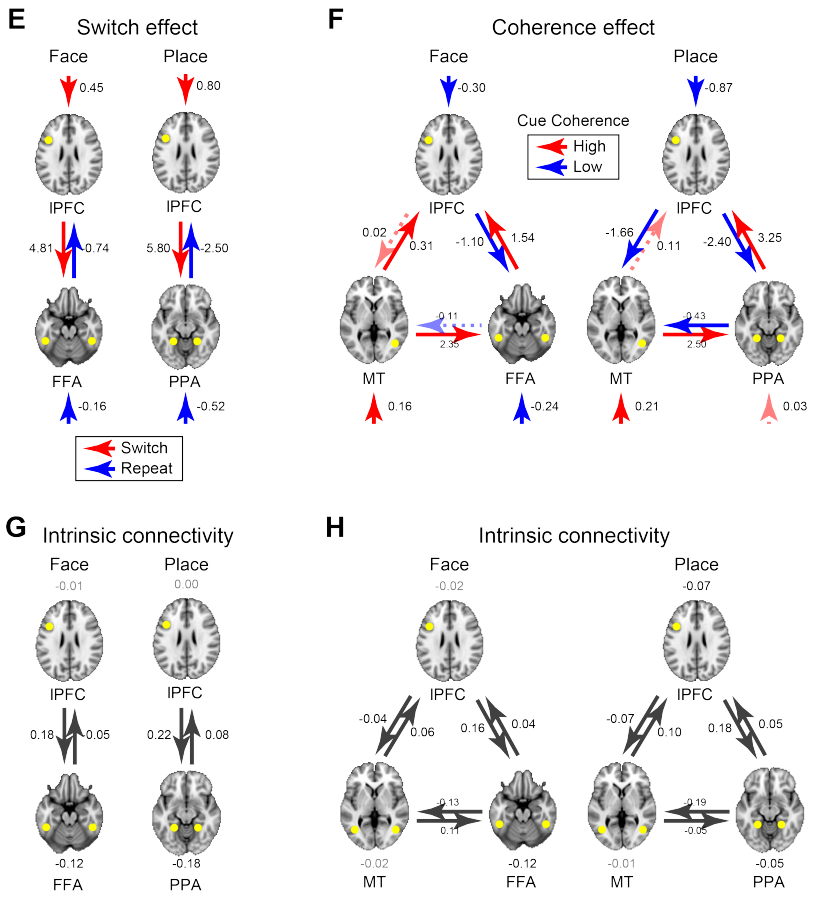


Fig. S1. Effective connectivity analysis. (A-B) Parameters estimated without parametric empirical Bayes. (A) Switching effect. (B) Coherence effect. (C-D) The numbers of ROIs in the model was changed. (C) Switching effect was examined in a model involving the lateral prefrontal cortex (lPFC), fusiform face area (FFA), and parahippocampal place area (PPA) for both of the face and place tasks. (D) Coherence effect was examined in a model involving lPFC, FFA, PPA, and middle temporal (MT) regions. (E-F) The definition of ROIs. (E) Switching effect. (F) Coherence effect. The other analysis procedures were identical to those in Figs. 2D-E. (G-H) Task unrelated intrinsic connectivity, estimations of "A" matrix in the dynamic causal modeling. (G) Switching effect (H) coherence effect. Gray solid lines indicate statistical significance.

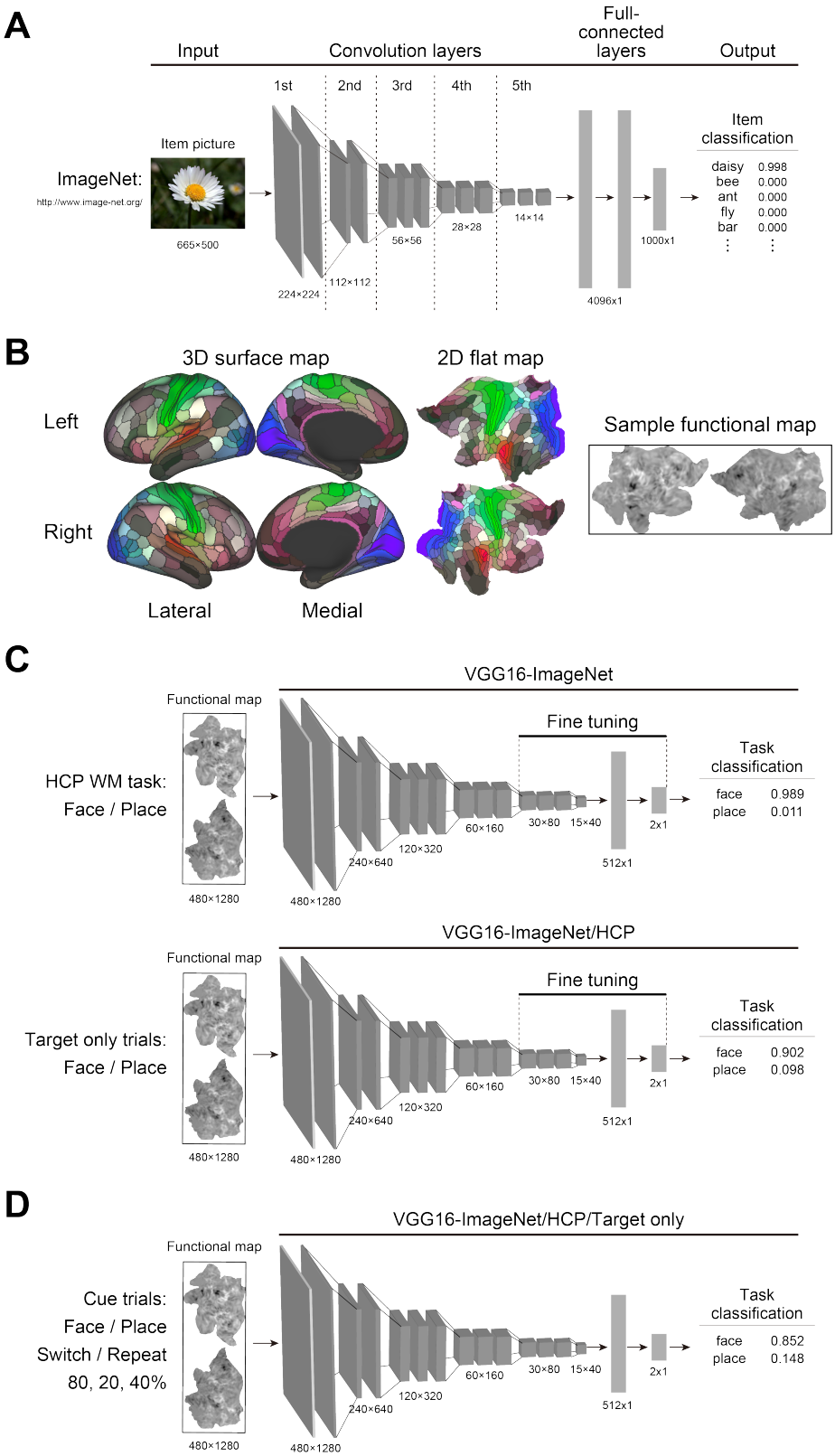


Fig. S2. Training and testing procedure for the convolutional neural network classifier (CNN). (A) VGG16 classifier pre-trained for concrete item pictures available from ImageNet. The classifier

consists of 5-tier convolution layers and 3-tier full-connected layers, and is capable of classifying pictures of concrete items into 1,000 categories. (B) 3D surface maps (*left*) are mapped to 2D flat maps (*middle*). For illustration purpose, colored areas are cortical subregions functionally and anatomically segmented in prior work (ref. 42). Sample flattened 2D z-maps for the contrast face vs. fixation and place vs. fixation (*right*). (C) Training procedures for 2D-flat images. ImageNet-pretrained VGG16 model (A) was first retrained such that it classified activation maps for the face and place task during working memory task distributed by Human Connectome Project (HCP) (top). The retrained model was further retrained such that it classifies activation maps for the face and place task during target only trials in the current experiment (*bottom*). (D) Testing procedure. The classifier trained by the two stepwise procedure (C) was tested for classification of face and place tasks during cued trials involving switch and repeat trials with different coherence levels.

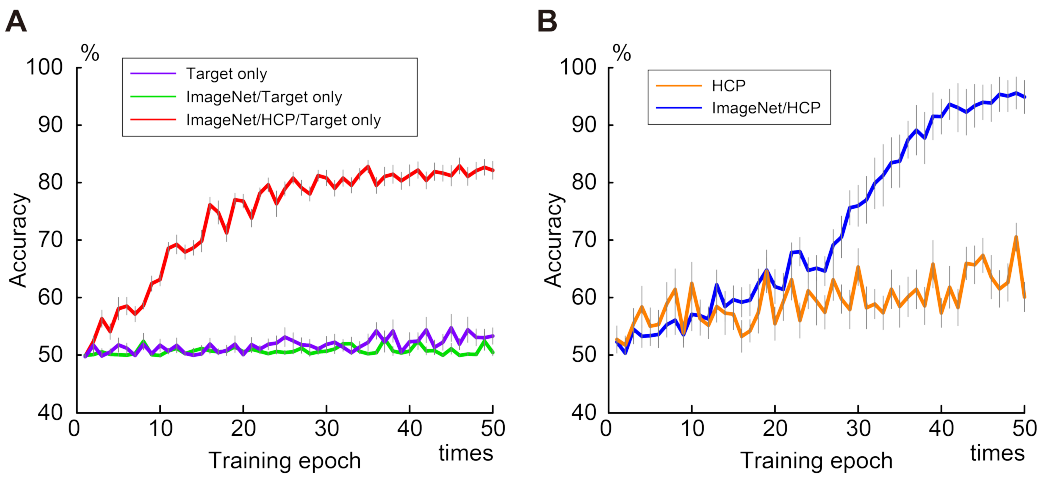


Fig. S3. Classification accuracy of convolutional neural network during training epoch. Vertical and horizontal axes indicate validation accuracy, and training epoch, respectively. (A) Validation accuracy for face/place classification of the current target-only trials. Purple line indicates the learning curve of a model trained based on the activation maps of target-only trials from a randomly initialized parameter. Green line indicates the learning curve of a model re-trained by the activation maps of target-only trials from pre-trained parameter by ImageNet. Red line indicates the learning curve of a model re-trained by the activation maps of target-only trials from the pre-trained parameter by Human Connectome Project (HCP) working memory maps and ImageNet (i.e. green line). Error bars indicate standard error of the mean across cross-validation testing. (B) Validation accuracy for face/place classification of working memory trials provided by HCP. Orange line indicates the learning curve of a model trained based on the activation maps of working memory trials from a randomly initialized parameter. Blue line indicates the learning curve of a model trained by activation maps of HCP working memory trials from a pre-trained parameter by ImageNet. Error bars indicate standard error of the mean across cross-validation testing.

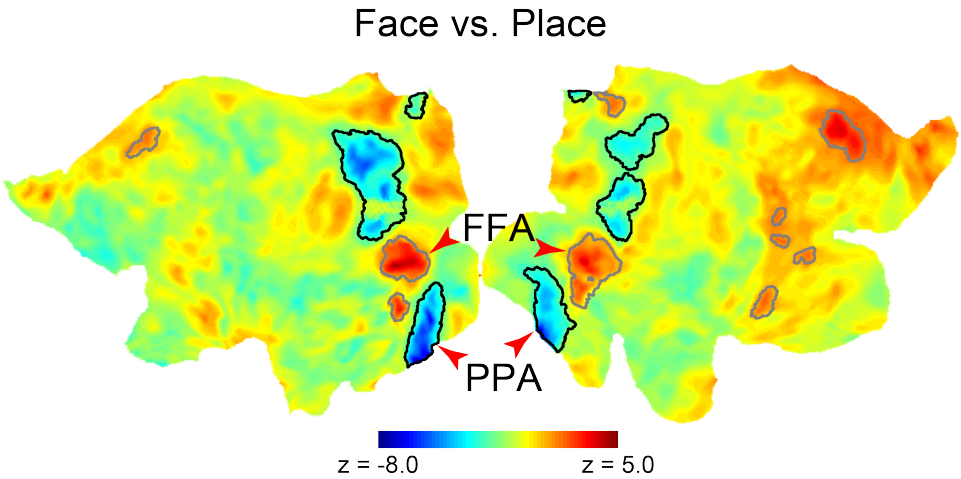


Fig. S4. Statistical activation z-map for contrast face vs. place tasks in univariate analysis. Data are derived from those shown in Fig. 2C, but maps are overlaid onto 2D-flat map of the cortical surface of the brain. Hot and cool colors indicate greater activity in face and place tasks, respectively. Closed lines indicate prominent activity during face (gray) and place (black), respectively, as identified by a group-level analysis (Table S3). The fusiform face area (FFA) and parahippocampal place area (PPA) are indicated by red arrow heads.

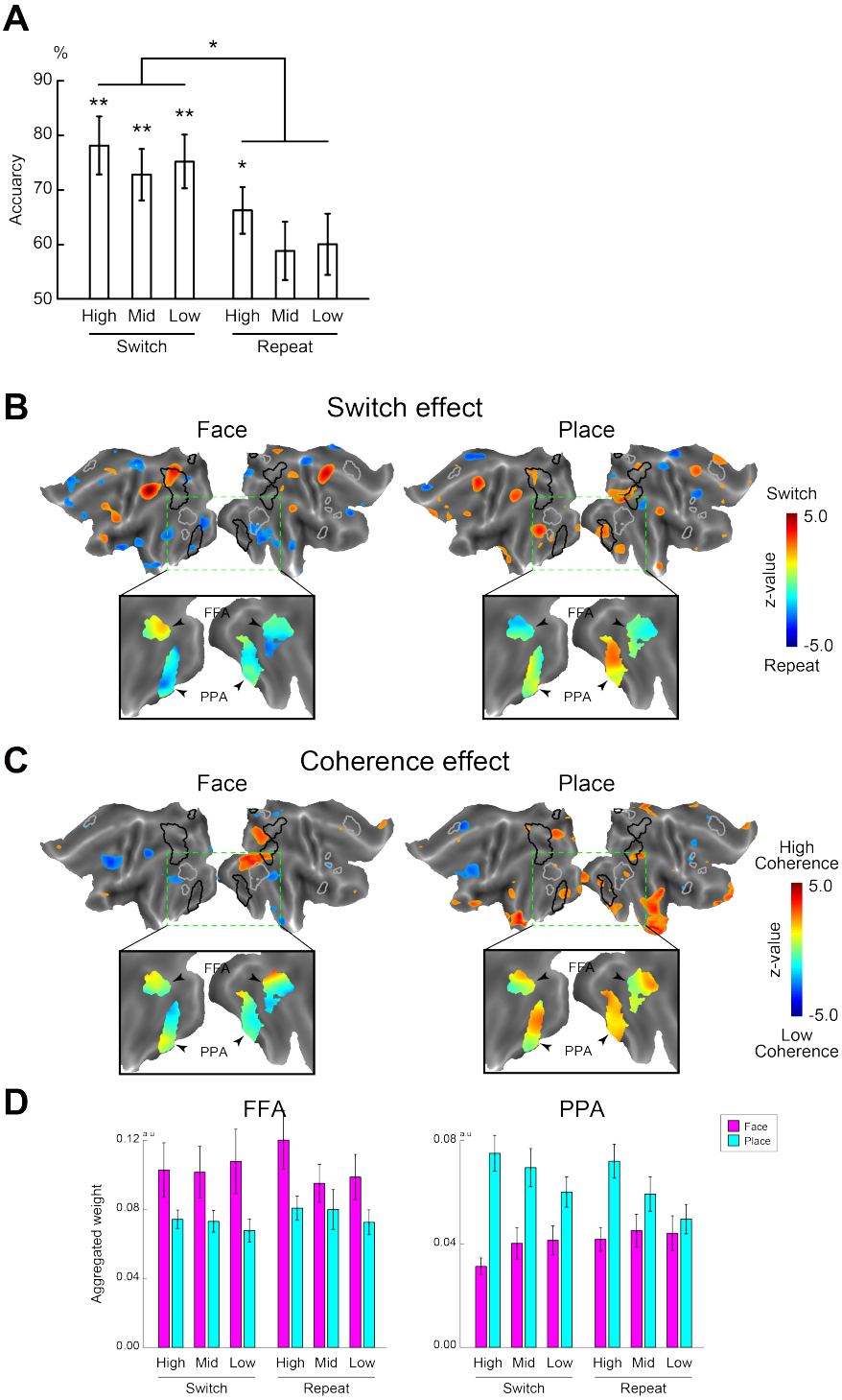


Fig. S5. Classification accuracy and weight maps of CNN classifier. A leave one-subject-out procedure was used in training and validation of the CNN classifier. (A) Classification accuracy for each task condition. (B) Visualization of aggregated weight contrasts for switch vs. repeat trials in face task (*left*) and place task (*right*). Occipitotemporal regions in rectangular boxes with green broken lines were expanded below. (C) Visualization of weight contrast for coherence effect in face task (*left*) and place task (*right*). (D) Regions of interests (ROIs) analysis. All formats are similar to those in Figs. 3B/D-F.

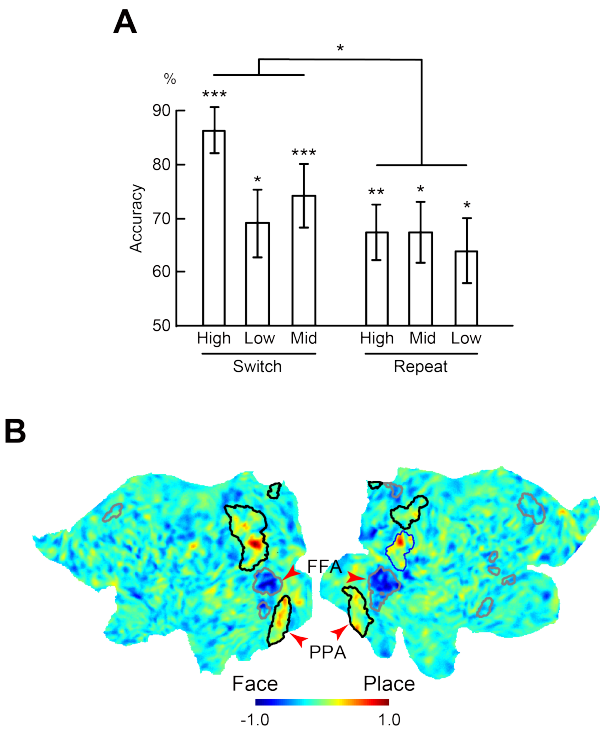


Fig. S6. The support vector machine classifier trained by target-only trials. (A) Classification accuracy for cue trials. Error bars indicate standard error of the mean across participants. Formats are similar to those in Fig. 2B. (B) Weight map of SVM. Maps are overlaid onto flat maps of the cortical brain. Hot and cool colors indicate high and low weights, respectively. Gray and black lines overlaid on flat map indicate clusters significantly activated during face and place tasks in univariate analysis, respectively (Figs. 2C and S4). FFA and PPA was indicated by red arrow heads.

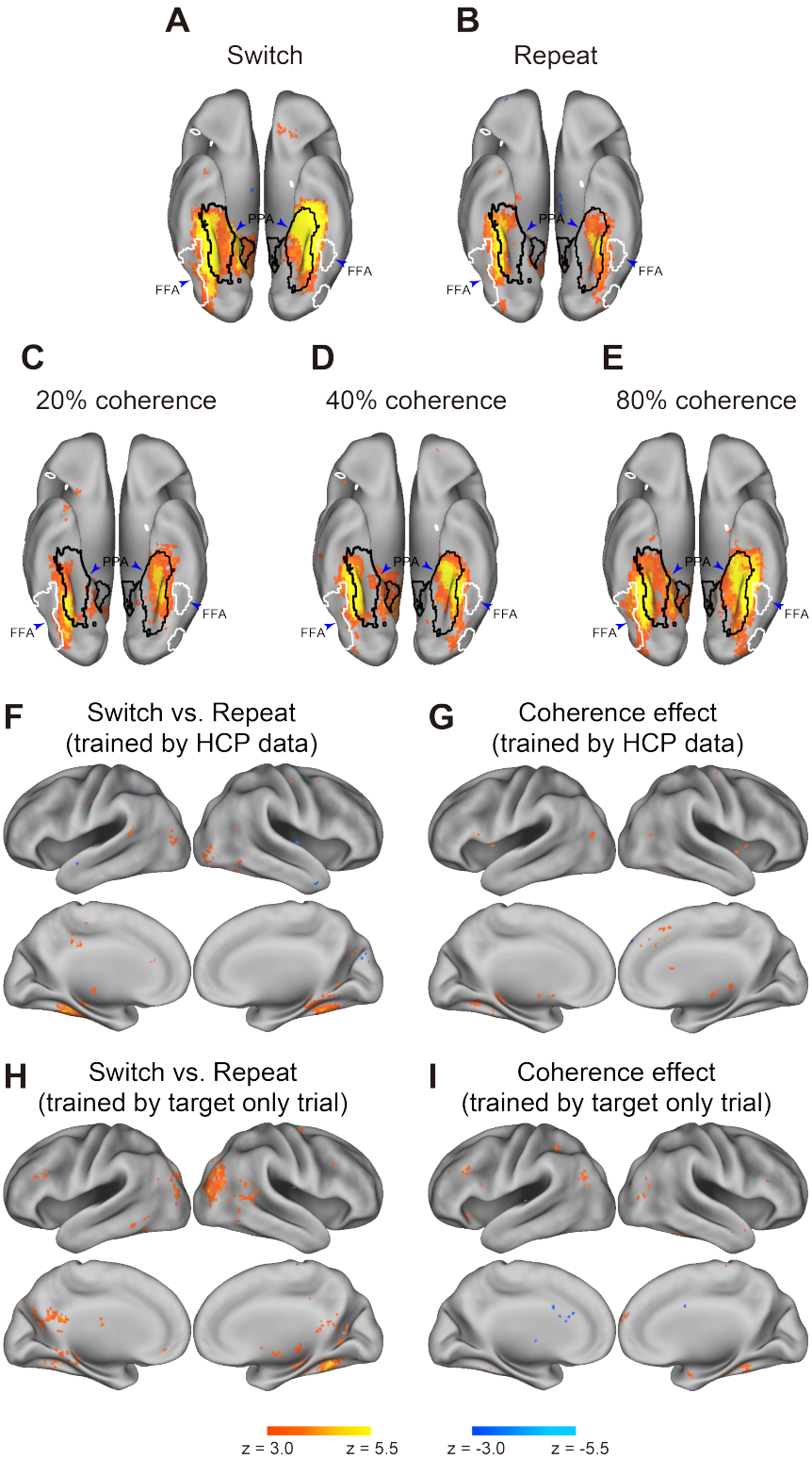


Fig. S7. Searchlight multi-variate pattern analysis. Hot and cool colors indicate statistical level for classification accuracy relative to chance level. Maps are overlaid onto 3D surface of the brain and displayed from a ventral view. White and black closed lines overlaid onto 3D surface of the brain indicate significant cluster for contrast face vs. place tasks in univariate analysis, respectively (Figs. 2C and S4).

The fusiform face area (FFA) and parahippocampal place area (PPA) are indicated by blue arrow heads. (A) Switch trial. (B) Repeat trial. (C) 20% coherence trial. (D) 40% coherence trial. (E) 80% coherence trial. (F) Accuracy difference between switch and repeat trials with a model trained by Human Connectome Project (HCP) data. Hot and cool colors indicate higher accuracy in switch and repeat trials, respectively. (G) Differential classification accuracy depending on coherence effect with a model trained by HCP data. Hot and cool colors indicate higher accuracy in high and low coherent trials, respectively. (H) Accuracy difference between switch and repeat trials with model trained by target-only trials of the current dataset. (I) Differential classification accuracy depending on the coherence effect with model trained data of target-only trials. The formats are similar to those in panel (F).

Supplementary tables

Table S1. Brain regions showing significant signal increase and decrease in the contrast of switch vs. repeat trials. Coordinates are listed in MNI space. Positive and negative z-values indicate increase in switch and repeat trials, respectively. BA indicates Brodmann areas and is approximate.

Area	x	y	z	z-value	BA
Frontal cortex	-30	0	54	5.32	6
	-6	8	52	4.98	6
	-18	12	62	4.80	6
	12	12	68	4.66	6
	-42	0	48	4.52	6
	0	28	32	4.42	32
	-6	4	70	4.12	6
	10	18	28	4.00	32
	10	14	56	3.59	6
	-40	12	26	4.62	44
	-38	-2	36	3.61	6
	22	0	48	4.16	6
	38	0	40	3.63	6
Parietal cortex	-28	-70	36	5.62	39
	0	-36	26	5.49	23
	20	-60	28	5.55	31
	-4	-16	32	5.35	23
	-10	-68	44	4.98	7
	-30	-56	44	4.67	39
	-50	-40	50	4.63	40
	-38	-48	50	4.58	40
	-14	-68	28	4.39	7
	6	-38	38	4.02	23
	16	-68	42	4.00	7
	16	-66	54	3.87	7
	-36	-82	26	3.64	19
	-4	-44	12	3.63	30
	12	-42	28	3.33	23

Table S2. Brain regions showing significant parametrical effect with motion coherence. Coordinates are listed in MNI space. Positive and negative z-values indicate increase in high and low coherent trials, respectively. *: corrected within motion-related brain regions (see Methods).

Area	x	y	z	z-value	BA
Frontal cortex	0	18	50	-5.90	8
	10	24	46	-5.59	8
	54	20	38	-5.43	9/44
	46	14	4	-5.43	44
	52	14	16	-5.21	44
	-56	12	28	-5.16	44
	36	4	54	-4.62	6
	48	6	34	-4.62	6
	-32	0	50	-4.49	6
	-38	10	32	-4.43	8
	8	30	28	-4.41	8
	42	24	28	-4.39	9
	2	42	-16	4.37	11
	4	30	-18	4.33	11
	20	12	46	-4.31	6
	-50	8	10	-4.25	44
	-42	24	26	-4.08	9/44
	38	12	20	-3.82	44
	40	36	-4	-3.82	47
	-8	46	-6	3.80	10
	24	8	62	-3.79	6
	34	34	34	-3.78	9
	-52	18	-6	-3.67	47
	48	38	14	-3.35	46
	-44	38	28	-3.21	9/10
Parietal cortex	-18	-56	74	4.70	7
	-16	-48	36	4.37	31
	-12	-68	30	-4.28	7
	-24	-64	38	-4.25	39
	-30	-52	46	-4.19	7/39
	14	-48	18	-4.19	23
	58	-46	36	-4.08	39/40
	-48	-40	48	-4.07	40
	26	-40	64	4.03	1/5
	20	-58	26	-4.02	23
	42	-50	52	-3.86	40
	-30	-42	56	3.85	5/7
	-40	-46	40	-3.66	39
	6	-46	32	3.63	23
	42	-46	38	-3.50	39
	32	-42	46	-3.27	7
Temporal cortex	-42	-54	-8	-5.45	37
	-30	-58	-10	-4.10	37
	56	-64	2	3.75*	37/19
Others	32	24	4	-5.66	insula
	-30	26	0	-5.56	insula
	0	-32	-2	-4.66	brain stem
	12	-18	-8	-4.30	brain stem
	8	-10	4	-3.83	thalamus
	-16	-28	-4	-3.60	thalamus

Table S3. Brain regions showing significant signal increase and decrease in the contrast of face vs. place in target-only trials. Coordinates are listed in MNI space. Positive and negative z-values indicate increase in face and place trials, respectively.

Area	x	y	z	z-value	BA
Frontal cortex	4	32	46	5.00	8
	38	28	12	4.97	45
	8	40	38	4.23	9
	46	2	28	3.96	6
	48	42	12	3.49	46
	48	18	40	3.40	8
	-6	36	36	3.38	8
Parietal cortex	-22	-74	44	-5.90	7
	-20	-58	12	-5.88	23
	-18	-70	58	-5.79	7
	-10	-76	50	-5.38	7
	18	-74	44	-4.78	7
	2	-60	30	4.60	31
	12	-68	62	-4.47	7
	40	-74	40	-4.37	39
	2	-48	18	4.09	23
Temporal cortex	26	-38	-14	-8.03	37
	-28	-44	-12	-7.46	37
	-28	-30	-20	-6.81	37
	42	-58	-14	5.72	37
	-44	-54	-20	4.91	37
	-10	-46	4	-4.38	18/30
	52	-64	0	4.06	37
	60	-64	-10	3.53	37
Occipital cortex	-40	-82	22	-6.69	19
	44	-72	-12	6.36	19
	40	-74	20	-6.27	19
	-48	-78	-10	5.75	19
	-20	-66	22	-5.72	19
	12	-46	4	-5.65	18
	18	-56	16	-5.61	18/23
	-26	-66	-12	-5.14	19
	24	-64	26	-4.18	19
	-32	-86	38	-4.08	19
	-48	-82	2	3.73	19
	-40	-70	-4	3.41	19
Others	-18	-6	-16	5.09	amygdala
	-30	2	-22	4.54	amygdala
	32	18	2	3.80	insula

Table S4. Brain regions showing significant accuracy in classifying task dimension (face/place) for target-only trials. Coordinates are listed in MNI space. Positive z-values indicate higher accuracy than chance level.

Area	x	y	z	z-value	BA
Frontal cortex	-10	36	42	5.66	8
	-6	24	32	4.05	32
Parietal cortex	24	-68	48	6.67	7
	-28	-70	26	6.57	39
	26	-82	46	6.40	7
	14	-78	56	6.10	7
	24	-68	66	5.78	7
	8	-64	22	5.60	31
	-24	-56	32	5.55	7/39
	-12	-46	24	5.40	23
	10	-84	46	5.34	7/19
	26	-54	28	5.23	23
	-10	-74	40	5.06	7
	-32	-78	38	4.96	39
	14	-60	62	4.86	7
	-26	-60	48	4.79	7
	42	-66	32	4.76	39
	36	-74	50	4.69	39
	-14	-64	56	4.68	7
	-14	-66	32	4.59	7
	50	-76	30	4.47	39
	-8	-76	52	4.44	7
	36	-82	26	4.41	19/39
	30	-68	32	4.32	7
Temporal cortex	32	-50	-18	7.86	37
	-28	-42	-20	7.60	37
	-40	-50	-10	7.46	37
	38	-38	-16	7.34	37
	-38	-32	-16	7.29	37
	-34	-26	-28	6.94	36
	-36	-46	4	6.40	21
	-40	-18	-20	6.36	20
	-40	-36	-2	6.28	21
	-26	-20	-20	6.19	36
	16	-40	0	6.00	18/30/36
	18	-26	-20	5.97	36
	14	-48	14	5.75	23
	44	-60	-16	5.56	37
	-34	8	-26	5.47	38
	-18	-48	2	5.43	30
	-48	-44	-24	5.18	37
	-10	-38	-4	5.01	36
	26	-14	-22	4.88	36
	-28	4	-42	4.82	38
	54	-50	-20	4.78	37
	54	-62	-8	4.75	37
	16	-8	-28	4.59	36
	6	-34	-4	4.28	36
	-38	-6	-20	4.12	22
Occipital cortex	34	-72	-12	7.26	19/37
	22	-70	12	6.96	17

	-44	-80	26	6.91	19
	16	-70	32	6.60	19
	22	-52	0	6.51	18/19
	30	-60	18	6.43	18
	-40	-68	26	6.37	19/39
	32	-84	-18	6.04	19
	-44	-76	14	6.01	19
	48	-76	8	6.00	19
	20	-64	0	5.95	19
	-30	-88	-10	5.94	18
	-32	-68	-10	5.90	19
	36	-84	-4	5.90	18
	20	-84	34	5.76	19
	46	-84	-14	5.67	19
	-32	-64	16	5.54	19
	-12	-64	20	5.43	18
	18	-90	22	5.41	18
	-32	-78	8	5.22	18
	-26	-56	-2	5.11	19
	-42	-86	0	4.95	18
	30	-96	10	4.50	18
	4	-60	8	4.49	18
	22	-76	-2	4.37	18
	54	-76	-4	4.35	19
	32	-84	8	4.27	19
Others	-24	-28	-8	7.66	hippocampus
	28	-42	4	6.64	hippocampus
	16	-46	-24	6.51	cerebellum
	-16	-32	-22	6.37	cerebellum
	-34	-68	-22	6.36	cerebellum
	-30	-56	-26	6.25	cerebellum
	38	-26	-8	6.14	hippocampus
	-4	-34	6	6.08	callosum
	26	-26	-8	5.63	hippocampus
	-48	-56	-32	5.59	cerebellum
	14	-58	-14	5.58	cerebellum
	-14	-14	-10	5.52	brain stem
	-26	-10	-10	5.48	hippocampus
	12	-38	-14	5.30	cerebellum
	-50	-68	-26	4.89	cerebellum
	10	-42	-40	4.88	cerebellum
	32	-32	-34	4.75	cerebellum
	-4	-24	-2	4.70	thalamus
	-22	-4	-22	4.70	hippocampus
	20	-70	-20	4.64	cerebellum
	16	-12	-12	4.63	hippocampus
	-32	6	-14	4.62	insula
	32	-72	-26	4.59	cerebellum
	16	2	-12	4.39	amygdala
	48	-56	-30	4.20	cerebellum
	28	-2	-14	4.15	amygdala
	-4	-38	-38	4.10	cerebellum
	8	-44	-56	3.96	cerebellum
	10	-54	-38	3.83	cerebellum
	-16	0	-4	5.05	globus pallidus

Table S5. Brain regions showing significant accuracy difference for face-place classification between switch and repeat trials. Coordinates are listed in MNI space. Positive and negative z-values indicate higher accuracy in switch and repeat trials, respectively.

Area	x	y	z	z-value	BA
Parietal cortex	34	-62	22	4.08	39
	14	-66	22	3.75	31
	22	-56	28	3.26	23/31
Temporal cortex	36	-50	-18	5.79	37
	-30	-48	-16	5.54	37
	-36	-22	-28	5.35	36
	22	-42	-16	5.24	37
	-34	-34	-16	4.93	36
	44	-54	-4	4.71	37
	36	-62	-8	4.33	37
	54	-62	-12	4.30	37
	20	-30	-12	3.88	36
	-40	-48	0	3.84	37
	26	-58	-14	3.80	37
	-26	-20	-20	3.38	36
	-42	-24	-16	3.26	20
	44	-50	8	3.25	37
Occipital cortex	22	-72	12	5.01	17
	-36	-72	4	4.91	18/19
	32	-74	-4	4.48	19
	-42	-84	10	4.05	19
	36	-88	-4	3.98	18
	44	-80	2	3.96	19
	22	-48	0	3.92	18
	-30	-82	-2	3.90	18
	24	-72	28	3.78	19
	18	-62	-4	3.28	19
Others	12	-52	-12	4.23	cerebellum
	-38	-34	-4	3.80	caudate
	28	-46	-30	3.69	cerebellum
	30	-38	-2	3.30	hippocampus

Table S6. Brain regions showing significant accuracy modulation for face and place tasks depending on coherence level. Coordinates are listed in MNI space. Positive and negative z-values indicate higher accuracy in high and low coherent trials, respectively.

Area	x	y	z	z-value	BA
Temporal cortex	-30	-62	-8	4.85	37
	-24	-38	-8	4.29	36
	-34	-52	-20	4.25	37
	-42	-40	-18	4.20	37
	-14	-28	-12	3.78	36
Occipital cortex	-18	-54	-8	3.78	19
Others	-38	-32	-4	4.18	caudate
	-6	-30	0	3.37	thalamus
	-34	-46	-32	3.17	cerebellum

Table S7. Brain regions showing significant accuracy to classify task dimension (face/place) in switch trials. Coordinates are listed in MNI space. Positive z-values indicate higher accuracy than chance level.

Area	x	y	z	z-value	BA
Parietal cortex	14	-66	20	6.45	31
	24	-54	10	5.48	23
	2	-52	20	5.43	23
	20	-70	34	5.35	19
	26	-58	32	5.10	31
	22	-74	52	5.08	7
	14	-54	26	4.77	23
	-46	-66	14	4.77	39
	-18	-52	22	4.27	23
	-14	-78	46	4.19	7
	42	-76	36	4.14	39
	50	-54	18	4.04	39
	34	-68	30	4.02	39
	-6	-64	20	3.92	31
	12	-70	60	3.87	7
	28	-60	62	3.84	7
	24	-58	46	3.68	7
	8	-70	44	3.64	7
	-34	-82	42	3.63	39
	34	-82	44	3.39	7
	-44	-68	38	3.37	39
	-8	-56	30	3.28	31
	34	-68	44	3.26	39
Temporal cortex	-32	-48	-12	8.39	37
	34	-50	-16	7.26	37
	-22	-28	-20	6.88	36
	-32	-38	-22	6.82	37
	16	-32	-8	6.73	36
	-34	-18	-24	6.64	36
	-42	-38	-4	6.51	21
	34	-64	-8	6.07	37
	34	-30	-24	5.61	37
	22	-32	-20	4.99	37
	42	-46	0	4.78	21
	54	-52	2	4.73	37
	8	-38	8	4.03	30
	58	-52	-20	4.03	37
	54	-68	-8	3.86	37
	42	-20	-28	3.59	20
Occipital cortex	16	-54	-4	6.28	19
	-34	-60	-4	6.18	19/37
	44	-58	-2	5.56	19/37
	26	-66	14	5.48	17/23
	-48	-84	8	5.46	19
	-36	-72	4	5.29	18/19
	32	-76	-10	5.24	19
	-28	-72	-8	5.13	19
	-48	-78	22	5.00	19
	-38	-74	28	4.98	19
	-20	-54	-2	4.90	19
	44	-80	2	4.77	19

	18	-78	24	4.74	18/19
	-32	-62	22	4.67	19
	-34	-88	6	4.59	18
	6	-66	8	4.54	18
	22	-82	38	4.51	19
	54	-68	22	4.44	19
	-28	-68	12	4.30	17
	56	-64	4	4.11	19/37
	36	-94	-2	4.09	18
	22	-90	22	4.03	19
	22	-80	12	3.92	17
	-28	-86	-10	3.75	18
	-46	-86	-4	3.66	18/19
	28	-90	-14	3.61	18
	-18	-74	16	3.41	17
	-24	-88	14	3.41	19
	32	-76	4	3.32	18
	-8	-46	2	3.25	18
	-36	-84	20	3.22	19
Others	-26	-32	-6	6.96	hippocampus
	28	-26	-8	6.34	hippocampus
	20	-44	-18	6.30	cerebellum
	-34	-28	-14	6.20	hippocampus
	30	-46	-28	6.06	cerebellum
	28	-64	-20	5.94	cerebellum
	-16	-60	-20	5.36	cerebellum
	-22	-46	-28	5.14	cerebellum
	34	-36	0	4.90	caudate
	-22	-42	2	4.86	hippocampus
	40	-28	-8	4.46	caudate
	2	-56	0	4.28	cerebellum
	-32	-80	-24	4.02	cerebellum
	-36	-54	-28	3.67	cerebellum
	38	-16	-16	3.30	hippocampus
	-40	-66	-22	3.20	cerebellum

Table S8. Brain regions showing significant accuracy to classify task dimension (face/place) in repeat trials. Positive z-values indicate higher accuracy than chance level.

Area	x	y	z	z-value	BA
Parietal cortex	0	-50	16	4.84	23
	12	-42	18	4.58	30
	10	-62	26	4.53	31
	36	-50	4	3.88	30
	-14	-56	32	3.66	31
	-6	-58	8	3.62	23
	-22	-52	22	3.53	23
	28	-56	12	3.46	23
	-32	-50	4	3.32	30
Temporal cortex	30	-42	-22	6.04	37
	-34	-50	-10	6.01	37
	30	-56	-18	5.96	37
	36	-26	-20	5.34	36
	38	-38	-14	5.28	37
	-38	-32	-18	4.20	20/36/37
	24	-20	-26	4.15	36
	-30	-16	-26	4.02	36
	-42	-54	-20	3.77	37
	44	-26	-2	3.42	22
	38	-76	-16	3.39	37
	22	-48	-12	3.16	37
Occipital cortex	34	-66	-6	5.06	19/37
	-36	-66	-16	4.92	19
	-46	-78	6	4.83	19
	16	-54	-2	4.79	19
	-32	-94	-14	4.67	18
	34	-88	-2	4.46	18
	-40	-66	20	4.21	19/39
	30	-90	-18	4.18	18/19
	-28	-56	30	4.16	7
	-26	-70	16	4.12	17
	-24	-74	-14	4.09	18
	-14	-64	22	4.08	18
	-34	-80	-10	4.06	19
	24	-70	20	3.84	19
	14	-66	6	3.67	17
	46	-86	6	3.52	19
	-46	-80	24	3.15	19
Others	-34	-36	-4	5.78	caudate
	-26	-40	-26	5.55	cerebellum
	28	-26	-8	5.46	hippocampus
	22	-32	-24	5.16	cerebellum
	30	-12	-8	4.61	putamen
	16	-52	-24	4.38	cerebellum
	36	-38	2	4.30	caudate
	-30	-20	-14	4.15	hippocampus
	-30	-82	-22	3.98	cerebellum
	-24	-60	-24	3.82	cerebellum
	12	-38	-16	3.53	cerebellum
	18	-16	-16	3.47	hippocampus
	30	-36	-34	3.46	cerebellum

Table S9. Brain regions showing significant accuracy to classify task dimension (face/place) in 20% coherence trials. Coordinates are listed in MNI space. Positive z-values indicate higher accuracy than chance level.

Area	x	y	z	z-value	BA
Frontal cortex	26	14	-18	4.54	47
	6	60	36	4.49	9
	14	54	44	3.98	9
	10	42	42	3.67	8/9
Parietal cortex	14	-66	22	4.87	31
	0	-52	18	4.81	23
	20	-52	30	4.63	23
	-48	-66	14	4.40	39
	-48	-76	26	4.40	39
	26	-60	10	4.07	23
	-40	-56	18	3.64	39
	4	-64	32	3.38	31
Temporal cortex	30	-56	-16	6.33	37
	-36	-50	-8	6.32	37
	-26	-46	-20	5.53	37
	32	-40	-24	4.85	37
	32	0	-30	4.49	36
	40	-32	-14	4.42	36
	-32	-54	4	4.19	18/30
	44	-24	-22	3.75	20
	30	14	-32	3.74	38
	58	-66	-12	3.68	37
	58	-58	2	3.39	37
	48	-68	-4	3.36	37
Occipital cortex	36	-64	-4	4.92	19
	-48	-78	8	4.89	19
	16	-54	-4	4.85	19
	32	-84	-4	4.66	18
	-24	-64	-12	4.60	19
	-16	-56	-4	4.41	19
	28	-60	22	4.29	18
	16	-42	0	4.07	18
	54	-72	6	4.03	19/37
	8	-60	10	3.94	18
	-34	-70	-2	3.87	19
	50	-82	0	3.76	19
	-32	-72	-18	3.51	19
	-44	-88	2	3.43	18
	-40	-84	36	3.38	19
Others	-32	-34	-10	6.06	hippocampus
	30	-70	-20	5.01	cerebellum
	-20	-30	-26	4.84	cerebellum
	16	-60	-18	4.71	cerebellum
	-32	-20	-20	4.60	hippocampus
	20	-48	-24	4.47	cerebellum
	22	-30	-26	4.41	cerebellum
	28	-28	-10	4.07	hippocampus
	32	-62	-30	3.52	cerebellum
	30	-42	4	3.52	hippocampus

Table S10. Brain regions showing significant accuracy to classify task dimension (face/place) in 40% coherence trials. Coordinates are listed in MNI space. Positive z-values indicate higher accuracy than chance level.

Area	x	y	z	z-value	BA
Parietal cortex	34	-60	-8	5.89	23
	6	-48	18	5.35	23
	-22	-68	48	5.10	7
	20	-46	10	5.08	30
	34	-46	4	4.84	30
	28	-68	50	4.75	7
	8	-60	20	4.75	31
	22	-58	42	4.54	7
	-8	-46	16	4.54	23
	32	-58	6	4.41	23
	-28	-64	60	4.34	7
	24	-52	20	4.32	23
	-16	-60	58	4.18	7
	-16	-56	22	4.17	23
	-28	-62	24	4.13	39
	-40	-74	30	4.08	39
	-2	-58	50	4.04	7
	46	-78	34	3.97	39
	6	-54	56	3.93	7
	-22	-82	46	3.91	7
	4	-46	50	3.74	31
	-14	-68	38	3.69	7
	36	-60	54	3.39	7
Temporal cortex	-34	-38	-10	6.99	36
	30	-52	-16	6.94	37
	-38	-50	-10	6.60	37
	-28	-34	-20	6.29	37
	-30	-60	-14	5.27	37
	38	-30	-24	5.11	37
	46	-42	-6	4.49	21
	-36	-38	-28	4.06	37
	-14	-26	-20	4.02	36
	-22	-50	-10	4.01	19/37
	44	-58	-22	3.82	37
Occipital cortex	-46	-80	10	6.51	19
	20	-70	14	5.35	17
	-34	-72	6	5.24	18/19
	-40	-68	16	5.01	19/39
	10	-60	8	4.97	18
	16	-60	-4	4.79	19
	18	-70	28	4.78	19
	-16	-76	22	4.73	18
	-32	-48	2	4.64	18/30
	-30	-82	-16	4.53	19
	-38	-66	-4	4.50	19
	-36	-82	-4	4.46	18
	36	-82	-2	4.43	18
	-18	-84	34	4.22	19
	30	-64	18	4.21	18
	54	-72	22	3.97	19

	38	-70	0	3.91	19
	32	-78	-14	3.74	19
	-34	-90	10	3.70	18
	28	-76	34	3.62	7/19
	4	-72	8	3.56	17
	28	-90	-12	3.32	18
	26	-70	0	3.20	19
Others	38	-32	-2	6.00	caudate
	-36	-72	-22	5.26	cerebellum
	26	-32	-2	5.25	hippocampus
	10	-48	-4	5.20	cerebellum
	16	-52	-16	5.16	cerebellum
	38	-24	-12	4.78	hippocampus
	14	-34	0	4.44	thalamus
	-32	-52	-30	4.20	cerebellum
	-18	-58	-24	4.13	cerebellum
	-12	-66	-12	4.10	cerebellum
	18	-34	-28	3.88	cerebellum
	4	-34	10	3.75	callosum
	10	-36	-18	3.32	cerebellum
	-32	-20	-16	3.30	hippocampus
	0	-46	4	3.16	cerebellum

Table S11. Brain regions showing significant accuracy to classify task dimension (face/place) in 80% coherence trials. Coordinates are listed in MNI space. Positive z-values indicate higher accuracy than chance level.

Area	x	y	z	z-value	BA
Parietal cortex	18	-62	30	5.46	31
	26	-52	16	5.40	23
	18	-60	48	5.23	7
	2	-50	18	5.04	23
	-16	-52	22	5.01	23
	-20	-56	34	4.93	7
	14	-50	14	4.92	23
	34	-56	28	4.87	39
	6	-66	36	4.28	7
	12	-76	52	4.22	7
	-28	-66	44	4.20	7/39
	46	-62	24	4.05	39
	14	-66	62	4.04	7
	26	-60	58	3.88	7
	-10	-74	50	3.86	7
	-28	-62	64	3.80	7
	28	-60	38	3.71	7
	-12	-46	34	3.58	23
	26	-74	46	3.53	7
	-8	-70	32	3.46	7
	-28	-78	44	3.41	7
	-50	-74	28	3.30	39
Temporal cortex	34	-50	-16	7.50	37
	-32	-48	-12	7.16	37
	-32	-62	-10	6.60	37
	34	-60	-8	6.51	37
	-44	-38	-10	6.06	37
	20	-36	-6	5.91	36
	22	-24	-12	5.74	36
	-40	-46	0	5.67	21/37
	-32	-26	-20	5.26	36
	26	-36	-22	5.10	37
	42	-48	0	4.92	37
	-44	-24	-16	4.78	20
	60	-54	10	4.33	37
	52	-62	-18	4.20	37
	56	-64	2	4.19	37
	26	-8	-32	3.92	36
Occipital cortex	14	-66	18	6.47	18
	-44	-74	14	5.88	19
	18	-50	0	5.79	18
	16	-56	-12	5.41	19
	-34	-72	2	5.34	18/19
	-22	-52	0	5.33	18
	-26	-80	-14	5.25	19
	38	-92	-2	5.14	18
	34	-76	-2	4.97	18/19
	10	-60	8	4.90	18
	-14	-64	22	4.77	18
	44	-80	4	4.66	19

	-36	-70	28	4.65	19
	20	-66	4	4.61	18
	30	-50	-2	4.55	19
	-30	-92	-14	4.42	18
	24	-72	24	4.36	19
	-30	-84	2	4.34	18
	-42	-84	30	4.25	19
	-26	-70	14	4.21	17
	-16	-78	22	4.16	18
	18	-86	20	3.96	18
	36	-68	14	3.92	19
	16	-68	-10	3.84	18/19
	16	-80	34	3.70	19
	58	-68	22	3.56	19
	-44	-80	-6	3.48	19
	28	-80	8	3.33	18
	-20	-66	-6	3.24	19
	-38	-92	-4	3.20	18
Others	-28	-28	-8	6.70	hippocampus
	36	-40	-28	5.94	cerebellum
	-22	-34	-26	5.66	cerebellum
	14	-28	-24	5.26	brain stem
	-26	-50	-26	5.26	cerebellum
	32	-32	-6	5.06	caudate
	16	-44	-24	4.57	cerebellum
	-26	-14	-22	4.49	hippocampus
	-10	-38	-22	4.23	cerebellum
	-14	-62	-16	4.05	cerebellum
	2	-36	4	3.92	brain stem
	36	-78	-22	3.86	cerebellum
	26	-14	-20	3.82	hippocampus
	-32	-74	-24	3.78	cerebellum
	-14	-12	-12	3.74	brain stem
	2	-50	0	3.69	cerebellum
	-18	-38	4	3.53	hippocampus
	22	-66	-24	3.46	cerebellum