The Architecture of Cognitive Control in the Human Prefrontal Cortex

Etienne Koechlin, Chrystele Ody, Frédérique Kouneiher

Supporting Online Material

Materials and Methods.

Mathematical model. The demand of cognitive control (*1*) required for selecting an action A is assumed to be given by the information (entropy) $h(A)$ associated with action A (*2*):

$$h(A) = -\log_2 p(A),$$

where $p(A)$ is the frequency of selecting action A among the set of alternative actions. Similarly , the control subserved by a signal X and involved in selecting action A is given by the mutual information $I(A,X)$ between X and A:

$$I(A,X) = \log_2 p(A,X) - \log_2 p(A)p(X)$$

where $p(X)$ is the probability that signal X occurs and $p(A,X)$ is the frequency of selecting action A when X occurs. The remaining demand of cognitive control required for selecting action A when signal X occurs is given by the conditional information (entropy) $h(A|X)$:

$$h(A|X) = -\log_2 p(A|X),$$

where $p(A|X)$ is the frequency of selecting action A given X. Then, from Information Theory, if action A depends upon signals X and Y only, we have (*3*):

$$\begin{aligned} h(A) &= I(A,X) + h(A|X) && \text{(E1)} \\ &= I(A,X) + I(A,Y|X) && \text{(E2)} \\ &= I(A,X) + h(Y|X) - h(Y|X, A), && \text{(E3)} \end{aligned}$$

where $I(A,Y|X)$ is the mutual information between A and Y given signal X.

In the cascade model, premotor regions are assumed to exert cognitive control $h(R)$ required for selecting a motor response R. Similarly, caudal LPFC regions are assumed to exert cognitive control $h(T|S)$ required for selecting a task-set T, when a stimulus S occurs. And rostral LPFC regions are assumed to exert additional control $h(T|S,C)$ required for selecting task-set T when stimulus S and contextual signal C occur. Furthermore, sensory control is given by the information $I(R,S)$ conveyed by stimulus S about response R, contextual control by the information $I(T,C|S)$ conveyed by contextual signal C about task-set T given stimulus S, and episodic control by the information $I(T,U|S,C)$ conveyed by episodic signal U about task-set T, given other signals. From equations (E1) (E2)(E3) above, we get:

$$h(T|S,C) = I(T,U|S,C),$$

and

$$h(T|S) = I(T,C|S) + I(T,U|S,C),$$

and

$$h(R) = I(R,S) + I(T,C|S) + I(T,U|S,C) - h(T|S,R),$$

where $h(T\mid S,R)$ simply measures the degree of congruency between task-sets (i.e. the proportion of task-sets including identical stimulus-response associations). Consequently, in the cascade model, the control exerted by rostral LPFC regions varies as episodic control only. The control exerted by caudal LPFC regions varies as the sum of contextual and episodic controls, while the control exerted by premotor regions varies as the sum of sensory, contextual and episodic controls.

Experimental paradigm. In the experimental paradigm, I_{stim} and I_{cont} respectively measure sensory control $I(R,S)$ and contextual control $I(T,C\mid S)$ averaged over each behavioral episode:

$$\begin{aligned} I_{stim} &= \Sigma_{(\mathrm{R,S})}\ p(R,S)\, I(R,S), \\ I_{cont} &= \Sigma_{(\mathrm{T,C,S})}\ p(T,C,S)\, I(T,C\mid S), \end{aligned}$$

where in each episode, $p(R,S)$ is the probability of selecting response R in response to stimulus S, and $p(T,C,S)$ is the probability of selecting task-set T for responding to stimulus S, when contextual signal C occurs. Similarly, I_{cues} measures episodic control $I(T,U\mid S,C)$ for each behavioral episode. For the episode preceded by instruction cue U we have*:*

$$\begin{aligned} I_{cues} &= \Sigma_{(\mathrm{T,C,S})}\ p(T,C,S)\, I(T,U\mid S,C) \\ &= \Sigma_{(\mathrm{T,C,S})}\ p(T,C,S)\, [\log_2 p(\mathrm{U}\mid T,S,C) - \log_2 p(\mathrm{U}\mid S,C)], \\ &= [\log_2 p(\mathrm{U}\mid T,S,C) - \log_2 p(\mathrm{U}\mid S,C)]. \end{aligned}$$

The last equation holds because the term in brackets is constant over each behavioral episode. Then we obtain:

$$I_{cues} = -\log_2 \boldsymbol{f}, \qquad \text{with } \boldsymbol{f} = p(\mathrm{U}\mid S,C)\, [p(\mathrm{U}\mid T,S,C)]^{-1}.$$

In the motor experiment, $\boldsymbol{f}$ was the proportion of episodes including identical stimuli that additionally involved congruent task-sets (i.e. same stimulus-response associations). In the task experiment, $\boldsymbol{f}$ was the proportion of episodes including identical stimuli and contextual signals that additionally involved congruent associations between contextual signals and task-sets (see below).

Subjects. Subjects (6 females and 6 males aged 19 - 29 years) provided written informed consent approved by the French Ethics Committee. Subjects were trained on each behavioral protocol a few days before each experimental session. Order of experiments was counterbalanced across subjects and genders.

Behavioral protocol. In both experiments, subjects responded to successively presented visual stimuli by pressing left or right hand-held response buttons. Each experiment was administered using a 8x8 Latin-square block design consisting of eight series of stimuli (scanning sessions) divided into eight blocks (episodes). A latin-square design was used in order to control for order of presentation of blocks and transitions between blocks. Each block included a series of 12 successive stimuli (Duration: 500 ms; onset asynchrony: 2400 ms) preceded 4800 ms earlier by an instruction cue (arbitrary, distinctive visual signals. Duration: 2000 ms). In each scanning session, the eight blocks formed four distinct experimental conditions crossing the episode factor with either the stimulus (motor experiment) or context (task experiment) factor.

In the motor experiment, stimuli were colored squares. Subjects had to ignore distrator stimuli and responded to other stimuli by pressing the left (L) or right (R) response button (Fig. S1). Subject performance depended upon instruction cues and varied therefore across blocks (instructions associated with each cue were prelearned by subjects):

Block #1: Squares were either green or white. White squares were distractors and subjects had to respond to green squares by pressing the left button (one forced-response episode).

Block #2: Squares were either red or white. White squares were distractors and subjects had to respond to red squares by pressing the right button (one forced-response episode).

Blocks #3 & #4: Squares were either green, red or white. Subjects had to respond to stimuli as in blocks #1 and #2 (two forced-response episodes).

Blocks #5: Squares were either yellow, blue or cyan. Yellow squares were distractors and subjects had to respond to blue and cyan squares by pressing the left button (one forced-response episode).

Blocks #6: Squares were either yellow, blue or cyan. Blue squares were distractors and subjects had to respond to yellow and cyan squares by pressing the right button (one forced-response episode).

Blocks #7 & #8: Squares were either yellow, blue or cyan. Cyan squares were distractors and subjects had to respond to yellow and blue squares by pressing the left and right buttons respectively (two forced-response episode).

Thus, blocks #1, 2, 5, 6 were one forced-response episodes ($\boldsymbol{I_{stim}}$ = 0 bit). Blocks #3, 4, 7, 8 were two forced-response episodes ($\boldsymbol{I_{stim}}$ = 1 bit). In addition, those instructions were chosen in order to parametrically vary the information $\boldsymbol{I_{cues}}$ conveyed by instruction cues across blocks. In blocks #1, 2, 3, 4, information $\boldsymbol{I_{cues}}$ was equal to 0 bit because in the protocol all blocks including the same stimuli (white, green and white squares) involved the same stimulus-response associations (formally, $\boldsymbol{f}$ = 100% and $\boldsymbol{I_{cues}} = -\log_2 \boldsymbol{f}$ = 0 bit, see above). In other words, intruction cues preceding blocks #1, 2, 3, 4 could have been omitted without altering subject performance. In blocks #7 & #8, information $\boldsymbol{I_{cues}}$ was equal to 1 bit, because in the protocol only 50% of blocks including the same stimuli (yellow, blue and cyan squares) involved the same stimulus-response associations ($\boldsymbol{f}$ = 50% and $\boldsymbol{I_{cues}} = -\log_2 \boldsymbol{f}$ = 1 bit). Finally, in blocks #5 & #6, information $\boldsymbol{I_{cues}}$ was equal to 2 bit, because in the protocol only 25% of blocks including the same stimuli (yellow, blue and cyan squares) involved the same stimulus-response associations ($\boldsymbol{f}$ = 25% and $\boldsymbol{I_{cues}} = -\log_2 \boldsymbol{f}$ = 2 bit).

In each block, sequences of stimuli were pseudorandomized so that the proportion of distractors was equal to 33%. In two forced-responses blocks, the ratio of left vs. right responses was equal to 1. The proportions of two successive trials including identical stimuli were also maintained constant across one forced-response blocks and across two forced-response blocks.

In the task experiment, stimuli were letters and contextual signals were colors of letters. According to contextual signals, subjects had to ignore stimuli or performed either a vowel/consonant (T1) or a lower/upper case (T2) discrimination task on letters (using the left and right response button) (Fig. S1). Again, subject performance depended upon instruction cues and varied therefore across blocks (instructions associated with each cue were prelearned by subjects):

Block #1: contextual signals were either green or white. White signals indicated subjects to ignore the letter. Green signals indicated subjects to perform task T1 (single task-set episode).

Block #2: contextual signals were either red or white. White signals indicated subjects to ignore the letter. Red signals indicated subjects to perform task T2 (single task-set episode).

Blocks #3 & #4: Contextual signals were either green, red or white. Subjects had to respond to letters as in blocks #1 and #2 (dual task-set episode).

Blocks #5: Contextual signals were either yellow, blue or cyan. Yellow signals indicated subjects to ignore letters. Blue and cyan signals indicated subjects to perform task T1 (single task-set episode).

Blocks #6: Contextual signals were either yellow, blue or cyan. Blue signals indicated subjects to ignore letters. Yellow and cyan signals indicated subjects to perform task T2 (single task-set episode).

Blocks #7 & #8: Contextual signals were either yellow, blue or cyan. Cyan signals indicated subjects to ignore letters. Yellow and Blue signals indicated subjects to perform tasks T1 and T2 respectively (dual task-set episode).

Thus, blocks #1, 2, 5, 6 were single task-set episodes (I_{cont} = 0 bit). Blocks #3, 4, 7, 8 were dual-task set episodes (I_{cont} = 1 bit). Again, those instructions were chosen in order to parametrically vary the information I_{cues} conveyed by instruction cues across blocks. In blocks #1, 2, 3, 4, information I_{cues} was equal to 0 bit because in the protocol all blocks including the same contextual signals (white, green and white colors) involved the same associations between contextual signals and task-sets T1 & T2 (f = 100% and $I_{cues} = -\log_2 f$ = 0 bit). In blocks #7 & #8, information I_{cues} was equal to 1 bit, because in the protocol only 50% of blocks including the same contextual signals (yellow, blue and cyan colors) involved the same associations between contextual signals and task-sets (f = 50% and $I_{cues} = -\log_2 f$ = 1 bit). Finally, in blocks #5 & #6, information I_{cues} was equal to 2 bit, because in the protocol only 25% of blocks including the same contextual signals (yellow, blue and cyan) involved the same associations between contextual signals and task-sets (f = 25% and $I_{cues} = -\log_2 f$ = 2 bit).

In each block, sequences of contextual signals were pseudorandomized so that the proportion of letters to be ignored was equal to 33%. In dual- task-set blocks, the ratio of trials associated with task-set T1 vs. task-set T2 was equal to 1. The proportions of two successive trials including identical contextual signals were maintained constant across single task-set blocks and across dual task-set blocks. Letters were pseudorandomly chosen from the set {A, E, I, O, a, e, i, o, C, G, K, P, c,g, k, p} so that in each block the ratio of left vs. right responses and the ratio of congruent vs. uncongruent letters (same vs. different responses for T1 and T2) were equal to 1.

Data acquisition. A 3 T Brucker whole-body and RF coil scanner were used to perform a structural scan for each subject followed by 8 series of 116 functional axial scans (TR 2.4s, TE 40 ms, flip angle 90 deg, FOV 24 cm, acquisition matrix 64x64, number of slices 18, thickness 6 mm). fMRI data were processed using the SPM99 software package (http://www.fil.ion.ucl.ac.uk/spm/) with standard image realignment, linear normalization to the stereotaxic Talairach atlas (MNI template)(*4*), spatial (3D Gaussian kernel: 10 mm) and temporal smoothing (Gaussian kernel: 4000 ms). The behavioral protocol was administered using the EXPE6 software package (*5*).

Computations of brain activations. Statistical parametric maps were computed from local fMRI signals using a linear multiple regression analysis with conditions (modeled as box-car functions convolved by the canonical hemodynamic response function) and scanning series as covariates (*6*). Brain regions exhibiting significant contrasts of parameter estimates across conditions were first identified using a fixed-effect model (voxel-wise threshold: Z = 4.3, $p < 0.05$ corrected for multiple comparisons; cluster-vise threshold: 576 mm^3, $p < 0.05$). Then, in order to account for between-subjects variability and to allow statistical inferences at the population level, regional activations were further confirmed using a random-effect model (voxel-wise threshold: $p < 0.05$, cluster-vise threshold: $p < 0.05$, corrected for multiple comparisons over the search volumes). The stimulus effect was cpmuted as larger activations in the two-forced than one-forced response episodes with I_{cues} = 0, the context effect as larger activations in the dual- than single- task-sets episodes with I_{cues} = 0, and the episode effect as activations that in both experiments parametrically varied as the episode factor I_{cues}.

Analyses of covariance. Mean activations in the premotor, caudal and rostral LPFC regions were separately entered in univariate repeated-measure ANCOVAs with hemisphere (left vs. right), experiment (motor vs. task), number of alternatives (one- vs. two- forced response or single- vs. dual- task-sets) as within-subject factors and episode (I_{cues} = 0, 1 or 2) as within-subjects covariates. Then, in each experiment, the same ANCOVAs were performed excluding the factor of experiment.

Effective connectivity. Structural equation modeling was processed using the MX software package (*7*). Subject-specific times series of mean regional activations identified in previous group analyses were averaged together and standardized in each condition. The resulting time series were then used for model estimation and statistical inference based on Maximum-Likehood statistics. The overall model fit was assessed by computing standard goodness-of-fit indexes including Normed Fit, Centrality and Relative Non-Centrality Indexes (all indexes > 0.9 indicating an appropriate fit) (*8*), Significant variations of path coefficients were assessed using a nested model approach (*8*). Coefficient variations related to the episode, context and stimulus factors were computed from variations in interregional correlation matrices observed between all episodes with $\boldsymbol{I_{cues}}$ = 0 vs. $\boldsymbol{I_{cues}}$ >0, single- vs. dual-task-sets episodes, and one- vs. two-forced response episodes respectively.

Subsequent analyses were performed to control that the significant variations of path coefficients computed in previous analyses did not result from a few outliers among subjects. For that purpose, the structural equation model was fitted and variations of path coefficients were computed for each subject separately using subject-specific times series of mean regional activations as described above. Then, Wilcoxon Signed Rank tests performed on subject-specific variations of path coefficients confirmed the results described in the main text. Path coefficients increasing with the episode factor were confirmed to connect bilateral rostral LPFC regions to premotor cortex through the right hemisphere (all Zs > 1.65, ps < 0.05, one-tailed). Path coefficients increasing with the context factor were confirmed to connect bilateral caudal LPFC regions to premotor cortex through the left hemisphere (all Zs > 2.04, ps < 0.021, one-tailed). And coefficients increasing with the stimulus factor were found on paths connecting left and right caudal LPFC regions and connecting left and right premotor regions (both Zs > 2.35, ps < 0.01, one-tailed).

Discussion

Our findings explain the pattern of prefrontal activations observed in several experimental paradigms including learning, episodic memory, working memory and task-switching paradigms. For instance, in learning paradigms, rostral LPFC activations were reported when subjects were learning action sequences by trials and errors, that is when subjects received feedback signals altering behaviors in subsequent episodes. In episodic memory paradigms, rostral LPFC activations were especially observed in retrieval phases, when subjects selected actions based on the occurrence of previous events. In working memory paradigms, caudal LPFC activations were reported, when subjects maintained task-sets in working memory, whereas rostral LPFC activations were observed when subjects had to select actions based on memorized information. In task-switching paradigms, caudal LPFC activations were observed when subjects switched between task-sets with respect to concomitant visual signals, whereas rostral LPFC activations were observed when subjects prepared task-sets for subsequent behaviors or when task-switching required additional control over several trials following switch signals.

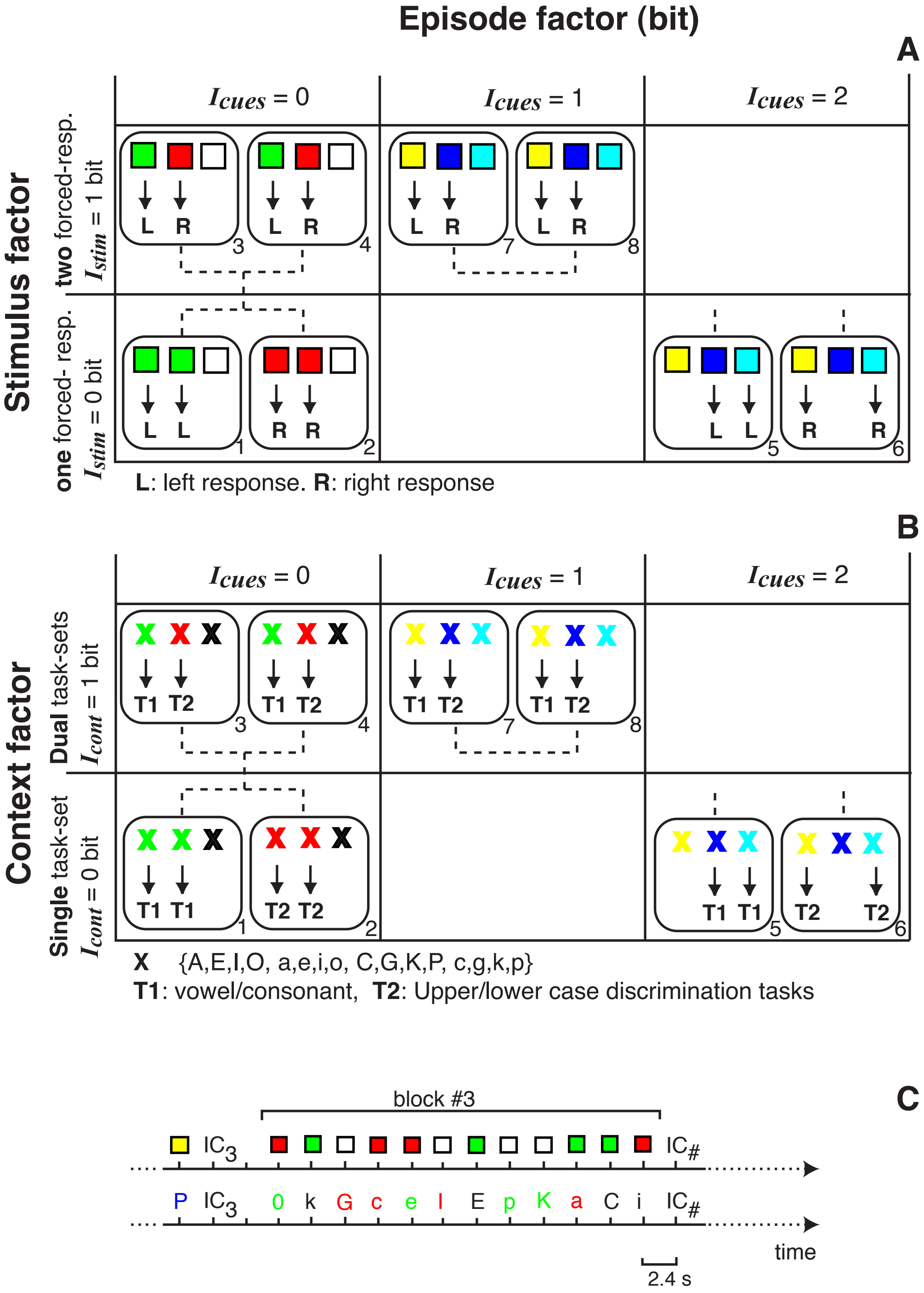

Fig. S1. Experimental design. Rounded boxes represent behavioral episodes (numbered from #1 to #8 as in materials and methods) with related stimuli and instructions. Episodes formed four distinct experimental conditions crossing the episode factor with either the stimulus (motor experiment, A) or context (task experiment, B) factor. In the motor experiment, as shown in rounded boxes, stimuli were colored squares. Subjects ignored distrator stimuli (no arrow) or responded by pressing the left (L) or right (R) response button. Dashed lines connect episodes involving congruent stimulus-response associations. In the task experiment, stimuli were letters (represented by the symbol X) and contextual signals were colors of letters. According to contextual signals, subjects ignored letters (no arrow) or performed either a vowel/consonant (T1) or a lower/upper case (T2) discrimination task on letters. Dashed lines connect episodes involving congruent associations between contextual signals and task-sets. C, typical examples of episodes (# 3) in the motor and task experiments. IC: instruction cues.

FIGURE S1

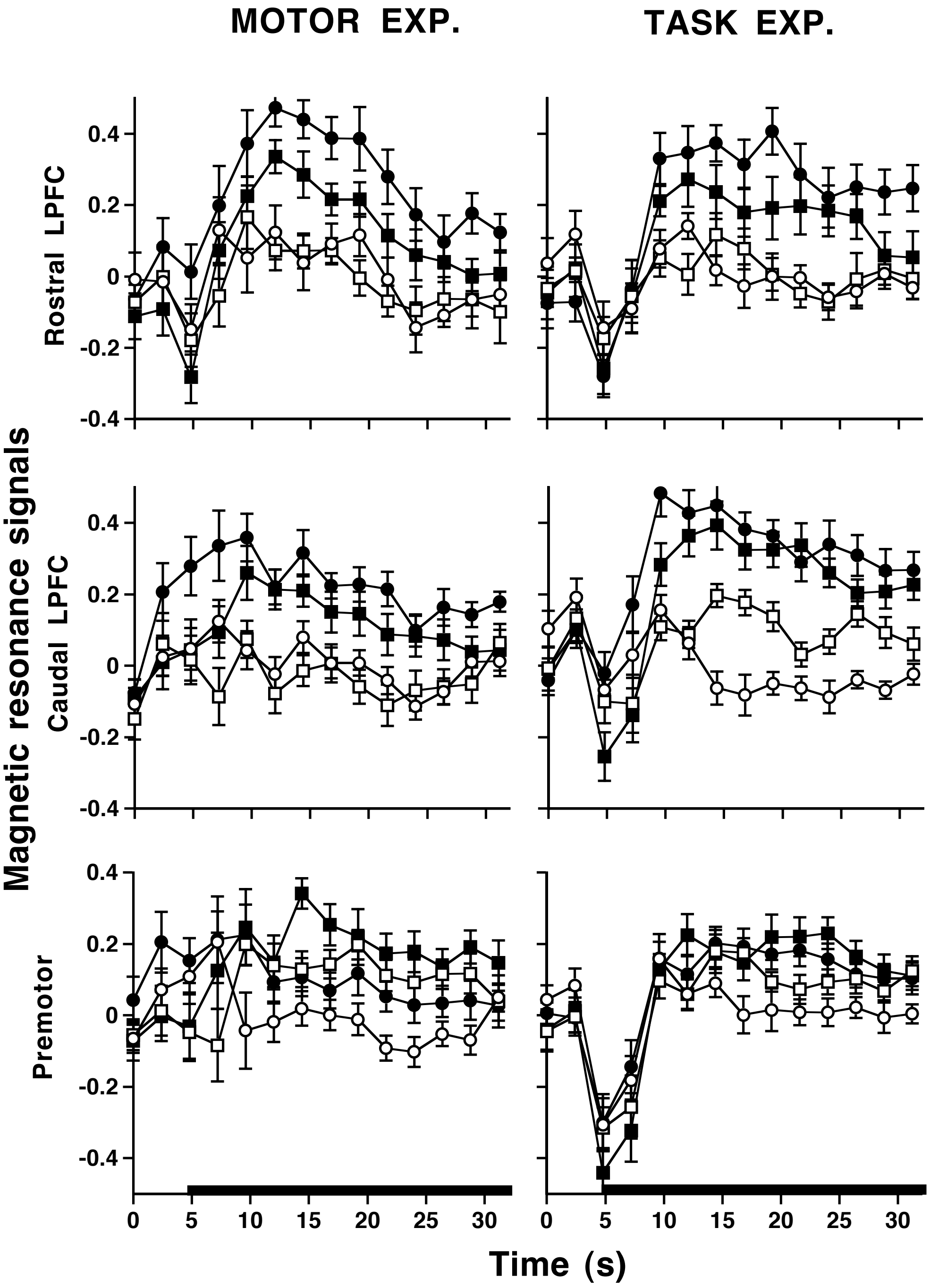

Fig. S2. Dynamic of frontal activations. Graphs show time-courses of magnetic resonance signals recorded during behavioral episodes in rostral LPFC (top), caudal LPFC (middle) and premotor regions (bottom) and averaged over both hemispheres and subjects (error bars indicate standard errors across subjects). Left, motor experiment. Circles: one forced-response conditions. Squares: two forced-response conditions. Right, task experiment. Circles: single task-set conditions. Squares: dual task-set conditions. In both experiments, open circles and squares are for conditions related to Icues = 0 bit; filled squares and circles for conditions related to Icues = 1 and 2 bit respectively. Data points are magnetic resonance signals relative to the average magnetic resonance signals recorded in baseline conditions (open circles, Y-axis origin). X-axes represent the time elapsed from onsets of instruction cues. Stimulus presentation started 4.8 s later as indicated by thick lines. Onset asynchrony of instruction cues was equal to 33.6 s.

Table S1. Brain activations outside the frontal lobes**.**

Brain regions	**Volumes (cm^3)**	**Talairach coordinates**	**Statistical effects** (Z-values, fixed-effect model)		
			Episode	*Context*	*Stimulus*
Stimulus effect					
L Inferior Parietal Lobule, SmG, BA40	2.8	-36, –40, 36	**4.6**	1.5	**5.7**
L Thalamus	1.4	-16, -16, 0	1.3	**4.6**	**5.1**
Context effect (excluding stimulus effect)					
R Precuneus BA 7	2.7	24, -60, 40	**18.8**	**6.0**	0.9
L Inferior Parietal Lobule, AnG, BA39	4.5	-32, -68, 32	**12.6**	**7.3**	1.3
R Putamen	0.7	24, 24, -4	**13.0**	**5.0**	-1.1
R Thalamus	5.5	8, -8, 8	**4.1**	**6.0**	1.3
Episode effect (excluding previous effects)					
R Inferior Parietal Lobule, SmG, BA40	5.3	36, -40, 30	**11.6**	1	0.2
L Inferior Parietal Lobule, SmG, BA40	1.3	-52, -48, 36	**9.3**	1.5	1.6
Precuneus, BA7	1.2	0, -56, 40	**10**	1.3	1.3
L mid. Temporal Gyrus, BA21/37	0.8	-56, -48, -4	**6.4**	-0.2	0.2

SmG: supramarginal gyrus. AnG: angular gyrus. L: left. R: right. BA: Brodman's area.
Talairach coordinates are those of maximal Z-scores.

TABLE S1

References

1. D. E. Berlyne, *Psychological Review* **64**, 329-339 (1957).
2. C. E. Shannon, *Bell System Technical Journal* **27**, 379-423, 623-656 (1948).
3. G. Deco, D. Obradovic, *An information-theoretic approach to neural computing* (Springer-Verlag, New-York, 1996).
4. J. Talairach, P. Tournoux, *Co-planar stereoaxic atlas of the human brain* (Thieme Medical Publishers, New York, 1988).
5. C. Pallier, E. Dupoux, X. Jeannin, *Behavior Research: Methods, Instruments, & Computers* **29**, 322-327 (1997).
6. K. J. Friston, C. D. Frith, P. F. Liddle, R. S. J. Frackowiak, *Journal of Cerebral Blood Flow and Metabolism* **11**, 690-699 (1991).
7. M. C. Neale, S. M. Boker, G. Xies, H. H. Maes, "Mx: Statistical modeling" *Tech. Report No. 6th Edition* (Department of Psychiatry, Richmond, VA, 2002).
8. R. O. Mueller, *Basic principles of structural equation modeling*, Springer texts in statistics (Springer-Verlag, New-York, 1996).