nature human behaviour

Article

https://doi.org/10.1038/s41562-024-01826-7

Orthogonal neural encoding of targets and distractors supports multivariate cognitive control

Received: 12 December 2022

Accepted: 15 January 2024

Published online: 08 March 2024

Check for updates

Harrison Ritz [1,2,3] & Amitai Shenhav [1,2]

The complex challenges of our mental life require us to coordinate multiple forms of neural information processing. Recent behavioural studies have found that people can coordinate multiple forms of attention, but the underlying neural control process remains obscure. We hypothesized that the brain implements multivariate control by independently monitoring feature-specific difficulty and independently prioritizing feature-specific processing. During functional MRI, participants performed a parametric conflict task that separately tags target and distractor processing. Consistent with feature-specific monitoring, univariate analyses revealed spatially segregated encoding of target and distractor difficulty in the dorsal anterior cingulate cortex. Consistent with feature-specific attentional priority, our encoding geometry analysis revealed overlapping but orthogonal representations of target and distractor coherence in the intraparietal sulcus. Coherence representations were mediated by control demands and aligned with both performance and frontoparietal activity, consistent with top-down attention. Together, these findings provide evidence for the neural geometry necessary to coordinate multivariate cognitive control.

We have remarkable flexibility in how we think and act. This flexibility is enabled by the array of mental tools we can bring to bear on challenges to our pursuit of goals[1–6]. For example, someone may respond to a mistake by becoming more cautious, enhancing task-relevant processing or suppressing task-irrelevant processing[7], and previous work has shown that people simultaneously deploy multiple such strategies in response to different task demands[3,8–10]. Flexibly coordinating multiple cognitive processes requires a control system that can monitor multiple forms of task demands and deploy multiple forms of control (also referred to as the necessity for observability and controllability[11]). These monitoring and regulation processes are fundamental to control and are thought to be underpinned by distinct cingulo-opercular and frontoparietal neural systems[12–19]. However, much is still unknown about how multiple forms of control are represented across these domains.

Past research on the neural mechanisms of cognitive control has often sought to identify representations that integrate over multiple different sources of task demands (that is, represent these different sources in alignment). For instance, previous studies have proposed that the dorsal anterior cingulate cortex (dACC) tracks integrative features such as response conflict, effort, value, error likelihood and time on task[20–27]. Because they integrate over different task features instead of differentiating between them, these forms of 'aligned encoding' (Fig. 1a) are ill suited for carrying out multidimensional control. Multidimensional cognitive control requires independent representations that can track multiple sources of difficulty and regulate multiple cognitive processes (for example, prioritizing multiple sources of information[28]).

An alternative to aligned encoding—one that would allow the brain to separately control multiple processes—is independent encoding,

[1]Cognitive, Linguistic & Psychological Science, Brown University, Providence, RI, USA. [2]Carney Institute for Brain Science, Brown University, Providence, RI, USA. [3]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA. e-mail: hritz@princeton.edu

https://doi.org/10.1038/s41562-024-01826-7

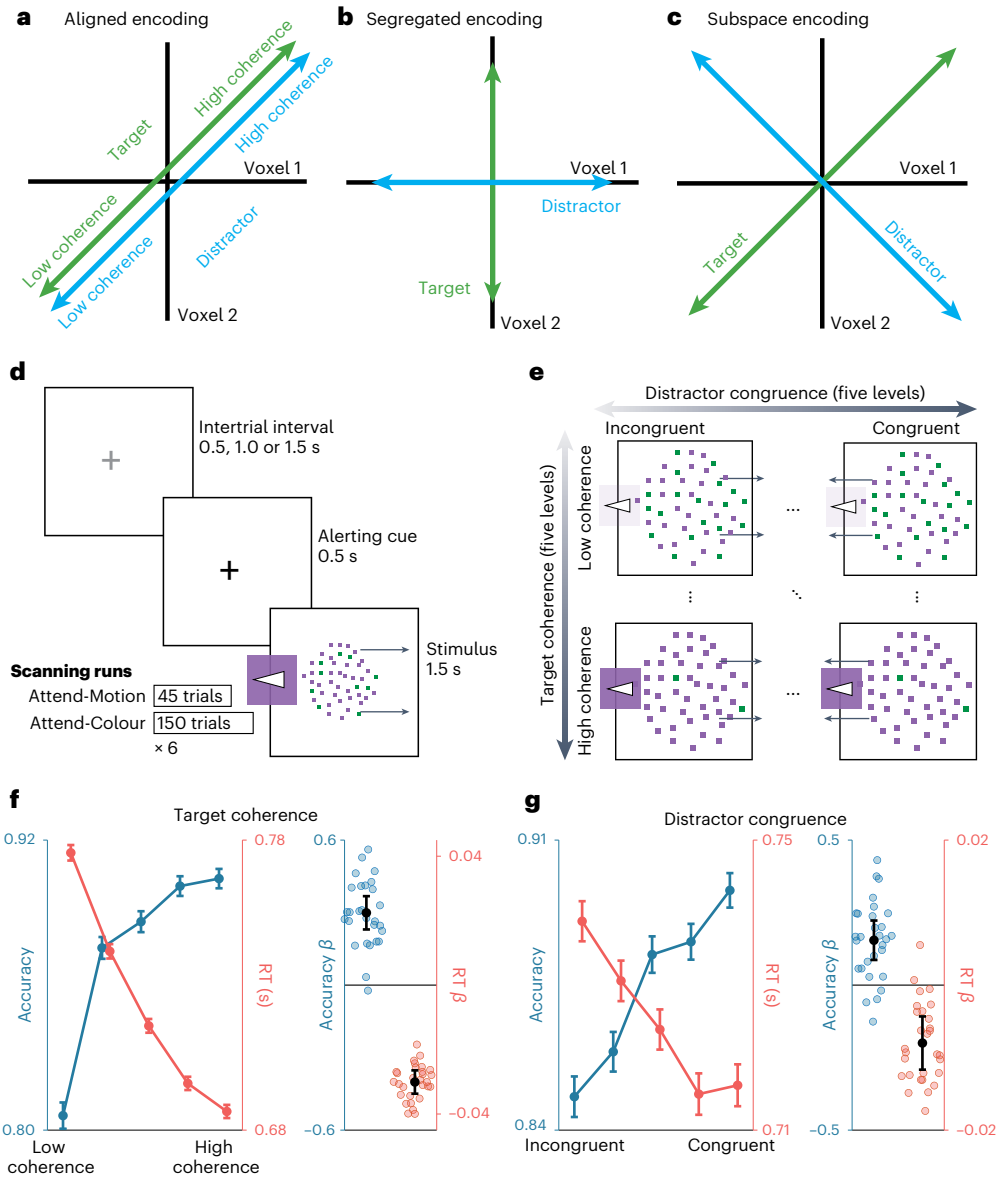


Fig. 1 | Task and behaviour. a–**c**, Three hypothesized encoding schemes. In aligned encoding, features are represented similarly—for example, when encoding performance variables such as error likelihood or time on task (**a**). In segregated encoding, features are encoded independently, in distinct voxel populations (that is, voxel-level pure selectivity[40]) (**b**). In subspace encoding, features are encoded independently, in overlapping voxel populations (that is, voxel-level mixed selectivity) (**c**). **d**, Participants responded to a colour–motion RDK with a button press. The participants responded either to the left/right motion direction of the RDK (Attend-Motion runs) or on the basis of the majority colour (Attend-Colour runs; critical condition). **e**, We parametrically and independently manipulated target coherence (the percentage of dots in the majority colour) and distractor congruence (motion coherence signed relative to the target response). **f**, Participants were faster and more accurate when the target was more coherent. **g**, Participants were faster and more accurate when the distractor was more congruent with the target. The error bars on the line plots reflect the mean and within-participant s.e.m.; the error bars for regression fixed-effect β values reflect 95% CIs ($N = 29$ for all figures).

which can come in at least two forms. One way the brain can have independent representations is by encoding different task features in spatially segregated neural populations ('segregated encoding'; Fig. 1b). For example, past work has shown that different subregions in the dACC encode distinct task demands, including various forms of errors and processing conflict[29–34]. The brain can instead have independent representations that are distributed across units within the same population, as has also been observed in the dACC[35–37]. Within a shared population, independent encoding of information occurs along a set of orthogonal dimensions or subspaces (Fig. 1c, 'subspace encoding'[38–41]). Despite this exciting recent work, it remains unclear to what extent different components of the cognitive control system leverage these aligned, segregated or orthogonal encoding strategies for monitoring multiple task demands and prioritizing multiple sources of information.

To gain new insight into the representations supporting cognitive control, we drew upon two key innovations. First, we leveraged an experimental paradigm we developed to tag multiple control processes[10]. Building on prior work[3,30,41,42], this task incorporates elements of perceptual decision-making (discrimination of a target feature) and inhibitory control (overcoming a salient and prepotent distractor). We have previously shown that we can separately tag target and distractor processing from participants' performance on this task and that target and distractor processing are independently controlled. For example, participants adjust target and distractor sensitivity in response to distinct task demands (for example, previous conflict or

incentives[10]). In conjunction with this process-tagging approach, our second innovation was to develop a multivariate functional MRI (fMRI) analysis for measuring relationships between feature encoding (that is, encoding geometry). Extending recent statistical approaches in systems neuroscience[35,43,44], we combined the strengths of multivariate encoding analyses and representation similarity analyses into a method we call 'encoding geometry analysis' (EGA). We used EGA to characterize whether putative markers of monitoring and prioritization leverage independent representations for targets and distractors.

In brief, we found that key nodes in the cognitive control network use orthogonal representations of target and distractor information to support cognitive control. In the dACC, encoding of target and distractor difficulty was spatially segregated and arranged along a rostrocaudal gradient. By contrast, in the intraparietal sulcus (IPS), encoding of target and distractor coherence was arranged along orthogonal neural subspaces. These regional distinctions are consistent with hypothesized roles in planning and implementing (multivariate) attentional policies[12,17]. Furthermore, we found that coherence encoding depended on control demands and was aligned with both task performance and frontoparietal activity, consistent with these coherence representations playing a critical role in cognitive control (for example, feature prioritization). Together, these results suggest that cognitive control uses representational formats that allow the brain to monitor and control multiple streams of information processing.

Results

Task overview

Twenty-nine human participants performed the Parametric Attentional Control Task[10] during fMRI. On each trial, the participants responded to an array of coloured moving dots (coloured random dot kinematogram (RDK); Fig. 1d). In the critical condition (Attend-Colour), the participants responded with a left/right keypress depending on which of two colours were in the majority. In alternating scanner runs, the participants instead responded on the basis of motion (Attend-Motion), which was designed to be less control-demanding due to the (Simon-like) congruence between motion direction and response hand[3,10]. Across trials, we independently and parametrically manipulated target and distractor information across five levels of target coherence (for example, the percentage of dots in the majority colour, regardless of which colour) and distractor congruence (for example, the percentage of dots moving either in the congruent or incongruent direction relative to the correct colour response; Fig. 1e). This task allowed us to 'tag' participants' sensitivity to each dimension by measuring behavioural and neural responses to independently manipulated target and distractor features. Unlike a similar task used to study post-error adjustments[3], our parametric manipulation of target and distractor coherence allows us to better measure feature-specific representations. Unlike similar tasks used to study contextual decision-making[30,41,45], this task pits more control-demanding responses (towards colour) against more automatic responses (towards motion), allowing comparisons between Attend-Colour and Attend-Motion tasks to isolate the contributions of cognitive control[46,47].

Performance depends on targets and distractors

The participants had overall good performance on the task, with a high level of accuracy (median accuracy, 89%; interquartile range, 84–92%) and a low rate of missed responses (median lapse rate, 2%; interquartile range, 0–5%). We used mixed-effects regressions to characterize how target coherence and distractor congruence influenced participants' accuracy and log-transformed correct reaction times (RTs). Replicating previous behavioural findings using this task, the participants were sensitive to both target and distractor information[10]. When target coherence was weaker, the participants responded slower ($t_{27.6} = 16.1$; $P < 0.001$; Cohen's $d = 3.01$; 95% confidence interval (CI), (0.0248, 0.0310)) and less accurately ($t_{28} = -8.90$; $P < 0.001$; $d = -1.65$; 95% CI, (−0.365, −0.233); Fig. 1f). When distractors were more incongruent, the participants also responded slower ($t_{28.8} = 5.09$; $P < 0.001$; $d = 0.942$; 95% CI, (0.00603, 0.0141)) and less accurately ($t_{28} = -4.66$; $P < 0.001$; $d = -0.865$; 95% CI, (−0.220, 0.0896); Fig. 1g). Further replicating prior findings with this task, interactions between targets and distractors were not significant for RT ($t_{28.2} = 0.143$; $P = 0.887$; $d = 0.0265$; 95% CI, (−0.00181 0.00208)) and had a weak influence on accuracy ($t_{28} = 2.36$; $P = 0.0257$; $d = 0.437$; 95% CI, (0.00581, 0.0634)). Models omitting target–distractor interactions provided a better complexity-penalized fit (RT ΔAIC = 17.7, accuracy ΔAIC = 1.38).

Segregated encoding of target and distractor difficulties

Past work has separately shown that the dACC tracks task demands related to perceptual discrimination (induced in our task when target information is weaker) and related to the need to suppress a salient distractor (induced in our task when distractor information is more strongly incongruent with the target[12,30–32,48]). Our task allowed us to test whether these two sources of increasing control demand are tracked within common regions of the dACC (reflecting an aggregated representation of multiple sources of task demands) or whether they are tracked by separate regions (potentially reflecting a specialized representation according to the nature of the demands).

Targeting a large region of the dACC—a conjunction of a cortical parcellation with a meta-analytic mask for 'cognitive control' (see 'fMRI univariate analyses' in Methods)—we found spatially distinct signatures of target difficulty and distractor congruence within the dACC. In caudal dACC, we found significant clusters encoding the parametric effect of target difficulty (Fig. 2a; the negative effect of target coherence is shown in green), and in more rostral dACC we found clusters encoding parametric distractor incongruence (the negative effect of distractor congruence is shown in blue). Supporting this dissociation, the spatial patterns of target and distractor regression weights were uncorrelated across dACC voxels ($t_{28.0} = 1.32$; $P = 0.197$; log Bayes factor (log(BF)) = −0.363; 95% CI, (−0.111, 0.515)). These analyses control for omission errors, and additionally controlling for commission errors produced the same whole-brain pattern at a reduced threshold (Extended Data Fig. 1). We also found that the most rostral portion of our dACC mask responded to target ease (Extended Data Fig. 2).

To further quantify how feature encoding changed along the longitudinal axis of the dACC, we used principal component analysis (PCA) to extract the axis positions of dACC voxels (see 'dACC longitudinal axis analyses' in Methods) and then regressed target and distractor β weights onto these axis scores. We found that targets had stronger difficulty coding in more caudal voxels ($t_{27.9} = 3.40$; $P = 0.00204$; $d = 0.631$; 95% CI, (10.6, 42.8)), with a quadratic trend ($t_{26.5} = 4.48$; $P < 0.001$; $d = 0.85$; 95% CI, (38.2, 103); Fig. 2b). In line with previous work on both perceptual and value-based decision-making[30,49–52], we found that signatures of target discrimination difficulty (negative correlation with target coherence) in caudal dACC were paralleled by signals of target discrimination ease (positive correlation with target coherence) within the rostral-most extent of our dACC region of interest (ROI) (Extended Data Fig. 3). In contrast to targets, distractors had stronger incongruence coding in more rostral voxels ($t_{28.0} = -2.87$; $P = 0.00781$; $d = -0.533$; 95% CI, (−55.6, −9.25)), without a significant quadratic trend. We used participants' random-effects terms to estimate the gradient location where target and distractor coding were at their most negative, finding that the target minimum was significantly more caudal than the distractor minimum (signed-rank test, $z_{28} = 2.41$, $P = 0.0159$). Target and distractor minima were uncorrelated across participants ($r_{27} = 0.0282$, $P = 0.880$, log(BF) = −0.839), again consistent with independent encoding of targets and distractors.

As additional evidence that target-related and distractor-related demands have a dissociable encoding profile, we found that the crossover between target and distractor encoding in the dACC occurred at the boundary between two well-characterized functional networks[53–55].

https://doi.org/10.1038/s41562-024-01826-7

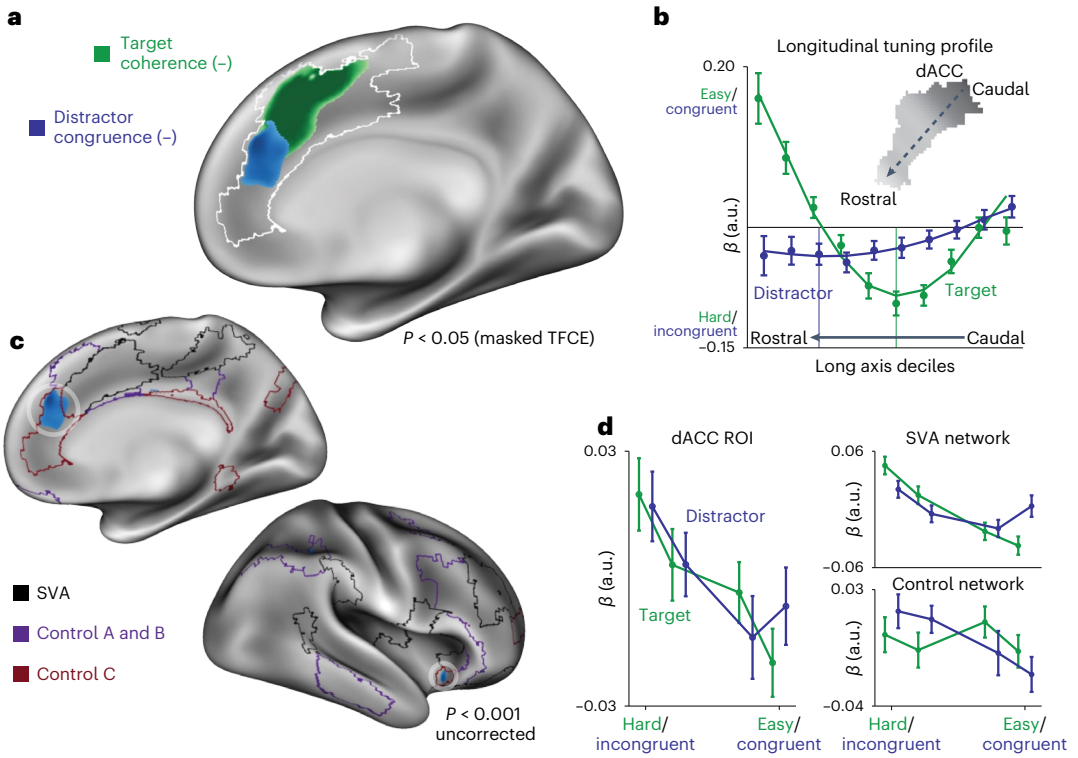


Fig. 2 | Distinct coding of target and distractor difficulty in the dACC. **a**, We looked for linear target coherence and distractor congruence signals within an a priori dACC mask (white outline; overlapping Kong22 parcels and medial 'cognitive control' Neurosynth mask). We found that voxels in the most caudal dACC reflected target difficulty (green), and more rostral voxels reflected distractor incongruence (blue). Note that these maps show only difficulty-coded effects; the most rostral portion of the dACC responded to target ease (Extended Data Fig. 2). The shading reflects two-tailed *t*-statistics, corrected for multiple comparisons using non-parametric threshold-free cluster enhancement (TFCE). **b**, We extracted the long axis of the dACC using a PCA of the voxel coordinates. We plotted the target coherence (green) and distractor congruence (blue) along the deciles of this long axis. The fit lines are the quantized predictions from a second-order polynomial regression. We used these regression β values to estimate the minima for target and distractor tuning (that is, the locations of the strongest difficulty effects), finding that the target difficulty peak (the vertical green line) was more caudal than the distractor incongruence peak (the vertical blue line). **c**, Plotting the uncorrected whole-brain response, the distractor incongruence responses (blue) were strongest in the Control C subnetwork (red), in both the dACC and anterior insula. **d**, Blood-oxygen-level-dependent responses across levels of target coherence and distractor congruence, plotted within the whole dACC ROI (left), or the SVA network and Control network parcels within the dACC ROI (right). GLMs: **a**–**c**, Feature UV; **d**, Difficulty Levels (Table 1). Throughout, the error bars reflect the mean and within-participant s.e.m. ($N = 29$).

Whereas distractor-related demands were more strongly encoded rostrally in the Control network (particularly in regions of the dACC and insula corresponding to the Control C subnetwork[54,56]), target-related demands were more strongly encoded caudally in the Salience / Ventral Attention (SVA) network (Fig. 2c,d). Including network membership alongside long axis location predicted target and distractor encoding better than models with either network membership or axis location alone (ΔBIC > 1,675).

Independent encoding of target and distractor coherence

We found that the dACC appeared to dissociably encode target and distractor difficulty through spatially segregated encoding, consistent with a role in monitoring different task demands and/or specifying different control signals[12]. To identify neural mechanisms for the implementation of this control through the prioritization of targets versus distractors, we next tested for regions that encode target and distractor coherence (the amount of information in a feature, regardless of which response it supports). On the basis of previous research, we might expect to find this form of selective attention in the posterior parietal cortex[17,57,58]. We explored whether target and distractor coherence share a common neural code (for example, as a global index of spatial salience), or whether these features are encoded distinctly (for example, as separate targets of control).

An initial whole-brain univariate analysis showed that overlapping regions throughout the occipital, parietal and prefrontal cortices track the feature coherence (the proportion of dots in the majority category) for both targets and distractors (Fig. 3a; conjunction in yellow). These regions showed elevated responses to lower target coherence and higher distractor coherence, potentially reflecting the relevance of each feature for task performance. Note that in contrast to distractor congruence, distractor coherence had an inconsistent relationship with task performance (RT: $t_{27.0} = 2.08$; $P = 0.0468$; $d = 0.394$; 95% CI, $(8.33 \times 10^{-5}, 0.0107)$; accuracy: $t_{28} = -0.845$; $P = 0.406$; $d = -0.157$; 95% CI, $(-0.085, 0.0338)$), suggesting that these neural responses are unlikely to reflect task difficulty per se.

While these univariate activations point towards widespread and coarsely overlapping encoding of the feature coherence (potentially consistent with aligned encoding; Fig. 1a), they lack information about how these features are encoded at finer spatial scales. To interrogate the relationship between target and distractor encoding, we developed a multivariate analysis that combines multivariate encoding analyses with pattern similarity analyses, which we term EGA. Whereas pattern similarity analyses typically quantify relationships between representations of specific stimuli or responses (for example, whether they could be classified[59]), EGA characterizes relationships between encoding subspaces (patterns of contrast weights) across different task features, consistent with recent analysis trends in systems neuroscience[35,36,43,60–62]. A stronger correlation between encoding subspaces (either positive or negative) indicates that features are similarly encoded (that is, that their representations are aligned and

thus confusable by a linear decoder; Fig. 1a), whereas weak correlations indicate that these representations are orthogonal (and thus distinguishable by a linear decoder[59]). In contrast to standard pattern similarity, the sign of these relationships is interpretable in EGA, reflecting how features are coded relative to one another. Relative to standard encoding analysis, simulations revealed that EGA maintains sensitivity under high levels of noise (Extended Data Fig. 3). We estimated this encoding alignment within each parcel, correlating unsmoothed and spatially pre-whitened patterns of parametric regression β values across scanner runs to minimize spatiotemporal autocorrelation[63–65]. This cross-validated similarity further allowed us to anchor our analysis on the measurement reliability of encoding profiles (that is, the self-correlation of encoding patterns across cross-validation folds[66,67]).

Focusing on regions that encoded both target and distractor information (parcels where both group-level $P < 0.001$), EGA revealed clear dissociations between regions that represent these features in alignment versus orthogonally. Within the mid-level visual cortex and the superior parietal lobule (SPL), target and distractor representations demonstrated significant negative correlations (Fig. 3b, red), reflecting (negatively) aligned encoding. In contrast, the early visual cortex and IPS (see Fig. 3c for the anatomical boundaries) demonstrated target–distractor correlations near zero (Fig. 3b, black), suggesting encoding along orthogonal subspaces.

To bolster our interpretation of the latter findings as reflecting orthogonal (that is, uncorrelated) representations rather than merely small but non-significant correlations, we employed Bayesian *t*-tests at the group level to estimate the relative ($\log_{10}$) likelihood that these encoding dimensions were orthogonal or correlated. Consistent with our previous analyses, we found strong evidence for correlation (positive log(BF)) in more medial regions of the occipital and posterior parietal cortex (for example, the SPL) and strong evidence for orthogonality (negative log(BF)) in more lateral regions of the occipital and posterior parietal cortex (for example, the IPS; Fig. 3d). Control analyses confirmed that coherence orthogonality was not due to encoding reliability, as a similar topography was observed with disattenuated correlations (normalizing correlations by their reliability; Supplementary Fig. 1). Further supporting these results, our BF analyses were robust to the choice of priors (Supplementary Fig. 2).

While our analyses support independent encoding of targets and distractors within the same parcel, we further explored whether feature information is reflected in overlapping voxels (that is, voxel-level mixed selectivity[40]). Simulations revealed that the alignment between absolute encoding weights can differentiate between pure and mixed selectivity, and parietal coherence representations bore this signature of voxel-level mixed selectivity (Extended Data Fig. 4), consistent with the subspace encoding hypothesis.

These results have focused on the coherence of different features regardless of the response they support, demonstrating that the SPL exhibits aligned representations of target and distractor coherence. Past decision-making research has separately demonstrated that the SPL tracks the amount of evidence supporting specific responses[42,68,69], which we found was also true for our task. In addition to encoding target and distractor coherence, the SPL and visual cortex tracked target and distractor 'evidence' (the proportion of dots supporting a rightward versus leftward response; Fig. 3e). EGA revealed orthogonal evidence representations between targets and distractors, in the same areas with aligned coherence representations (compare Fig. 3d,e), consistent with previous observations of multiple decision-related signals in the SPL[68].

We complemented our whole-brain analyses with ROI analyses in areas exhibiting reliable encoding of key variables, focusing on core frontal regions linked with cognitive control (the dACC and lateral prefrontal cortex (lPFC)) and parietal regions linked with decision-making and attention (the SPL and IPS[12,15]). Consistent with our analyses above, we found that target and distractor coherence encoding was aligned in the SPL but not in the IPS (Fig. 4a, compare with Fig. 3d), whereas the SPL encoded target and distractor evidence. Directly comparing these regions (Supplementary Table 1), we found stronger encoding of target evidence in the SPL, stronger encoding of target coherence in the IPS and stronger target–distractor coherence alignment in the SPL. Unlike our univariate results, we did not find distractor congruence encoding in the dACC (though this was found in the lPFC and IPS). Instead, the dACC showed multivariate encoding of target coherence and evidence.

To further characterize how feature coherence and evidence are encoded across these regions, we performed multidimensional scaling over each region's task representations (Fig. 4b and refs. 64,70). Briefly, this method allows us to visualize—in a non-parametric manner—the relationships between representations of different feature levels (for example, levels of target coherence), by estimating each feature level separately within a general linear model (GLM) and then using singular value decomposition to project these patterns into a 2D space (see Methods for additional details). We found that coherence and evidence axes naturally emerge in the top two principal components in this analysis within the dACC, SPL and IPS. Coherence axes (light to dark shading) are parallel between left (blue) and right (brown) responses, suggesting response-independent encoding. In these components, evidence encoding appeared to be binary, in contrast to parametric coherence encoding (we found similar whole-brain encoding maps for binary-coded evidence; Supplementary Fig. 3). Critically, whereas coherence encoding axes in the SPL were aligned between targets (circles) and distractors (diamonds; confirming aligned encoding), in the IPS these representations formed perpendicular lines (confirming orthogonal encoding). When we visualized higher dimensions, we found that the IPS did appear to have weak encoding alignment between target and distractor coherence in higher dimensions (Extended Data Fig. 5). Nevertheless, the orthogonal encoding in the first two principal components is sufficient for a downstream region to have an independent read-out of feature-specific coherence. These analyses both help visualize cross-region dissociations in encoding profiles and validate the idea that task features are encoded in a monotonic fashion.

Finally, to explore the divisions between the SVA and Control networks evident in the univariate analyses, we split up our two prefrontal ROIs by their network membership (Extended Data Fig. 6). In the dACC, we found that SVA parcels tended to have stronger feature encoding than Control parcels. Interestingly, in these SVA parcels, several features were aligned with the target evidence dimension, consistent with recent human electrophysiology findings[35]. In the lPFC, we found that Control parcels, but not SVA parcels, encoded distractor congruence (Control: $t_{28} = 3.60$; two-tailed $P = 0.0012$; log(BF) = 1.45; 95% CI, (0.0037, 0.0135); SVA: $t_{28} = 0.57$; $P = 0.57$; log(BF) = −0.64; 95% CI, (−0.0046, 0.0082); Control − SVA: $t_{28} = 3.27$; $P = 0.0029$; log(BF) = 1.12; 95% CI, (0.0025, 0.0111)). This distractor congruence encoding was present in the lPFC in Control A/B parcels ($t_{28} = 3.66$; $P = 0.001$; log(BF) = 1.51; 95% CI, (0.0041, 0.0146)), but not significantly in Control C parcels ($t_{28} = 1.86$; $P = 0.073$; log(BF) = −0.0448; 95% CI, (−0.0006, 0.0136)). This network-selective encoding of congruence is consistent with the univariate results in the dACC (Fig. 2).

Control demands dissociate coherence and evidence encoding

Our findings thus far demonstrate two sets of dissociations within and across brain regions. In the dACC, we found that distinct regions encode the control demands related to discriminating targets (caudal dACC) versus overcoming distractor incongruence (rostral dACC). In the posterior parietal cortex, we found that overlapping regions track the coherence of these two stimulus features but that distinct regions represent these features in alignment (SPL) versus orthogonally (IPS). While these findings suggest that this set of regions is involved in translating between feature information and goal-directed responding, they only focus on the information that was presented to the participant on a given trial. To provide a more direct link between feature-specific

https://doi.org/10.1038/s41562-024-01826-7

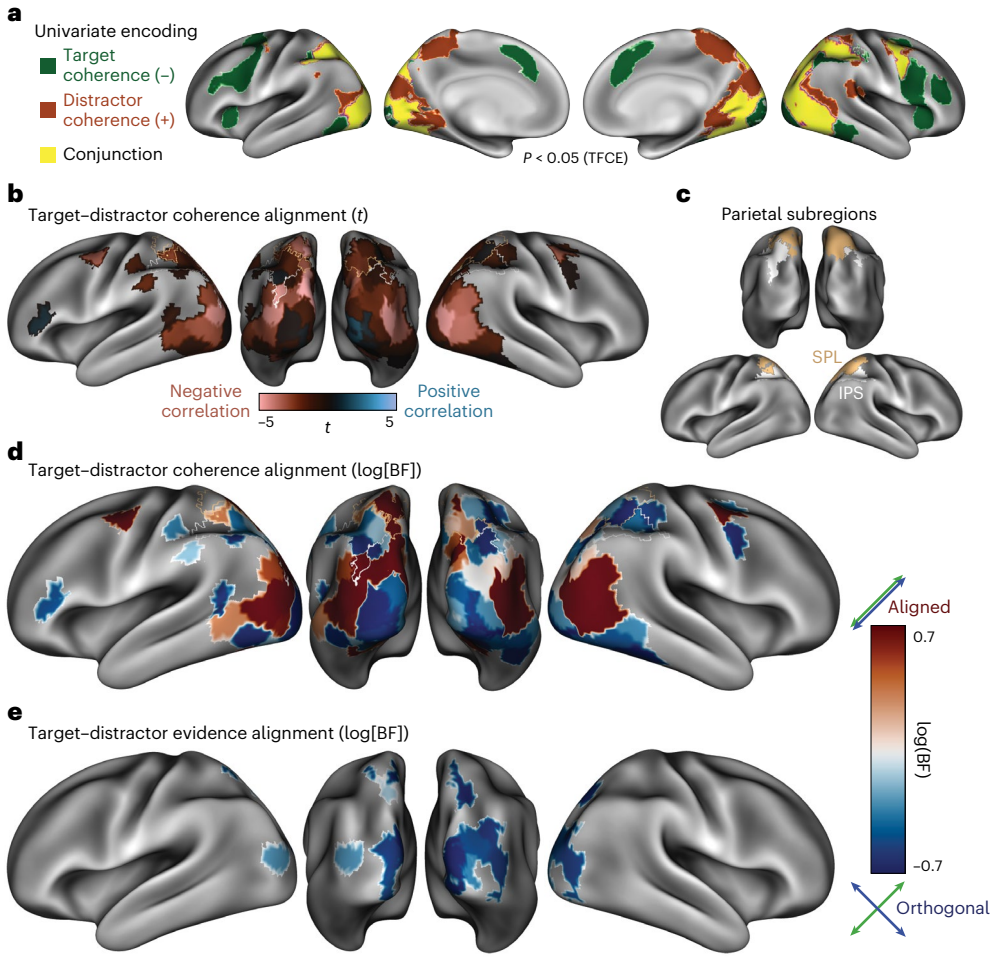


Fig. 3 | EGA dissociates target and distractor encoding. a, Parametric univariate responses to weak target coherence (green; the percentage of dots in the majority colour), strong distractor coherence (orange; percentage of dots with coherent motion) and their conjunction (yellow). Statistical tests (two-tailed *t*-tests) are corrected for multiple comparisons using non-parametric TFCE. **b**, Alignment between target and distractor coherence (two-tailed *t*-test on correlation values), within parcels where both were jointly reliable (two-tailed $P < 0.001$, uncorrelated). Representations were negatively correlated in the SPL (in gold; Kong22 labels) and uncorrelated in the IPS (in white; Kong22 labels). **c**, Anatomical labels for parietal regions, based on the labels in the Kong22 parcellation. **d**, Bayesian analyses provide explicit evidence for orthogonality in the IPS (that is, negative BF; theoretical minima, −0.71). **e**, Coherence coded in terms of evidence (that is, supporting a left versus right choice). Target and distractor evidence encoding overlapped in the visual cortex and SPL and was represented orthogonally. GLMs: **a**, Feature UV; **b**–**e**, Feature MV (Table 1).

encoding and control, we examined how the encoding of feature coherence differed between matched tasks that placed stronger or weaker demands on cognitive control. So far, our analyses have focused on conditions in which the participants needed to respond to the colour feature while ignoring the motion feature (Attend-Colour task), but on alternating scanner runs the participants instead responded to the motion dimension and ignored the colour dimension (Attend-Motion task). These tasks were matched in their visual properties (identical stimuli) and motor outputs (left/right responses) but critically differed in their control demands. Attend-Motion was designed to be much easier than Attend-Colour, as the left/right motion directions are compatible with the left/right response directions (that is, Simon facilitation[3,10]). Comparing these tasks allows us to disambiguate bottom-up attentional salience from the top-down contributions to attentional priority[47,71–73].

Consistent with previous work[10], performance on the Attend-Motion task was better overall (mean RT, 565 ms versus 725 ms; sign-rank $P < 0.001$; mean accuracy, 93.7% versus 87.5%; sign-rank $P < 0.001$). Unlike the Attend-Colour task, performance was not impaired by distractor incongruence (that is, colour distractors; RT: $t_{28} = -1.39$; $P = 0.176$; $d = -0.0438$; 95% CI, (−0.00629, 0.000577); accuracy: $t_{28} = 0.674$; $P = 0.506$; $d = 0.0847$; 95% CI, (−0.0913, 0.147)). To investigate these task-dependent feature representations, we fit a GLM that included both tasks. To control for performance differences across tasks, we analysed only accurate trials and included trial-wise RT as a nuisance covariate, concatenating RT across tasks.

Whereas the encoding of both colour and motion coherence was widespread during the Attend-Colour task (Fig. 3), coherence encoding was consistently weaker during the less demanding Attend-Motion task (Fig. 5a). Coherence encoding was weaker during Attend-Motion whether classifying according to goal relevance (comparing targets or distractors) or the features themselves (comparing motion or colour). Task-relevant ROIs revealed that coherence encoding was effectively absent during the easy Attend-Motion task (Fig. 5b), consistent with coherence encoding in these regions depending on the control demands of the Attend-Color task[47,74].

In contrast to these stark task-related differences in coherence encoding, we found that neural encoding of the target evidence (colour evidence in the Attend-Colour task and motion evidence in the Attend-Motion task) was preserved across tasks, including within the dACC, lPFC, SPL and IPS (Fig. 5b). Consistent with previous experiments examining context-dependent decision-making[36,41,42,45,73,75,76], we found stronger target evidence encoding relative to distractor evidence encoding, in our case in the evidence-encoding SPL (Attend-Colour:

https://doi.org/10.1038/s41562-024-01826-7

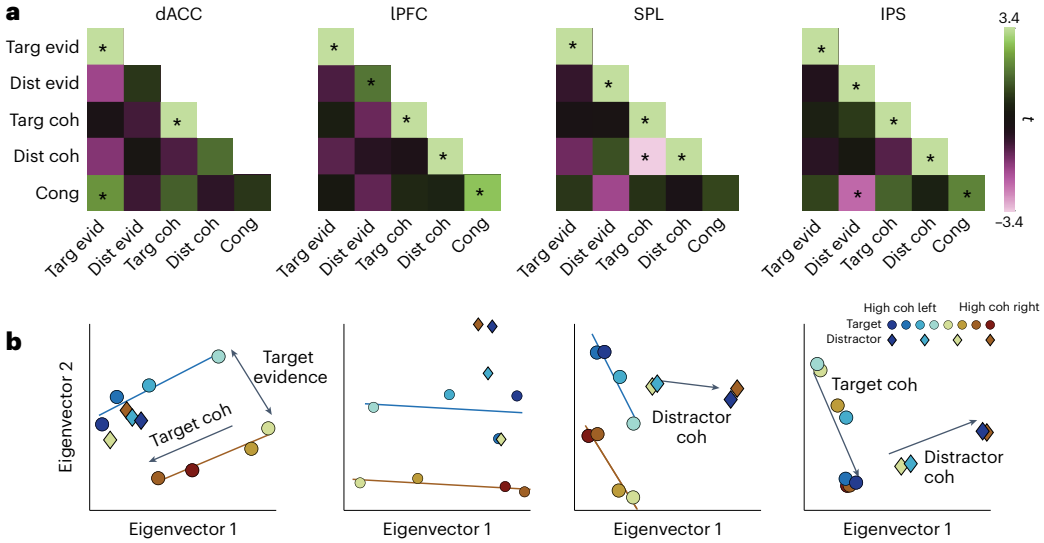


Fig. 4 | Region-specific feature encoding. a, Similarity matrices for the dACC, lPFC, SPL and IPS, correlating feature evidence (evid), feature coherence (coh) and feature congruence (cong). Encoding strength is shown on the diagonal (right-tailed *t*-test); encoding alignment is on the off-diagonal (two-tailed *t*-test). Asterisks indicate $p < .05$ (uncorrected). Targ, target; dist, distractor. **b**, Classical multidimensional scaling embedding of target (circles) and distractor (diamonds) representations at different levels of evidence. The colours denote responses; hues denote coherence. GLMs: **a**, Feature MV; **b**, Evidence Levels (Table 1).

$t_{28} = 4.26$, right-tailed $P < 0.001$, $d = 0.790$; Attend-Motion: $t_{28} = 2.37$, right-tailed $P = 0.0124$, $d = 0.4403$). We also found that target evidence encoding during Attend-Motion was aligned with that during Attend-Colour, both for motion evidence encoding ('stimulus axis'; SPL: $t_{28} = 2.08$; right-tailed $P = 0.0236$; $d = 0.386$; 95% CI, (0.0009, 0.0095); IPS: $t_{28} = 2.24$; right-tailed $P = 0.0167$; $d = 0.416$; 95% CI, (0.0016, 0.0114)) and target evidence encoding ('decision axis'; SPL: $t_{28} = 5.87$; right-tailed $P < 0.001$; $d = 1.09$; 95% CI, (0.0109, 0.0199); IPS: $t_{28} = 3.64$; right-tailed $P = 0.0011$; $d = 0.676$; 95% CI, (0.0056, 0.0154)). These axis alignments are again in agreement with previous experiments, though note that target evidence is often manipulated separately from the motor response. Whereas our experiment replicates previous observations of the neural representations supporting contextual decision-making, we now extended these findings to understand how putative attention signals (that is, feature coherence) are encoded in response to the asymmetric inference that is characteristic of cognitive control[77].

Feature coherence aligns with task performance

Feature coherence encoding (that is, feature strength, regardless of response or congruence) depends on task demands, consistent with a role in cognitive control. To further understand this relationship between coherence encoding and control, we next explored how coherence encoding was related to task performance. We tested this question by determining whether feature coherence representations were aligned with performance representations (that is, alignment between stimulus and behavioural subspaces[78]). Specifically, we included trial-level RT and accuracy in our first-level GLMs. Encoding of performance was itself highly robust: most parcels encoded RT and accuracy, with the strongest encoding in cognitive control regions (Extended Data Fig. 7). Across the cortex, RT and accuracy were negatively correlated, again most prominently across the cognitive control network. To explore the behavioural relevance of coherence representations, we tested whether coherence encoding was aligned with the voxel patterns encoding task performance.

We found that the encoding of target and distractor coherence was aligned with performance across frontoparietal and visual regions (Fig. 6a–b). If a region's encoding of target coherence reflects how sensitive the participant was to target information on that trial (for example, due to top-down priority), we would expect target encoding to be positively aligned with performance on a given trial, such that stronger target coherence encoding is associated with better performance and weaker target coherence encoding is associated with poorer performance. We would also expect distractor encoding to demonstrate the opposite pattern—stronger encoding associated with poorer performance and weaker encoding associated with better performance. We found evidence for both patterns of feature–performance alignment across the visual and frontoparietal cortex: target encoding was aligned with better performance (faster RTs and higher accuracy; Fig. 6a), whereas distractor encoding was aligned with worse performance (slower RTs and lower accuracy; Fig. 6b).

Next, we examined whether performance–coherence alignment reflected individual differences in participants' task performance in our main task-related ROIs (Figs. 3 and 4). In particular, we tested whether the alignment between features and behaviour reflects specific relationships with speed or accuracy, or whether it reflects overall increases in evidence accumulation (for example, faster responding and higher accuracy). Within each ROI, we correlated feature–RT alignment with feature–accuracy alignment across participants. We found that in the dACC and IPS, participants showed the negative correlation between performance alignment measures predicted by an increase in processing speed (Fig. 6c). People with stronger alignment between target coherence and shorter RTs tended to have stronger alignment between target coherence and higher accuracy, with the opposite found for distractors. While these between-participant correlations were present within targets and distractors, we did not find any significant correlations across features (between-feature: all $P > 0.10$), again consistent with feature-specific processing. These analyses were qualitatively similar after partialing out the reliability of coherence and performance encoding (Supplementary Table 2). While between-participant analyses using small sample sizes warrant a note of caution, these findings are consistent across features and regions. In conjunction with our within-participant evidence that feature coherence representations are aligned with performance efficiency, these findings support a role for coherence encoding in adaptive control.

Feature coherence aligns with frontoparietal activity

Across the frontal, parietal and visual cortex, encoding of target and distractor coherence depended on task demands and was aligned with

https://doi.org/10.1038/s41562-024-01826-7

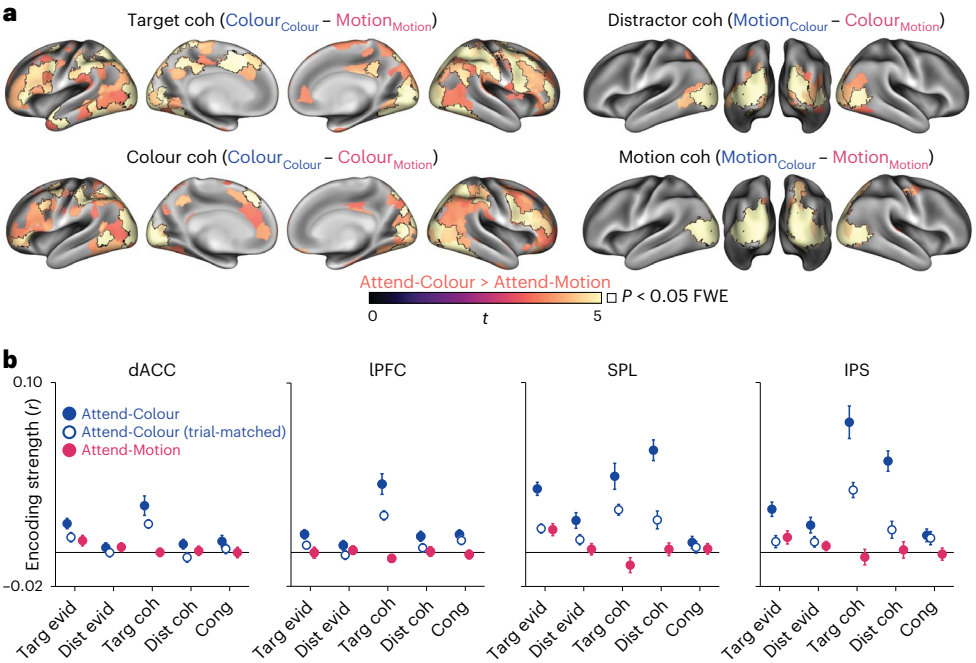


Fig. 5 | Task-dependent encoding strength. a, Across the cortex, feature coherence encoding was stronger during Attend-Colour than Attend-Motion, matched for the same number of trials. Attend-Colour had stronger encoding when comparing target coherence (top left), distractor coherence (top right), colour coherence (bottom left) and motion coherence (bottom right). The parcels are thresholded at $P < 0.001$ (two-tailed, uncorrected); the outlined parcels are significant at $P < 0.05$ (family-wise error (FWE) corrected using two-tailed max-statistic randomization tests). The condition labels in parentheses are coded 'Feature$_{Task}$'. **b**, Target and distractor coherence information was encoded more strongly during Attend-Colour than Attend-Motion in the dACC, lPFC, SPL and IPS. Attend-Colour encoding is plotted from the whole sample (blue fill) and a trial-matched sample (the first 45 trials of each run; white fill). In Attend-Motion runs, only target evidence was significantly encoded (magenta). The error bars reflect the mean and within-participant s.e.m. ($N = 29$). GLM: Between-Task (Table 1).

performance. Since this widespread encoding of task information probably reflects distributed network involvement in cognitive control[77,79,80], we sought to understand how frontal and parietal systems interact. We focused our analyses on the IPS and lPFC, linking the core parietal site of orthogonal coherence encoding (IPS) to a prefrontal site that previous work suggests provides top-down feedback during cognitive control[58,79,81,82]. Previous work has found that the IPS attentional biases lower-level stimulus encoding in visual cortices[83,84], and that the IPS mediates directed connectivity between the lPFC and visual cortex during perceptual decision-making[42]. Here we extended these experiments to test how the IPS mediates the relationship between prefrontal feedback and stimulus encoding.

To investigate these putative cortical interactions, we developed a multivariate connectivity analysis to test whether coherence encoding was aligned with prefrontal activity and whether this lPFC–coherence alignment was mediated by the IPS. We first estimated the voxel-averaged residual time series in the lPFC (SPM12's eigenvariate) and then included this residual time series alongside task predictors in a whole-brain regression analysis (Extended Data Fig. 8). This analysis can be schematized as:

$$\beta_{seed} = \mathrm{GLM}\left(Y_{seed}, X\right) \quad (1)$$

$$e_{seed} = \mathrm{PCA}\left(Y_{seed} - X\beta_{seed}\right) \quad (2)$$

$$\beta_{all} = \mathrm{GLM}\left(Y_{all}, [X, e_{seed}]\right) \quad (3)$$

The GLM function performs regression on multivariate voxel time series Y using design matrix X, and the PCA function extracts the first principal component of the residuals. Finally, we used EGA to test whether there was alignment between patterns encoding lPFC functional connectivity (that is, β values from the residual time series predictor e_{seed}) and patterns encoding target and distractor coherence. Note that these analyses depend on functional connectivity, a correlational measure that can be subject to confounding[85].

We found that lPFC connectivity patterns were aligned with coherence-encoding patterns in the visual cortex (Fig. 7a). Stronger prefrontal functional connectivity was aligned with weaker target coherence and stronger distractor coherence, consistent with prefrontal recruitment during difficult trials. Notably, IPS connectivity was also aligned with target and distractor coherence in overlapping parcels, even when controlling for lPFC connectivity. These effects were liberally thresholded for visualization, as significant direct and indirect effects are not necessary for significant mediation[86].

Our critical test was whether the IPS mediated the relationship between lPFC activity and coherence encoding. We compared regression estimates between a model that only included lPFC residuals ('Solo' model) and a model that included both lPFC and IPS residuals ('Both' model). Comparing the strength of lPFC–coherence alignment with and without the IPS is a test of whether the parietal cortex mediates lPFC–coherence alignment[86]. These models can be schematized as:

$$\beta_{Solo} = \mathrm{GLM}\left(Y_{all}, [X, e_{lPFC}]\right) \quad (4)$$

$$\beta_{Both} = \mathrm{GLM}\left(Y_{all}, [X, e_{lFPC}, e_{IPS}]\right) \quad (5)$$

We found that this mediation was strongest in early visual cortex, where the alignment between the lPFC and feature coherence was reduced in a model that included the IPS relative to a model without the IPS (Fig. 7b). The negatively correlated target–lPFC relationship

https://doi.org/10.1038/s41562-024-01826-7

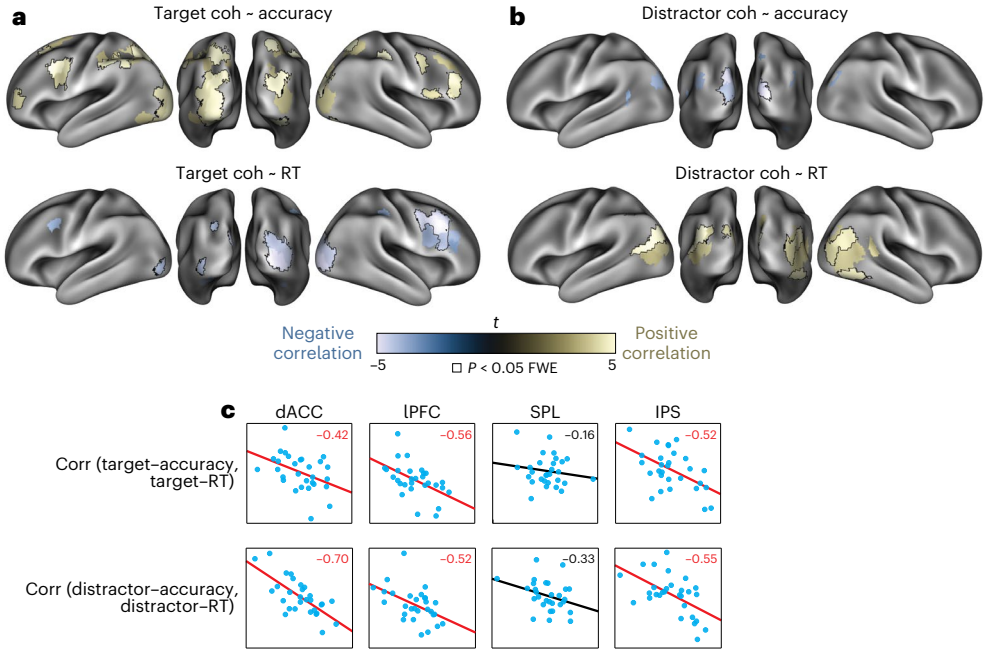


Fig. 6 | Alignment between feature and performance encoding. a, Alignment between encoding of target coherence and performance (accuracy is shown in the top row and RT in the bottom row). **b**, Alignment between encoding of distractor coherence and performance (accuracy is shown in the top row and RT in the bottom row). Across **a**,**b**, the parcels are thresholded at $P < 0.001$ (two-tailed uncorrected t-test, in jointly reliable parcels at $P < 0.001$). The outlined parcels are significant at $P < 0.05$ (two-tailed max-statistic randomization test across jointly reliable parcels). **c**, Individual differences in feature–RT alignment correlated with feature–accuracy alignment across regions (Pearson correlation values are shown in top right; $P < 0.05$ in red). See Supplementary Table 2 for partial correlations controlling for reliability. GLM: Performance (Table 1).

became more positive when the IPS was included (top), and the positively correlated distractor–PFC relationship became more negative when the IPS was included (bottom). Critically, we found that the IPS reduced prefrontal–coherence alignment in early visual cortex more than the lPFC reduced parietal–coherence alignment (Fig. 7b inset and Supplementary Fig. 4a,b), consistent with frontal-to-parietal directed connectivity in previous research[42,81]. Looking within colour- and motion-sensitive parcels (determined using task-free localizer runs; Methods), we found that this mediation was significant in colour-sensitive cortex. The opposite relationship, lPFC mediation of IPS connectivity, appeared in higher-level visual cortex for distractor coherence (Supplementary Fig. 4c,d), though these effects were not reliable in explicit contrasts and may reflect projections from both regions. Note that we did not see any significant mediation of first-order target or distractor coherence encoding by the IPS.

We were primarily interested in alignment with the lPFC due to previous work implicating these regions in top-down control (for reviews, see refs. 12,87), but for completeness we also examined how different subnetworks in both the lPFC and dACC aligned with coherence encoding. In the lPFC, we found that the SVA and Control subnetworks had similar patterns of alignment (Supplementary Fig. 5). In the dACC, we found that the SVA subnetwork had a qualitatively similar profile of coherence alignment as the lPFC, but this alignment was absent in the Control subnetwork. Whereas this seed–coherence alignment was similar across the lPFC and the SVA dACC, unlike the lPFC, we found that the SVA dACC failed to demonstrate strong evidence for mediation by the IPS (Supplementary Fig. 6).

A final set of analyses examined whether the SPL and IPS demonstrated different patterns of task-related functional connectivity with other regions, given that we found that these regions differentially encoded evidence and coherence. When seeding our connectivity analyses with SPL activity, we found that SPL activity aligned with evidence encoding in the bilateral motor cortex (Extended Data Fig. 9). In contrast, IPS activity did not significantly align with evidence encoding, and this seed–evidence alignment in the motor cortex was stronger for the SPL than for the IPS, consistent with a putative role for the SPL in response selection[68].

Discussion

In this experiment, we explored whether neural control systems use representations with the same dimensionality as the processes they regulate[2,5,11]. Inspired by behavioural evidence that participants can independently control their sensitivity to targets and distractors[10], we set out to understand whether the neural correlates of monitoring and prioritization leverage independent encoding for feature-selective control (Fig. 1a–c). We found that key nodes of canonical cognitive control networks had orthogonal neural representations of targets and distractors. Within the dACC, orthogonal representations of target and distractor difficulty arose from segregated encoding along a rostrocaudal axis. Within the IPS, orthogonal representations of target and distractor coherence arose from orthogonal subspaces in overlapping voxels. Consistent with a role in attentional priority, coherence representations depended on control demands, task performance and frontoparietal activity. Together, these results reveal a neural mechanism for how cognitive control prioritizes multiple streams of information during decision-making.

Neurocomputational theories have proposed that the dACC is involved in planning control across multiple levels of abstraction[12,88–90]. Past work has found that control abstraction is hierarchically organized along the dACC's rostrocaudal axis, with more caudal dACC involved in lower-level action control and more rostral dACC involved in higher-level strategy control[30–32,34], an organization that may reflect a more general hierarchy of abstraction within the PFC[31,91–93]. Consistent with this account, we found that caudal dACC tracked the coherence of the target and distractor dimensions, especially within the SVA network. In contrast, more rostral dACC tracked incongruence between targets and distractors, especially within the Control network. Speculatively, our results are consistent with caudal dACC

https://doi.org/10.1038/s41562-024-01826-7

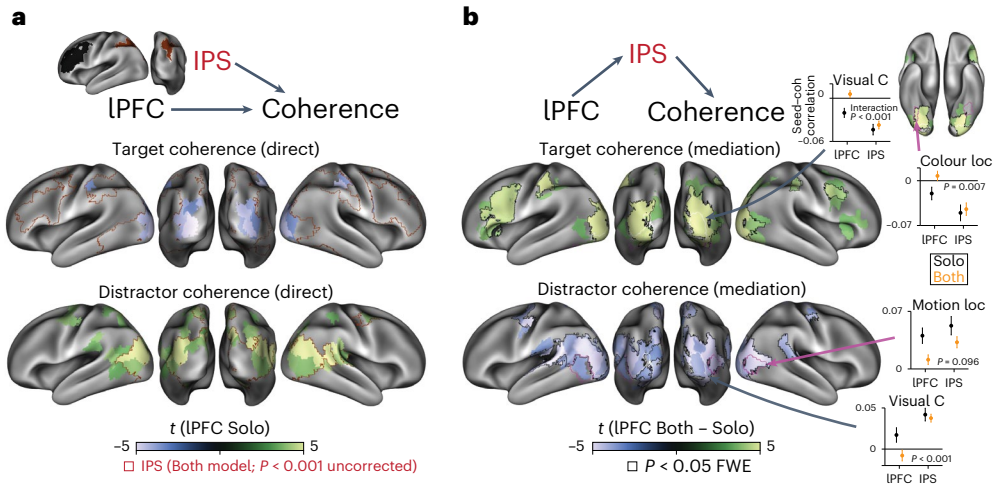


Fig. 7 | The IPS mediates alignment between the lPFC and feature encoding. **a**, Connectivity patterns from the lPFC (colour) and IPS (red outline) were aligned with target and distractor coherence patterns (two-tailed $P < 0.001$ uncorrected, in jointly reliable parcels at $P < 0.001$). IPS effects are outlined to show overlap, with all effects in a consistent direction to the lPFC. **b**, lPFC–feature alignment contrasted between the lPFC-only model ('Solo') and the lPFC + IPS model ('Both'). Including the IPS reduced the alignment between the lPFC and feature encoding (compare the sign of the main effect in **a** with the contrast in **b**). The parcels are thresholded at two-tailed $P < 0.001$ (uncorrected, jointly reliable parcels), and the outlined parcels are significant at $P < 0.05$ (two-tailed max-statistic randomization test across jointly reliable parcels). Insets: seed–coherence alignment in the Solo models (black) and the Both model (orange) across visual regions. 'Visual C' is defined by our parcellation[54]; Colour and Motion localizers (loc) are parcels near the peak response identified during feature localizer runs (Methods). In general, lPFC alignment was more affected by the IPS than IPS alignment was affected by the lPFC. Throughout, the error bars reflect the mean and within-participant s.e.m. ($N = 29$). GLM: Performance CX (Table 1).

tracking the first-order difficulty arising from the relative salience of feature-specific information, and more rostral dACC tracking the second-order difficulty arising from cross-feature (in)compatibility[92], the latter of which may require additional disengagement from distractor-dependent attentional capture.

Whereas the dACC encoded feature difficulty (for example, distractor incongruence), in the parietal cortex we found overlapping representations of feature coherence (for example, distractor coherence). In the SPL, features had correlated coherence encoding (similarly representing low target coherence and high distractor coherence), consistent with this region's transient and non-selective role in attentional control[94–99]. In contrast, the IPS had orthogonal representations of feature coherence, consistent with selective prioritization of task-relevant information[47,71–73,81,83,94–96,99,100]. While the IPS primarily encoded features orthogonally (that is, in the largest components of our multidimensional scaling analysis), the total coherence across features could also be read out at higher dimensions. The ability of the IPS to communicate both orthogonal and aligned coherence representations is consistent with the diverse roles of the IPS in attentional control.

Our previous work has demonstrated behavioural evidence for independent control over target and distractor attentional priority in this task[10], with different task variables selectively enhancing target or distractor sensitivity (see also refs. 4,101). Orthogonal feature representation in the IPS may offer a mechanism for this feature-selective control, consistent with theoretical accounts of the IPS implementing a priority map that combines stimulus- or value-dependent salience with goal-dependent feedback from the PFC[17,57,58,80,102].

In the dACC, we found that target and distractor difficulty encoding was consistent with the segregated encoding hypothesis, with features evoking univariate responses in distinct but adjacent regions. Interestingly, we did not find corresponding encoding of distractor congruence in our multivariate analyses within the dACC, potentially reflecting the spatial smoothness of this response. However, we did find multivariate encoding of distractor congruence in the lPFC and multivariate encoding of target and distractor coherence in the IPS. These multivariate profiles were consistent with our subspace encoding hypothesis. The reason for a mix of segregated and subspace encoding across the cortex is unclear, but this may speculatively reflect the segregation across functional networks. Like in the dACC, distractor congruence had stronger encoding in the lPFC Control network, albeit without the feature segregation (lPFC Control parcels also encoded target coherence in an orthogonal subspace). It is possible that these network segregations help bind related control processes[15,18,80], a hypothesis that future experiments should test with targeted paradigms (for example, with participant-specific functional networks).

By comparing two different task goals (Attend-Colour versus Attend-Motion), our study was able to test whether coherence representations reflect control-dependent prioritization of information processing. Previous research has shown that these tasks differ dramatically in their control demands[10]. As in previous work, task performance was much better in Attend-Motion runs than in Attend-Colour runs, and the participants were not sensitive to colour distractors. Consistent with previous work on context-dependent decision-making, target evidence had similarly strong encoding across tasks, with generalizable encoding dimensions for choice and motion directions[36,41,45]. In contrast to these putative decision representations, we found that coherence representations disappeared in the easier Attend-Motion task. On its own, weaker encoding of colour distractors in Attend-Motion could be explained by the weaker bottom-up salience of the colour dimension. However, the stark drop in the encoding of target (motion) coherence in these blocks cannot be similarly accounted for—these differences in target coherence encoding showed the opposite relationship expected from salience: better encoding of low-salience colour targets (hard Attend-Colour task) and weaker encoding of high-salience motion targets (easy Attend-Motion task). Instead, this encoding profile is consistent with previous research finding that feature decoding is stronger for more difficult tasks[47,71,72,103] or when people are incentivized to use cognitive control[104,105].

Critically, the stimuli and responses were matched across tasks, helping rule out alternative accounts of coherence encoding based on bottom-up stimulus salience, decision-making or eye movements. Difficulty-dependent coherence encoding may instead reflect the involvement of an attention control system that can separately regulate

target and distractor processing, speculatively indexing the top-down 'gain' or 'priority' on these features[17,58,102]. Supporting this account, coherence representations in cognitive control regions such as the IPS were aligned with performance representations, with target encoding strength aligned with better performance and distractor encoding strength aligned with poorer performance. Individual difference in feature–performance alignment was correlated across features, consistent with these representations reflecting the underlying processes (for example, priority) that give rise to behaviour, rather than performance monitoring or surprise (which would probably have the opposite relationship—for example, high target coherence aligned with poorer performance).

Classic models of prefrontal involvement in cognitive control[77,82,106] propose that the PFC biases information processing in sensory regions. In line with this macro-scale organization, we found that coherence encoding in the visual cortex was related to functional connectivity with the frontoparietal network. In particular, coherence encoding in the visual cortex was aligned with patterns of functional connectivity to the lPFC, and this feature–seed relationship was mediated by the IPS. The results of this multivariate connectivity analysis are consistent with previous research supporting a role for the IPS in top-down control of visual encoding[83,84,107], as well as a Granger-causal PFC–IPS–visual pathway during a similar decision-making task[42]. Here we demonstrate stable 'communication subspaces' between the visual cortex and PFC[108,109], which can plausibly communicate feedback adjustments to feature gain. With that said, while our interpretation of the direction of communication is therefore supported by prior work, these connectivity methods are correlational[85] and cannot rule out the possibility that our mediation findings reflect a bottom-up pattern of communication (for example, visual–IPS–PFC). The asymmetric mediation between regions (that is, the IPS mediates the lPFC more than the lPFC mediates the IPS; Supplementary Fig. 4) rules out a range of potential confounders, and these regions were selected on the basis of the anatomical connectivity within the frontoparietal network, notably through the superior longitudinal fasciculus[110]. Future research should use temporally precise neuroimaging to account for directionality and causal manipulations to account for causality (for example, ref. 111) and should explore the higher-dimensional connectivity subspaces that link different regions[103,109]. These considerations notwithstanding, our findings are consistent with the IPS, a critical site for orthogonal feature representations, playing a key role in linking the PFC with early perceptual processing.

Collectively, our findings provide new insights into how the brain may control multiple streams of information processing. While evidence for multivariate control has a long history in attentional tracking[28,112], including parametric relationships between attentional load and IPS activity[113–117], little is known about how the brain coordinates multiple control signals[2,5]. Future experiments should further elaborate on this frontoparietal control circuit—for instance, by interrogating how incentives influence different task representations[104,105,118–120] or how neural and behavioural indices of control causally depend on perturbations of neural activity[111]. Future experiments should also use fast-timescale neural recording technologies such as (i)EEG or (OP-)MEG to better understand the within-trial dynamics of multivariate control[10,121]. In sum, this experiment provides new insights into the large-scale neural networks involved in multivariate cognitive control and points towards new avenues for developing a richer understanding of goal-directed attention.

Methods

Participants

Twenty-nine individuals (17 females; age: mean, 21.2 years; s.d., 3.4 years) provided informed consent and participated in this experiment for compensation (US$40; institutional review board approval code: 1606001539). All participants had self-reported normal colour vision and no history of neurological disorders. Two participants missed one Attend-Colour block (see below) due to a scanner removal, and one participant missed a motion localizer due to a technical failure, but all participants were retained for analysis. This study was approved by Brown University's institutional review board.

Task

The main task closely followed our previously reported behavioural experiment[10]. On each trial, the participants saw an RDK against a black background. This RDK consisted of coloured dots that moved left or right, and the participants responded to the stimulus with button presses using their left or right thumbs.

In Attend-Colour blocks (six blocks of 150 trials), the participants responded depending on which colour was in the majority. Two colours were mapped to each response (four colours total), and the dots were a mixture of one colour from each possible response. The dot colours were approximately isolument (uncalibrated RGB: (239, 143, 143), (191, 239, 143), (143, 239, 239), (191, 143, 239)), and we counterbalanced their assignment to responses across participants.

In Attend-Motion blocks (six blocks of 45 trials), the participants responded on the basis of the dot motion instead of the dot colour. Dot motion consisted of a mixture between dots moving coherently (either left or right) and dots moving in random directions. Attend-Motion blocks were shorter because they acted to reinforce motion sensitivity and provide a test of stimulus-dependent effects.

Critically, the dots always had colour and motion, and we varied the strength of colour coherence (the percentage of dots in the majority) and motion coherence (the percentage of dots moving coherently) across trials. Our previous experiments have found that in Attend-Colour blocks, participants are still influenced by motion information, introducing a response conflict when colour and motion are associated with different responses[10]. Target coherence (for example, colour coherence during Attend-Colour) was linearly spaced between 65% and 95% with five levels, and distractor congruence (signed coherence relative to the target response) was linearly spaced between −95% and 95% with five levels. To increase the salience of the motion dimension relative to the colour dimension, the display was large (~10 degrees of visual angle), and the dots moved quickly (~10 degrees of visual angle per second).

The participants had 1.5 seconds from the onset of the stimulus to make their response, and the RDK stayed on the screen for this full duration to avoid confusing RT and visual stimulation (the fixation cross changed from white to grey to register the response). The intertrial interval was uniformly sampled from 1.0, 1.5 or 2.0 seconds. This intertrial interval was relatively short to maximize the behavioural effect, and because efficiency simulations showed that it increased power to detect parametric effects of target and distractor coherence (for example, relative to a more standard 5-second intertrial interval). The fixation cross changed from grey to white for the last 0.5 seconds before the stimulus to provide an alerting cue.

Procedure

Before the scanning session, the participants provided consent and practised the task in a mock MRI scanner. First, the participants learned to associate four colours with two button presses (two colours for each response). After being instructed on the colour–button mappings, the participants practised the task with feedback (correct, error or 1.5-second time-out). Errors or time-out feedback were accompanied with a diagram of the colour–button mappings. The participants performed 50 trials with full colour coherence and then 50 trials with variable colour coherence, all with 0% motion coherence. Next, the participants practised the motion task. After being shown the motion mappings, the participants performed 50 trials with full motion coherence and then 50 trials with variable motion coherence, all with 0% colour coherence. Finally, the participants practised 20 trials of the

Attend-Colour task and 20 trials of the Attend-Motion task with variable colour and motion coherence (the same as the scanner task).

Following the 12 blocks of the scanner task, the participants underwent localizers for colour and motion, based on the tasks used in our previous experiments[30]. Both localizers were block designs, alternating between 16 seconds of feature present and 16 seconds of feature absent for seven cycles. For the colour localizer, the participants saw an aperture the same size as the task, filled with either coloured squares that were resampled every second during stimulus-on ('Mondrian stimulus') or luminance-matched grey squares that were similarly resampled during stimulus-off. For the motion localizer, the participants saw white dots that were moving with full coherence in a different direction every second during stimulus-on or still dots during stimulus-off. No responses were required during the localizers.

MRI sequence

We scanned the participants with a Siemens Prisma 3T MR system. We used the following sequence parameters for our functional runs: field of view (FOV), 211 mm × 211 mm (60 slices); voxel size, 2.4 mm^3; repetition time (TR), 1.2 s with interleaved multiband acquisitions (acceleration factor 4); echo time (TE), 33 ms; flip angle (FA), 62°. Slices were acquired anterior to posterior, with an auto-aligned slice orientation tilted 15° relative to the AC/PC plane. At the start of the imaging session, we collected a high-resolution structural MPRAGE with the following sequence parameters: FOV, 205 mm × 205 mm (192 slices); voxel size, 0.8 mm^3; TR, 2.4 s; TE1, 1.86 ms; TE2, 3.78 ms; TE3, 5.7 ms; TE4, 7.62; FA, 7°. At the end of the scan, we collected a field map for susceptibility distortion correction (TR, 588 ms; TE1, 4.92 ms; TE2, 7.38 ms; FA, 60°).

fMRI preprocessing

We preprocessed our structural and functional data using fMRIPrep (v.20.2.6)[122], based on the Nipype platform[123]. We used FreeSurfer and ANTs to nonlinearly register structural T1w images to the MNI152NLin6Asym template (resampling to 2 mm). To preprocess functional T2w images, we applied susceptibility distortion correction using fMRIPrep, co-registered our functional images to our T1w images using FreeSurfer and slice-time corrected to the midpoint of the acquisition using AFNI. We then registered our images into MNI152NLin6Asym space using the transformation that ANTs computed for the T1w images, resampling our functional images in a single step. For univariate analyses, we smoothed our functional images using a Gaussian kernel (8 mm full width at half maximum, as dACC responses often have a large spatial extent). For multivariate analyses, we worked in the unsmoothed template space (see below).

fMRI univariate analyses

We used SPM12 (v.7771) for our univariate GLM analyses. Due to high trial-to-trial collinearity from to our short intertrial intervals, we performed all analyses across trials rather than extracting single-trial estimates. Our regression models used whole trials as events (that is, a 1.5-second boxcar aligned to the stimulus onset). We parametrically modulated these events with standardized trial-level predictors (for example, linear-coded target coherence or contrast-coded errors) and then convolved these predictors with SPM's canonical hemodynamic response function, concatenating our voxel time series across runs. We included nuisance regressors to capture (1) run intercepts and (2) the average time series across white matter and cerebrospinal fluid (as segmented by fMIRPrep). To reduce the influence of motion artefacts, we used robust weighted least-squares[124,125], a procedure for optimally down-weighting noisy TRs.

We estimated contrast maps at the participant level, which we then used for one-sample *t*-tests at the group level. We controlled for family-wise error rate using TFCE[126], testing whether voxels have an unlikely degree of clustering under a randomized null distribution (implemented in PALM[127]; 10,000 randomizations). To improve the specificity of our coverage (for example, reducing white-matter contributions) and to facilitate our inference about functional networks (see below), we limited these analyses to voxels within the Kong2022 whole-brain parcellation[54,55]. This parcellation assigns regional labels to parcels (for example, whether parcels are in the 'SPL' or 'IPS'), which was used throughout to generate ROIs. Surface renders were generated using surfplot[128,129], projecting from MNI space to the Human Connectome Project's fsLR space (164,000 vertices).

Table 1 | fMRI models

Model name	Trial selection	Predictors (*z*-scored)
Feature UV	No omission errors; run-concatenated	Target coherence, distractor coherence, target evidence, distractor evidence, distractor congruence; omission errors (run-concatenated)
Difficulty Levels	No omission errors; run-concatenated	Separate levels (1, 2, 4, 5) of target coherence, separate levels (1, 2, 4, 5) of distractor congruence; omission errors (run-concatenated)
Feature MV	No errors; run-separated	Target coherence, distractor coherence, target evidence, distractor evidence, distractor congruence; errors (run-concatenated)
Evidence Levels	No errors; run-separated	Levels (1–5, 6–10) of target evidence, levels (1, 2, 4, 5) of distractor evidence; errors (run-concatenated)
Between-Task	No errors; run-separated	Target coherence, distractor coherence, target evidence, distractor evidence, distractor congruence; errors (run-concatenated); RT (run-concatenated)
Performance	No omission errors; run-separated	Target coherence, distractor coherence, target evidence, distractor evidence, distractor congruence, RT, accuracy; omission errors (run-concatenated)
Performance CX	No omission errors; run-separated	Target coherence, distractor coherence, target evidence, distractor evidence, distractor congruence, RT, accuracy; omission errors (run-concatenated); seed time series (run-separated)

First-level GLMs used for univariate and multivariate fMRI analyses are shown. Coherence is the percentage of dots supporting the same response ('unsigned coherence'). Evidence is the percentage of dots supporting a rightward versus leftward response ('signed coherence'). Distractor congruence is the percentage of dots supporting the same response as the target dimension. All predictors were *z*-scored within their run. For difficulty and feature levels, we included each level as a separate predictor, with collinearity with the block intercept preventing all levels from being included. For Evidence Levels, targets had greater granularity due to distractors being coded relative to targets (five levels of congruence led to five levels of coherence). For Performance CX, seed time series were included as run-separated regressors (see 'Multivariate connectivity analysis').

dACC longitudinal axis analyses

To characterize the spatial organization of target difficulty and distractor congruence signals in dACC, we constructed an analysis mask that provided broad coverage across the cingulate cortex and preSMA. This mask was constructed by getting a meta-analytic mask of cingulate responses co-occurring with 'cognitive control' (Neurosynth uniformity test[130]) and taking any parcels from the whole-brain Schaefer parcellation (400 parcels[54,55]) that had a 50-voxel overlap with this meta-analytic mask. We used this parcellation because it provided more selective grey-matter coverage than the Neurosynth mask alone and it allowed us to categorize voxel membership in putative functional networks.

To characterize the spatial organization within the dACC, we first performed PCA on the masked voxel coordinates (y and z), getting a score for each voxel's position on the longitudinal axis of this ROI. We then regressed the voxels' gradient scores against their regression weights from a model including linear target coherence and distractor congruence (both coded −1 to 1 across difficulty levels). We used linear mixed-effects analysis to partially pool across participants and accommodate within-participant correlations between voxels. Our model predicted gradient score from the linear and quadratic expansions of the target and distractor βs (gradientScore ~ 1 + target + target^2 + distractor + distractor^2 + (1 + target + target^2 + distractor + distractor^2 | participant)). To characterize the network-dependent organization of target and distractor encoding, we complexity-penalized fits between models that either (1) predicted target or distractor βs from linear and quadratic expansions of gradient scores or (2) predicted target/distractor βs from dummy-coded network assignment from the Schaefer parcellation, comparing these models against a model that used both network and gradient information.

Encoding Geometry Analysis (EGA)

We adapted functions from the pcm-toolbox and rsatoolbox packages for our multivariate analyses[65,131]. We first fit whole-brain GLMs without spatial smoothing, separately for each scanner run. These GLMs estimated the parametric relationship between task variables and blood-oxygen-level-dependent response (for example, linearly coded target coherence), with a pattern of these parametric βs across voxels reflecting linear encoding subspace[59]. Within each Schaefer parcel (N = 400), we spatially pre-whitened these β maps, reducing noise correlations between voxels that can inflate pattern similarity and reduce reliability[63]. We then computed the cross-validated Pearson correlation, estimating the similarity of whitened patterns across scanner runs. We used a correlation metric to estimate the alignment between encoding subspaces, rather than distances between condition patterns, to normalize biases and scaling across stimuli (for example, greater sensitivity to targets versus distractors) and across time (for example, representational drift). Note that this analysis approach is related to 'Parallelism Scores'[43], but here we use parametric encoding models and emphasize not only deviations from parallel/orthogonal but also the direction of alignment between features (for example, Figs. 5 and 7).

We computed subspace alignment between contrasts of interest within each participant and then tested these against zero at the group level. Since our correlations were less than r = |0.5|, we did not transform the correlations before analysis. We used a Bayesian t-test to test for orthogonality (bayesFactor toolbox in MATLAB, based on ref. 132). The BF from this t-test gives evidence for either non-orthogonality (BF_{10} further from zero) or orthogonality (BF_{10} closer to zero, often defined as the reciprocal, BF_{01}). Using a standard prior (Cauchy, width = 0.707), our strongest possible evidence for the orthogonality is BF_{01} = 5.07 or equivalently log(BF) = −0.705 (that is, the BF when t_{28} = 0).

Our measure of encoding strength was whether encoding subspaces were reliable across blocks (that is, whether within-feature encoding pattern correlations across runs were significantly above zero at the group level). We used pattern reliability as a geometric proxy for how well a linear encoder would predict held-out brain data, as reliability indicates the similarity between the cross-validated model and the best linear unbiased estimator of the within-sample data. We confirmed through simulations that pattern reliability is a good proxy for the traditional encoding metric of predicting held-out time series[59]. However, we found that pattern reliability is more powerful, due to it being much less sensitive to the magnitude of residual variance (these two methods are identical in the noise-free case; Extended Data Fig. 3).

When looking at alignment between two subspaces across parcels, we first selected parcels that significantly encoded both factors ('jointly reliable parcels', both P < 0.001 uncorrected). This selection process acts as a thresholded version of classical correlation disattenuation[66,67], and we confirmed through simulations that this selection procedure does not increase the type 1 error rate. We corrected for multiple comparisons using non-parametric max-statistic randomization tests across parcels[133]. These randomization tests determine the likelihood of an observed effect under a null distribution generated by randomizing the sign of alignment correlations across participants and parcels 10,000 times. Within each randomization, we saved the maximum and minimum group-level effect sizes across all parcels, estimating the strongest parcel-wise effect we would expect if there was no systematic group-level effect.

Some of our first-level models had non-zero levels of multicollinearity, due to conditioning on trials without omission errors or when including feature coherence and performance in the same model. Multicollinearity was far below standard thresholds for concern, generally (much) less than 5; a standard threshold is 30 (the ratio between the largest and smallest singular values in the design matrix, using MATLAB colintest[134]). However, we wanted to confirm that predictor correlations would not bias our estimates of encoding alignment. We simulated data from a pattern component model[131] in which two variables were orthogonal (generated by separate variance components with no covariance) but were generated from a design matrix with correlated predictors. These simulations confirmed that cross-validated similarity measures were not biased by predictor correlations (Extended Data Fig. 10).

To provide further validation for our parametric analyses, we estimated encoding profiles using an analysis with fewer parametric assumptions. First, we fit a GLM with separate predictors for levels of target and distractor evidence (the 'Evidence Levels' GLM in Table 1). Next, we estimated a traditional (cross-validated) representational dissimilarity matrix across all feature levels. Finally, we visualized these encoding profiles using classical multidimensional scaling (eigenvalue decomposition; Fig. 4b and Extended Data Fig. 5).

Multivariate connectivity analysis

To estimate what information is plausibly communicated between cortical areas, we measured the alignment between multivariate connectivity patterns (that is, the 'communication subspace' with a seed region[108]) and local feature encoding patterns. We first residualized our Performance GLM (Table 1) from a seed region's time series and then extracted the variance-weighted average time course (that is, the leading eigenvariate from SPM12's volume-of-interest function). We then re-estimated our Performance GLM, including the block-specific seed time series as a covariate, and performed EGA between seed and coherence patterns (equations (1)–(3)). We found convergent results when we residualized a quadratic expansion of our Performance GLM from our seed region, helping confirm that connectivity alignment was not due to underfitting. Note that our cross-validated EGA helps avoid false positives due to any correlations in the design matrix (see above). We localized this connectivity analysis to colour- and motion-sensitive cortex by finding the bilateral Kong22 parcels that roughly covered the area of strongest block-level contrast during our localizer runs. Note that these analyses reflect 'functional connectivity', which is susceptible to unmodelled confounders[85].

To estimate the mediation of lPFC connectivity by the IPS, we compared models in which just the lPFC or just the IPS were used for EGA against a model where both seeds were included as covariates in the same model[86] (equations (4) and (5)). Our test of mediation was the group-level difference in lPFC seed–coherence alignment before and after including the IPS. While these analyses are inherently cross-sectional (that is, the lPFC and IPS are measured at the same time), we supplemented these analyses by showing that the mediating effect of the IPS on the lPFC was much larger than the mediating effect of the lPFC on the IPS (Fig. 7b and Supplementary Fig. 4). Unlike traditional mediation analyses looking at the first-order change in regression estimates, our analysis looks at the second-order change in the multivariate alignment between regression estimates, using the same core rationale.

Reporting summary
Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability
The unprocessed fMRI data are available at https://doi.org/10.18112/openneuro.ds004909.v1.1.0. The behavioural data and event timing are available at https://github.com/shenhavlab/PACT_fMRI_public.

Code availability
The analysis pipeline and code are available at https://github.com/shenhavlab/PACT_fMRI_public. The software versions used are MATLAB v.2020a, fMRIPrep v.20.2.6, SPM12 (v.7771), rwls v.4.1, PALM v.a119, rsatoolbox_matlab v.1.0, bayesFactor v.1.1, surfplot v.0.1.0 and ScientificColourMaps7.

References

1. Musslick, S., Shenhav, A., Botvinick, M. & Cohen, J. A Computational model of control allocation based on the expected value of control. In *2nd Multidisciplinary Conference on Reinforcement Learning and Decision Making* (RLDM, 2019).
2. Badre, D., Bhandari, A., Keglovits, H. & Kikumoto, A. The dimensionality of neural representations for control. *Curr. Opin. Behav. Sci.* **38**, 20–28 (2021).
3. Danielmeier, C., Eichele, T., Forstmann, B. U., Tittgemeyer, M. & Ullsperger, M. Posterior medial frontal cortex activity predicts post-error adaptations in task-related visual and motor areas. *J. Neurosci.* **31**, 1780–1789 (2011).
4. Egner, T. Multiple conflict-driven control mechanisms in the human brain. *Trends Cogn. Sci.* **12**, 374–380 (2008).
5. Ritz, H., Leng, X. & Shenhav, A. Cognitive control as a multivariate optimization problem. *J. Cogn. Neurosci.* **34**, 569–591 (2022).
6. Friedman, N. P. & Miyake, A. Unity and diversity of executive functions: individual differences as a window on cognitive structure. *Cortex* **86**, 186–204 (2017).
7. Danielmeier, C. & Ullsperger, M. Post-error adjustments. *Front. Psychol.* **2**, 233 (2011).
8. Fischer, A. G., Nigbur, R., Klein, T. A., Danielmeier, C. & Ullsperger, M. Cortical beta power reflects decision dynamics and uncovers multiple facets of post-error adaptation. *Nat. Commun.* **9**, 5038 (2018).
9. Leng, X., Yee, D., Ritz, H. & Shenhav, A. Dissociable influences of reward and punishment on adaptive cognitive control. *PLoS Comput. Biol.* **17**, e1009737 (2021).
10. Ritz, H. & Shenhav, A. Humans reconfigure target and distractor processing to address distinct task demands. *Psychol. Rev.* https://doi.org/10.1037/rev0000442 (2023).
11. Kalman, R. E. On the general theory of control systems. *IFAC Proc.* **1**, 491–502 (1960).
12. Shenhav, A., Botvinick, M. M. & Cohen, J. D. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron* **79**, 217–240 (2013).
13. MacDonald, A. W. 3rd, Cohen, J. D., Stenger, V. A. & Carter, C. S. Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science* **288**, 1835–1838 (2000).
14. Smith, E. H. et al. Widespread temporal coding of cognitive control in the human prefrontal cortex. *Nat. Neurosci.* **22**, 1883–1891 (2019).
15. Menon, V. & D'Esposito, M. The role of PFC networks in cognitive control and executive function. *Neuropsychopharmacology* https://doi.org/10.1038/s41386-021-01152-w (2021).
16. Kerns, J. G. et al. Anterior cingulate conflict monitoring and adjustments in control. *Science* **303**, 1023–1026 (2004).
17. Gottlieb, J., Cohanpour, M., Li, Y., Singletary, N. & Zabeh, E. Curiosity, information demand and attentional priority. *Curr. Opin. Behav. Sci.* **35**, 83–91 (2020).
18. Gordon, E. M. et al. Precision functional mapping of individual human brains. *Neuron* **95**, 791–807.e7 (2017).
19. Gratton, C., Laumann, T. O., Gordon, E. M., Adeyemo, B. & Petersen, S. E. Evidence for two independent factors that modify brain networks to meet task goals. *Cell Rep.* **17**, 1276–1288 (2016).
20. Kragel, P. A. et al. Generalizable representations of pain, cognitive control, and negative emotion in medial frontal cortex. *Nat. Neurosci.* **21**, 283–289 (2018).
21. Fu, Z. et al. The geometry of domain-general performance monitoring in the human medial frontal cortex. *Science* **376**, eabm9922 (2022).
22. Vermeylen, L. et al. Shared neural representations of cognitive conflict and negative affect in the medial frontal cortex. *J. Neurosci.* **40**, 8715–8725 (2020).
23. Brown, J. W. & Braver, T. S. Learned predictions of error likelihood in the anterior cingulate cortex. *Science* **307**, 1118–1121 (2005).
24. Rushworth, M. F. & Behrens, T. E. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat. Neurosci.* **11**, 389–397 (2008).
25. Grinband, J. et al. The dorsal medial frontal cortex is sensitive to time on task, not response conflict or error likelihood. *NeuroImage* **57**, 303–311 (2011).
26. Yarkoni, T., Barch, D. M., Gray, J. R., Conturo, T. E. & Braver, T. S. BOLD correlates of trial-by-trial reaction time variability in gray and white matter: a multi-study fMRI analysis. *PLoS ONE* **4**, e4257 (2009).
27. Mumford, J. A. et al. The response time paradox in functional magnetic resonance imaging analyses. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-023-01760-0 (2023).
28. Pylyshyn, Z. W. & Storm, R. W. Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spat. Vis.* **3**, 179–197 (1988).
29. Beldzik, E. & Ullsperger, M. A thin line between conflict and reaction time effects on EEG and fMRI brain signals. Preprint at *bioRxiv* https://doi.org/10.1101/2023.02.14.528515 (2023).
30. Shenhav, A., Straccia, M. A., Musslick, S., Cohen, J. D. & Botvinick, M. M. Dissociable neural mechanisms track evidence accumulation for selection of attention versus action. *Nat. Commun.* **9**, 2485 (2018).
31. Taren, A. A., Venkatraman, V. & Huettel, S. A. A parallel functional topography between medial and lateral prefrontal cortex: evidence and implications for cognitive control. *J. Neurosci.* **31**, 5026–5031 (2011).
32. Venkatraman, V., Rosati, A. G., Taren, A. A. & Huettel, S. A. Resolving response, decision, and strategic control: evidence for a functional topography in dorsomedial prefrontal cortex. *J. Neurosci.* **29**, 13158–13164 (2009).
33. Fu, Z. et al. Single-neuron correlates of error monitoring and post-error adjustments in human medial frontal cortex. *Neuron* **101**, 165–177.e5 (2019).
34. Zarr, N. & Brown, J. W. Hierarchical error representation in medial prefrontal cortex. *NeuroImage* **124**, 238–247 (2016).
35. Ebitz, B. R. et al. Human dorsal anterior cingulate neurons signal conflict by amplifying task-relevant information. Preprint at *bioRxiv* https://doi.org/10.1101/2020.03.14.991745 (2020).
36. Flesch, T., Juechems, K., Dumbalska, T., Saxe, A. & Summerfield, C. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron* **110**, 1258–1270 (2022).
37. Minxha, J., Adolphs, R., Fusi, S., Mamelak, A. N. & Rutishauser, U. Flexible recruitment of memory-based choice representations by the human medial frontal cortex. *Science* **368**, eaba3313 (2020).

38. Ebitz, R. B. & Hayden, B. Y. The population doctrine in cognitive neuroscience. *Neuron* https://doi.org/10.1016/j.neuron.2021.07.011 (2021).
39. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* **17**, 1500–1509 (2014).
40. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
41. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
42. Kayser, A. S., Erickson, D. T., Buchsbaum, B. R. & D'Esposito, M. Neural representations of relevant and irrelevant features in perceptual decision making. *J. Neurosci.* **30**, 15778–15789 (2010).
43. Bernardi, S. et al. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954–967 (2020).
44. Panichello, M. F. & Buschman, T. J. Shared mechanisms underlie the control of working memory and attention. *Nature* https://doi.org/10.1038/s41586-021-03390-w (2021).
45. Takagi, Y., Hunt, L. T., Woolrich, M. W., Behrens, T. E. & Klein-Flügge, M. C. Adapting non-invasive human recordings along multiple task-axes shows unfolding of spontaneous and over-trained choice. *eLife* **10**, e60988 (2021).
46. Cohen, J. D., Dunbar, K. & McClelland, J. L. On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychol. Rev.* **97**, 332–361 (1990).
47. Woolgar, A., Hampshire, A., Thompson, R. & Duncan, J. Adaptive coding of task-relevant information in human frontoparietal cortex. *J. Neurosci.* **31**, 14592–14599 (2011).
48. Nee, D. E., Wager, T. D. & Jonides, J. Interference resolution: insights from a meta-analysis of neuroimaging tasks. *Cogn. Affect. Behav. Neurosci.* **7**, 1–17 (2007).
49. Shenhav, A., Straccia, M. A., Botvinick, M. M. & Cohen, J. D. Dorsal anterior cingulate and ventromedial prefrontal cortex have inverse roles in both foraging and economic choice. *Cogn. Affect. Behav. Neurosci*. https://doi.org/10.3758/s13415-016-0458-8 (2016).
50. Fleming, S. M., van der Putten, E. J. & Daw, N. D. Neural mediators of changes of mind about perceptual decisions. *Nat. Neurosci.* **21**, 617–624 (2018).
51. Shenhav, A. & Karmarkar, U. R. Dissociable components of the reward circuit are involved in appraisal versus choice. *Sci. Rep.* **9**, 1958 (2019).
52. Clairis, N. & Pessiglione, M. Value, confidence, deliberation: a functional partition of the medial prefrontal cortex demonstrated across rating and choice tasks. *J. Neurosci.* **42**, 5580–5592 (2022).
53. Yeo, B. T. T. et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).
54. Kong, R. et al. Individual-specific areal-level parcellations improve functional connectivity prediction of behavior. *Cereb. Cortex* **31**, 4477–4500 (2021).
55. Schaefer, A. et al. Local–global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex* **28**, 3095–3114 (2018).
56. Kong, R. et al. Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion. *Cereb. Cortex* **29**, 2533–2551 (2019).
57. Bisley, J. W. & Mirpour, K. The neural instantiation of a priority map. *Curr. Opin. Psychol.* **29**, 108–112 (2019).
58. Yantis, S. & Serences, J. T. Cortical mechanisms of space-based and object-based attentional control. *Curr. Opin. Neurobiol.* **13**, 187–193 (2003).
59. Kriegeskorte, N. & Diedrichsen, J. Peeling the onion of brain representations. *Annu. Rev. Neurosci.* **42**, 407–432 (2019).
60. Cohen, M. R. & Maunsell, J. H. R. A neuronal population measure of attention predicts behavioral performance on individual trials. *J. Neurosci.* **30**, 15241–15253 (2010).
61. Libby, A. & Buschman, T. J. Rotational dynamics reduce interference between sensory and memory representations. *Nat. Neurosci*. https://doi.org/10.1038/s41593-021-00821-9 (2021).
62. Kimmel, D. L., Elsayed, G. F., Cunningham, J. P. & Newsome, W. T. Value and choice as separable and stable representations in orbitofrontal cortex. *Nat. Commun.* **11**, 3466 (2020).
63. Walther, A. et al. Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage* **137**, 188–200 (2016).
64. Diedrichsen, J. & Kriegeskorte, N. Representational models: a common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Comput. Biol.* **13**, e1005508 (2017).
65. Nili, H. et al. A toolbox for representational similarity analysis. *PLoS Comput. Biol.* **10**, e1003553 (2014).
66. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **100**, 441–471 (1987).
67. Thornton, M. A. & Mitchell, J. P. Consistent neural activity patterns represent personally familiar people. *J. Cogn. Neurosci.* **29**, 1583–1594 (2017).
68. Hunt, L. T. et al. Mechanisms underlying cortical activity during value-guided choice. *Nat. Neurosci.* **15**, 470–476 (2012).
69. Kayser, A. S., Buchsbaum, B. R., Erickson, D. T. & D'Esposito, M. The functional anatomy of a perceptual decision in the human brain. *J. Neurophysiol.* **103**, 1179–1194 (2010).
70. Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl Acad. Sci. USA* **103**, 3863–3868 (2006).
71. Woolgar, A., Williams, M. A. & Rich, A. N. Attention enhances multi-voxel representation of novel objects in frontal, parietal and visual cortices. *NeuroImage* **109**, 429–437 (2015).
72. Woolgar, A., Afshar, S., Williams, M. A. & Rich, A. N. Flexible coding of task rules in frontoparietal cortex: an adaptive system for flexible cognitive control. *J. Cogn. Neurosci.* **27**, 1895–1911 (2015).
73. Jackson, J., Rich, A. N., Williams, M. A. & Woolgar, A. Feature-selective attention in frontoparietal cortex: multivoxel codes adjust to prioritize task-relevant information. *J. Cogn. Neurosci.* **29**, 310–321 (2017).
74. Woolgar, A., Thompson, R., Bor, D. & Duncan, J. Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. *NeuroImage* **56**, 744–752 (2011).
75. Aoi, M. C., Mante, V. & Pillow, J. W. Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. *Nat. Neurosci.* https://doi.org/10.1038/s41593-020-0696-5 (2020).
76. Pagan, M. et al. A new theoretical framework jointly explains behavioral and neural variability across subjects performing flexible decision-making. Preprint at *bioRxiv* https://doi.org/10.1101/2022.11.28.518207 (2022).
77. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).
78. Stringer, C. et al. Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364**, 255 (2019).
79. Goldman-Rakic, P. S. Topography of cognition: parallel distributed networks in primate association cortex. *Annu. Rev. Neurosci.* **11**, 137–156 (1988).
80. Corbetta, M. & Shulman, G. L. Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* **3**, 201–215 (2002).
81. Suzuki, M. & Gottlieb, J. Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nat. Neurosci.* **16**, 98–104 (2013).

82. Kastner, S. & Ungerleider, L. G. Mechanisms of visual attention in the human cortex. *Annu. Rev. Neurosci.* **23**, 315–341 (2000).
83. Kay, K. N. & Yeatman, J. D. Bottom-up and top-down computations in word- and face-selective cortex. *eLife* **6**, e22341 (2017).
84. Saalmann, Y. B., Pigarev, I. N. & Vidyasagar, T. R. Neural mechanisms of visual attention: how top-down feedback highlights relevant locations. *Science* **316**, 1612–1615 (2007).
85. Reid, A. T. et al. Advancing functional connectivity research from association to causation. *Nat. Neurosci.* **22**, 1751–1760 (2019).
86. MacKinnon, D. P., Fairchild, A. J. & Fritz, M. S. Mediation analysis. *Annu. Rev. Psychol.* **58**, 593–614 (2007).
87. Friedman, N. P. & Robbins, T. W. The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology* https://doi.org/10.1038/s41386-021-01132-0 (2021).
88. Holroyd, C. B. & McClure, S. M. Hierarchical control over effortful behavior by rodent medial frontal cortex: a computational model. *Psychol. Rev.* **122**, 54–83 (2015).
89. Holroyd, C. B. & Yeung, N. in *Neural Basis of Motivational and Cognitive Control* (eds Mars, R. B. et al.) 332–349 (MIT Press, 2011); https://doi.org/10.7551/mitpress/9780262016438.003.0018
90. Vassena, E., Deraeve, J. & Alexander, W. H. Predicting motivation: computational models of PFC can explain neural coding of motivation and effort-based decision-making in health and disease. *J. Cogn. Neurosci.* **29**, 1633–1645 (2017).
91. Koechlin, E. & Summerfield, C. An information theoretical approach to prefrontal executive function. *Trends Cogn. Sci.* **11**, 229–235 (2007).
92. Badre, D. & D'Esposito, M. Is the rostro-caudal axis of the frontal lobe hierarchical? *Nat. Rev. Neurosci.* **10**, 659–669 (2009).
93. Badre, D. & Nee, D. E. Frontal cortex and the hierarchical control of behavior. *Trends Cogn. Sci.* **22**, 170–188 (2018).
94. Serences, J. T. & Yantis, S. Spatially selective representations of voluntary and stimulus-driven attentional priority in human occipital, parietal, and frontal cortex. *Cereb. Cortex* **17**, 284–293 (2007).
95. Yantis, S. et al. Transient neural activity in human parietal cortex during spatial attention shifts. *Nat. Neurosci.* **5**, 995–1002 (2002).
96. Greenberg, A. S., Esterman, M., Wilson, D., Serences, J. T. & Yantis, S. Control of spatial and feature-based attention in frontoparietal cortex. *J. Neurosci.* **30**, 14330–14339 (2010).
97. Esterman, M., Chiu, Y.-C., Tamber-Rosenau, B. J. & Yantis, S. Decoding cognitive control in human parietal cortex. *Proc. Natl Acad. Sci. USA* **106**, 17974–17979 (2009).
98. Serences, J. T., Schwarzbach, J., Courtney, S. M., Golay, X. & Yantis, S. Control of object-based attention in human cortex. *Cereb. Cortex* **14**, 1346–1357 (2004).
99. Molenberghs, P., Mesulam, M. M., Peeters, R. & Vandenberghe, R. R. C. Remapping attentional priorities: differential contribution of superior parietal lobule and intraparietal sulcus. *Cereb. Cortex* **17**, 2703–2712 (2007).
100. Adam, K. C. S. & Serences, J. T. History modulates early sensory processing of salient distractors. *J. Neurosci.* https://doi.org/10.1523/JNEUROSCI.3099-20.2021 (2021).
101. Soutschek, A., Stelzel, C., Paschke, L., Walter, H. & Schubert, T. Dissociable effects of motivation and expectancy on conflict processing: an fMRI study. *J. Cogn. Neurosci.* **27**, 409–423 (2015).
102. Bisley, J. W. & Goldberg, M. E. Attention, intention, and priority in the parietal lobe. *Annu. Rev. Neurosci.* **33**, 1–21 (2010).
103. Rust, N. C. & Cohen, M. R. Priority coding in the visual system. *Nat. Rev. Neurosci.* https://doi.org/10.1038/s41583-022-00582-9 (2022).
104. Etzel, J. A., Cole, M. W., Zacks, J. M., Kay, K. N. & Braver, T. S. Reward motivation enhances task coding in frontoparietal cortex. *Cereb. Cortex* **26**, 1647–1659 (2016).
105. Hall-McMaster, S., Muhle-Karbe, P. S., Myers, N. E. & Stokes, M. G. Reward boosts neural coding of task rules to optimize cognitive flexibility. *J. Neurosci.* **39**, 8549–8561 (2019).
106. Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* **18**, 193–222 (1995).
107. Lauritzen, T. Z., D'Esposito, M., Heeger, D. J. & Silver, M. A. Top-down flow of visual spatial attention signals from parietal to occipital cortex. *J. Vis.* **9**, 18 (2009).
108. Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M. & Kohn, A. Cortical areas interact through a communication subspace. *Neuron* **102**, 249–259.e4 (2019).
109. Srinath, R., Ruff, D. A. & Cohen, M. R. Attention improves information flow between neuronal populations without changing the communication subspace. *Curr. Biol.* **31**, 5299–5313 (2021).
110. Petrides, M. & Pandya, D. N. Efferent association pathways originating in the caudal prefrontal cortex in the macaque monkey. *J. Comp. Neurol.* **498**, 227–251 (2006).
111. Jackson, J. B., Feredoes, E., Rich, A. N., Lindner, M. & Woolgar, A. Concurrent neuroimaging and neurostimulation reveals a causal role for dlPFC in coding of task-relevant information. *Commun. Biol.* **4**, 588 (2021).
112. Vul, E., Alvarez, G., Tenenbaum, J. & Black, M. in *Advances in Neural Information Processing Systems 22* (eds Bengio, Y. et al.) 1–9 (Curran Associates, 2009).
113. Ritz, H., Wild, C. J. & Johnsrude, I. S. Parametric cognitive load reveals hidden costs in the neural processing of perfectly intelligible degraded speech. *J. Neurosci.* **42**, 4619–4628 (2022).
114. Culham, J. C., Cavanagh, P. & Kanwisher, N. G. Attention response functions: characterizing brain areas using fMRI activation during parametric variations of attentional load. *Neuron* **32**, 737–745 (2001).
115. Culham, J. C. et al. Cortical fMRI activation produced by attentive tracking of moving targets. *J. Neurophysiol.* **80**, 2657–2670 (1998).
116. Howe, P. D., Horowitz, T. S., Morocz, I. A., Wolfe, J. & Livingstone, M. S. Using fMRI to distinguish components of the multiple object tracking task. *J. Vis.* **9**, 10.1–11 (2009).
117. Jovicich, J. et al. Brain areas specific for attentional load in a motion-tracking task. *J. Cogn. Neurosci.* **13**, 1048–1058 (2001).
118. Peck, C. J., Jangraw, D. C., Suzuki, M., Efem, R. & Gottlieb, J. Reward modulates attention independently of action value in posterior parietal cortex. *J. Neurosci.* **29**, 11182–11191 (2009).
119. Wisniewski, D., Reverberi, C., Momennejad, I., Kahnt, T. & Haynes, J.-D. The role of the parietal cortex in the representation of task–reward associations. *J. Neurosci.* **35**, 12355–12365 (2015).
120. Parro, C., Dixon, M. L. & Christoff, K. The neural basis of motivational influences on cognitive control. *Hum. Brain Mapp.* **39**, 5097–5111 (2018).
121. Weichart, E. R., Turner, B. M. & Sederberg, P. B. A model of dynamic, within-trial conflict resolution for decision making. *Psychol. Rev.* https://doi.org/10.1037/rev0000191 (2020).
122. Esteban, O. et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
123. Gorgolewski, K. et al. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. *Front. Neuroinform.* **5**, 13 (2011).
124. Diedrichsen, J. & Shadmehr, R. Detecting and adjusting for artifacts in fMRI time series data. *NeuroImage* **27**, 624–634 (2005).
125. Jones, M. S., Zhu, Z., Bajracharya, A., Luor, A. & Peelle, J. E. A multi-dataset evaluation of frame censoring for task-based fMRI. *Aperture Neuro.* https://doi.org/10.52294/apertureneuro.2022.2.nxor2026 (2022).
126. Smith, S. M. & Nichols, T. E. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage* **44**, 83–98 (2009).

127. Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. & Nichols, T. E. Permutation inference for the general linear model. *NeuroImage* **92**, 381–397 (2014).
128. Vos de Wael, R. et al. BrainSpace: a toolbox for the analysis of macroscale gradients in neuroimaging and connectomics datasets. *Commun. Biol.* **3**, 103 (2020).
129. Gale, D. J., Vos de Wael., R., Benkarim, O. & Bernhardt, B. Surfplot: publication-ready brain surface figures. *Zenodo* https://doi.org/10.5281/zenodo.5567926 (2021).
130. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670 (2011).
131. Diedrichsen, J., Yokoi, A. & Arbuckle, S. A. Pattern component modeling: a flexible approach for understanding the representational structure of brain activity patterns. *NeuroImage* **180**, 119–133 (2018).
132. Rouder, J. N., Morey, R. D., Speckman, P. L. & Province, J. M. Default Bayes factors for ANOVA designs. *J. Math. Psychol.* **56**, 356–374 (2012).
133. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* **15**, 1–25 (2002).
134. Belsley, D. A., Kuh, E. & Welsch, R. E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* Wiley Series in Probability and Statistics (John Wiley & Sons, 1980).

Acknowledgements

This work was supported by NIH grant no. R01MH124849 (A.S.), NSF CAREER Award no. 2046111(A.S.), NIH grant no. S10OD025181 (A.S.) and the C.V. Starr Postdoctoral Fellowship (H.R.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank J. Kim for her assistance in data collection and M. J. Frank, M. N. Nassar, J. Cohen, M. Esterman, R. Frömer, J. Diedrichsen, A. Bhandari, D. Yee, S. Nastase, C. Jahn and the Shenhav Lab for helpful discussions.

Author contributions

Both authors designed the experiment, planned the analyses and wrote the manuscript. H.R. collected the data and conducted the analyses.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41562-024-01826-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-024-01826-7.

Correspondence and requests for materials should be addressed to Harrison Ritz.

Peer review information *Nature Human Behaviour* thanks Tobias Egner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

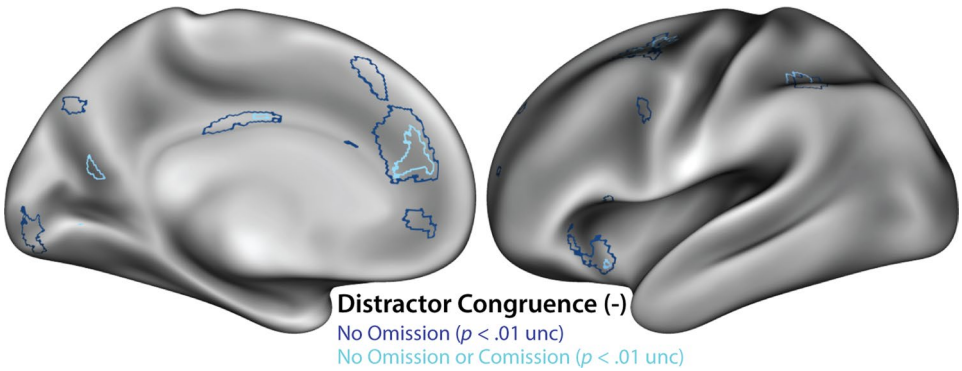
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

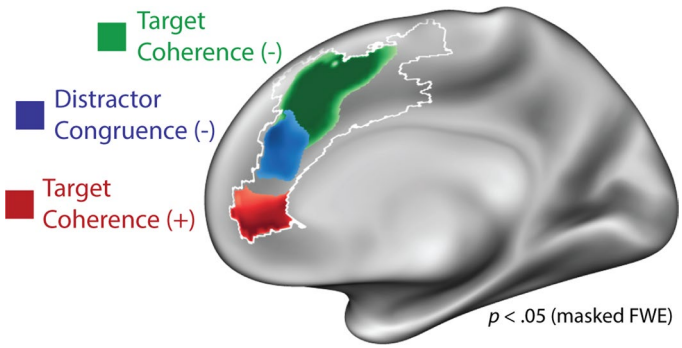
© The Author(s), under exclusive licence to Springer Nature Limited 2024

https://doi.org/10.1038/s41562-024-01826-7



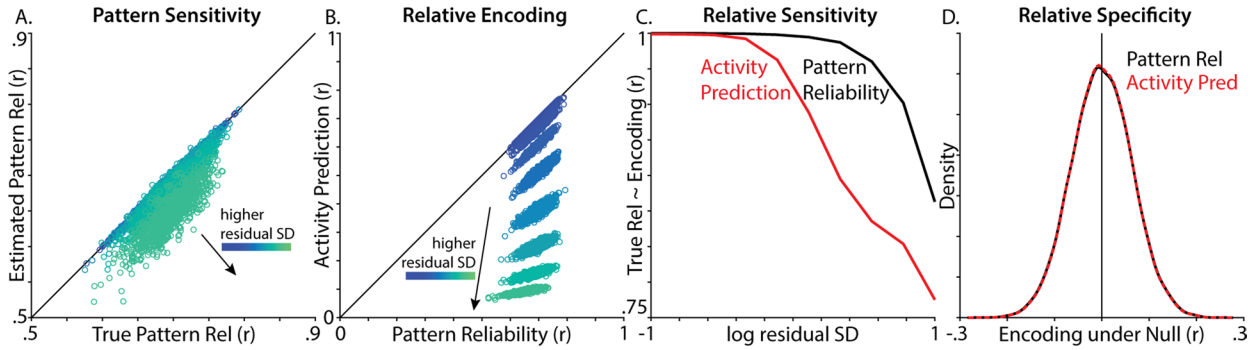
Extended Data Fig. 1 | Error control analysis. Distractor congruence effect when controlling for different types of errors (two-tailed *t*-test, thresholded at $p < 0.01$ uncorrected). Our primary analysis only analyzed trials without omission errors (navy), here plotted at a liberal uncorrected threshold. When we analyze trials without omission errors and commission errors (cyan), we see a consistent whole-brain topography, albeit at a lower statistical threshold. In both cases, relevant errors trials were included as nuisance events.

https://doi.org/10.1038/s41562-024-01826-7



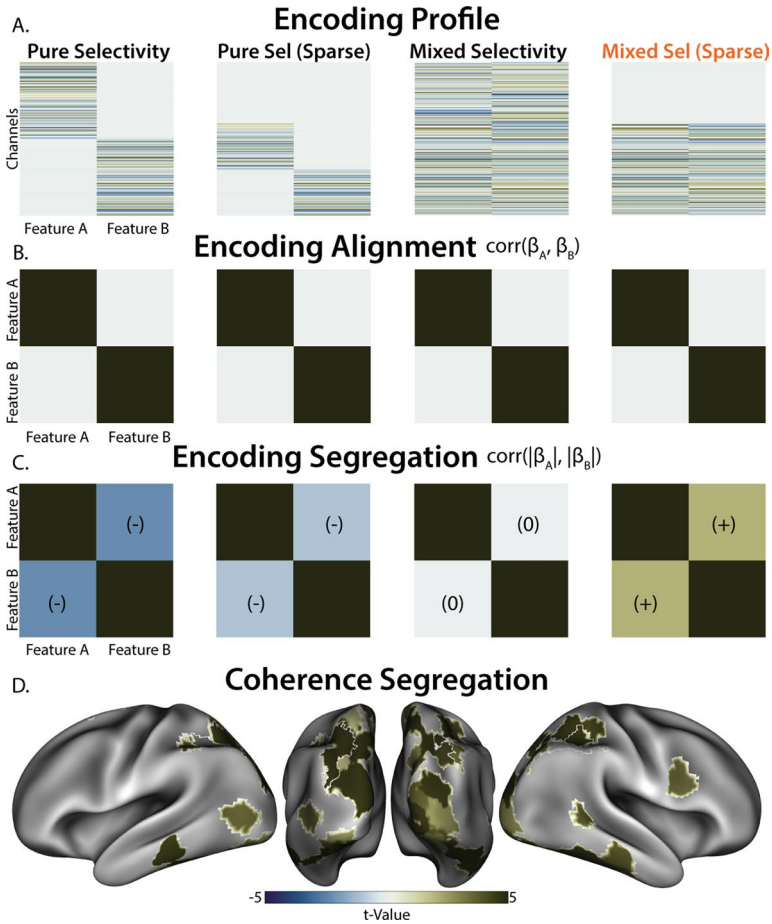
Extended Data Fig. 2 | Univariate fMRI response to target ease. Parametric effects of target coherence and distractor congruence (two-tailed *t*-test, corrected using threshold-free cluster enhancement). Here we included the rostral effect of target ease (positive relationship with target coherence) in red. Compare to Fig. 2.

https://doi.org/10.1038/s41562-024-01826-7



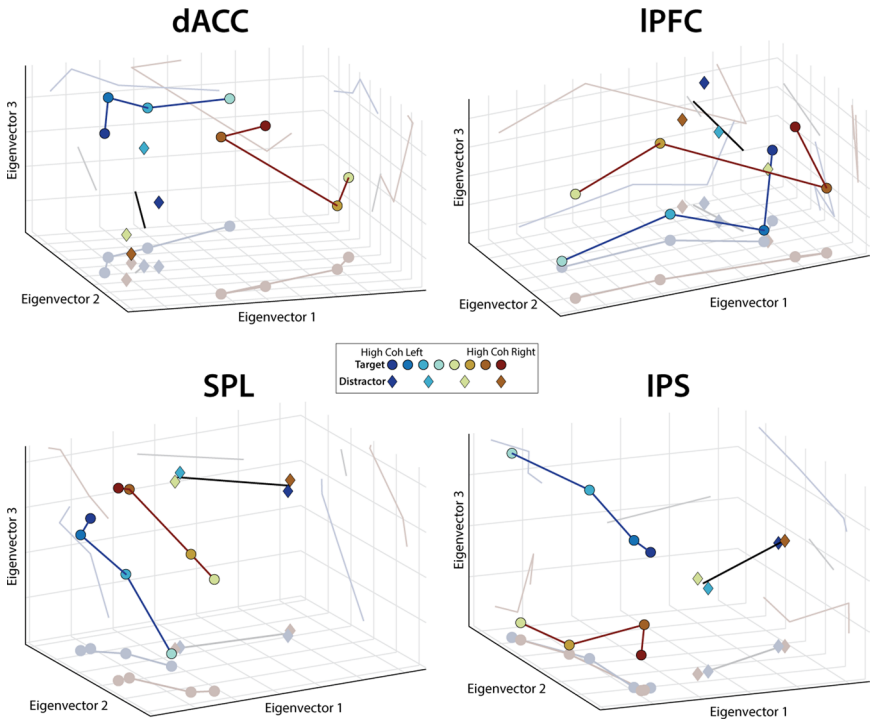
Extended Data Fig. 3 | Encoding Geometry Analysis (EGA) validation. We validated how well we could recover the similarity between linear Gaussian models (training: $Y = XB + \Sigma$, test: $Y' = X'B' + \Sigma$). Y is the [1000 × 250] activity timeseries, X is the [1000 × 1] design matrix, B is the [1 × 250] encoding profile, and Σ reflects IID Gaussian noise. In each of our 1000 simulations, we used two different methods to recover the similarity between the true training encoding profile (B) and the true test encoding profile ($B' = B + N(0,1)$), based on noisy activity timeseries ($Y = XB + N(0,\sigma_Y)$; $Y' = X'B' + N(0,\sigma_Y)$). The first method was *pattern reliability* (that is, our EGA method), correlating the encoding profile estimated during training ($\hat{B} = X^{\dagger}Y$, † indicates pseudoinverse) with the encoding profile estimated during test ($\overline{B'} = X'^{\dagger}Y'$). The second method was *activity prediction* (that is, the traditional encoding approach), correlating the ground-truth test activity (Y') with the predicted test activity ($\overline{Y'} = X'\hat{B}$) after vectorizing both multivariate timeseries. To simulate the high measurement noise inherent to fMRI, we compared these methods under different levels of residual SD (σ_Y). **a)** Estimated pattern reliability tracked the true pattern reliability (that is, the true correlation between B' and B) across the full range of residual SD, with some attenuation at high levels of noise **b)** Unlike pattern reliability, activity prediction became much poorer as residual SD increased. **c)** Correlating the true pattern reliability (correlation between B and B') and estimated encoding strength (that is, pattern reliability or activity prediction), we found pattern reliability was better correlated with the true reliability, particularly at higher levels of noise. **d)** Both methods had similar performance in the absence of a signal ($B'_{null} = \mathcal{N}(0,1)$).

https://doi.org/10.1038/s41562-024-01826-7



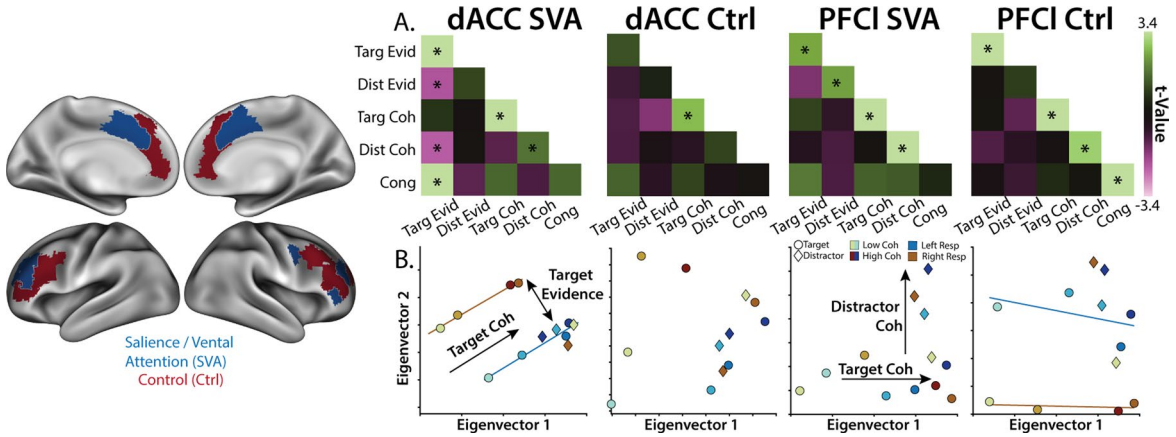
Extended Data Fig. 4 | Segregation Analysis. a) we used pattern component modelling[131] to simulate different candidate encoding profiles. 'Pure Selectivity' reflects the segregated encoding hypothesis, with different voxels (rows) encoding different features (columns). 'Mixed Selectivity' reflects the orthogonal subspace hypothesis, with the same voxels encoding both features. 'Sparse' models include non-selective voxels. **b)** By design, all of these encoding profiles had the same orthogonal encoding alignment (uncorrelated encoding weights), highlighting that this measure is unable to adjudicate between candidate encoding profiles. **c)** These models can be differentiated by correlating their absolute encoding weights, testing whether the sensitivity of a voxel to one feature is related to its sensitivity to the other feature, ignoring the direction of encoding. Pure selective encoding predicts a negative relationship, mixed selective encoding predicts no relationship, and sparse mixed selective encoding predicts a positive relationship. Similarity matrices averaged over 10,000 simulations. **d)** Correlating the absolute encoding weights, we found that the IPS profile was consistent with sparse mixed selective encoding.

https://doi.org/10.1038/s41562-024-01826-7



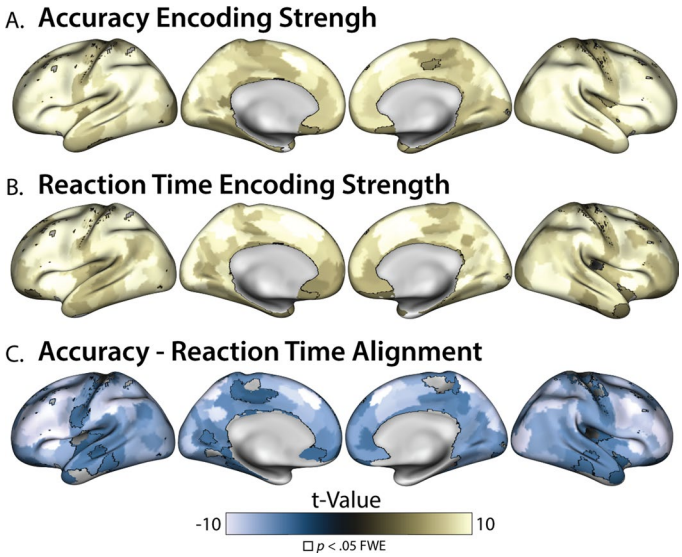
Extended Data Fig. 5 | 3D multidimensional scaling. The first three principal components of region-averaged condition similarity. Dark lines highlight the encoding geometry (connecting target coherence circles and showing the average direction for distractor coherence diamonds). Gray lines reflect the projection of these trends on different planes of the representational space. See legend and Fig. 4b for figure details. Note that in IPS, whereas targets and distractors are encoded orthogonally in the first two dimensions (floor), there appears to be some alignment in higher dimensions (right wall). In SPL, features appear to be aligned in all dimensions.

https://doi.org/10.1038/s41562-024-01826-7



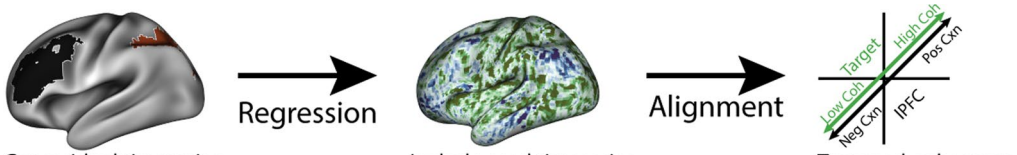
Extended Data Fig. 6 | Feature encoding in frontal networks. a) Similarity matrices for 'Salience / Ventral Attention (SVA)' and 'Control' networks in dACC and lPFC, correlating feature evidence ('Evid'), feature coherence ('Coh'), and feature congruence ('Cong'). Encoding strength on diagonal (right-tailed *p*-value), encoding alignment on off-diagonal (two-tailed *p*-value). **b)** Classical MDS embedding of target (circle) and distractor (diamond) representations at different levels of evidence. Colors denote responses, hues denote coherence. GLMs: A: Feature MV, B: Evidence Levels, see Table 1.

https://doi.org/10.1038/s41562-024-01826-7



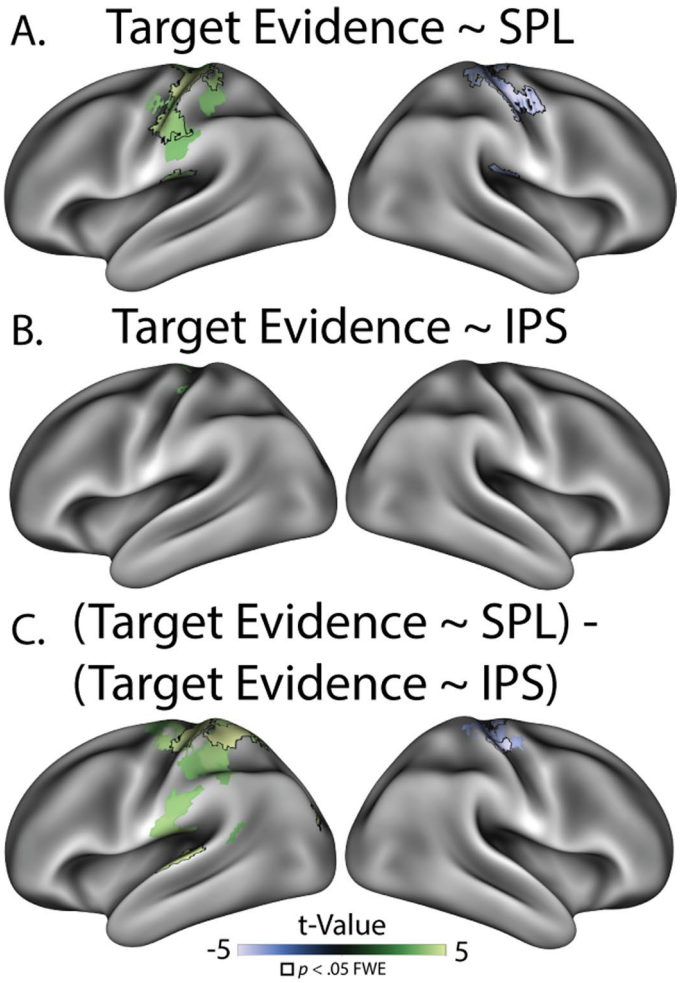
Extended Data Fig. 7 | Multivariate encoding of task performance. Encoding Strength (across-run reliability) for **a)** Accuracy and **b)** Reaction Time (B). **C)** Alignment between Accuracy and Reaction Time encoding. Outlined parcels are significant at $p < 0.05$ FWE (two-tailed max-statistic randomization test). Parcels in C are thresholded based on the reliability in A and B (both two-tailed $p < 0.001$ uncorrected). GLM: Performance.

https://doi.org/10.1038/s41562-024-01826-7



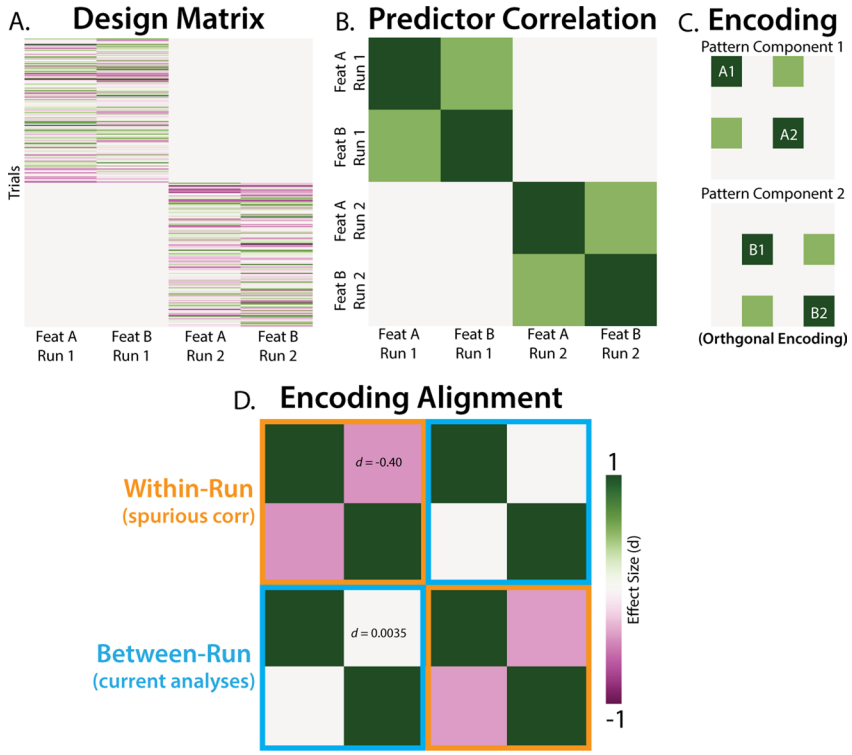
Extended Data Fig. 8 | Connectivity Alignment Schematic. We estimated connectivity encoding by getting the aggregated residual timeseries from our seed regions (eigenvariate; left), including these timeseries in our whole-brain GLM (middle), and then testing the alignment between connectivity encoding patterns and task encoding patterns (right).

https://doi.org/10.1038/s41562-024-01826-7



Extended Data Fig. 9 | SPL alignment with evidence encoding. a) Alignment between SPL activity and target evidence encoding. **b)** Alignment between IPS activity and target evidence encoding. **c)** Differences between SPL-evidence alignment and IPS-evidence alignment, showing stronger SPL connectivity. Note that target evidence encoding is signed according to the right-hand response (contralateral motor cortex should have a positive response). Colors reflect two-tailed $p < 0.001$ (uncorrected), outlines reflect $p < 0.05$ (corrected with two-tailed max-statistic randomization test).

https://doi.org/10.1038/s41562-024-01826-7



Extended Data Fig. 10 | Cross-validation prevents feature correlations from biasing alignment. We used pattern component modeling[131] to simulate neural data, testing whether feature correlations could spuriously create encoding alignment. **a)** Our design matrix had two simulate runs of two feature timeseries. **b)** Our features were correlated by design (that is, the columns of the design matrix were correlated). **c)** Despite correlation in the design matrix, these features were independently encoding in our simulated neural population (that is, in two distinct pattern components, which were each reliable across runs). **d)** Correlating our estimated encoding profiles, we found that within-run alignment (orange) had a spurious negative correlation (the opposite direction of the feature correlations). Critically, our analyses used between-run alignment (cyan), which avoids this biasing effect of feature correlations. Intuitively, since features are not correlated across runs (that is, they come from different trials), they do not produce spurious correlations. Effect sizes are computed across 10,000 simulations.

nature portfolio

Corresponding author(s): Harrison Ritz

Last updated by author(s): Jan 3, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	MATLAB 2020a; Psychtoolbox 3.0.15
Data analysis	MATLAB 2020a; fMRIPrep 20.2.6; SPM12 (v7771); rwls 4.1; PALM a119; rsatoolbox_matlab 1.0; bayesFactor 1.1; surfplot 0.1.0; ScientificColourMaps7

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Unprocessed fMRI data is available at https://doi.org/10.18112/openneuro.ds004909.v1.1.0. Behavioral data, event timing, and analysis code are available at: https://github.com/shenhavlab/PACT_fMRI_public.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	17 females and 12 males
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	Participants were recruited through the Brown University subject pool and local advertisements. Sample biases are not expected to influence our results.
Ethics oversight	This experiment was approved by Brown University's institutional review board (IRB approval code: 1606001539).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Our same size was chosen to match recent standards for similar task-based fMRI experiments (e.g., Danielmeier et al 2011: n=20, 336 trials; Jiang et al., 2018, n=22, 450 trials; Li et al., 2018: n=20, 585; Shenhav et al., 2018: n=34, 576 trials). Our experiment had n=29 and 1170 trials, due in large part to longer 90 min scanning sessions. Analyses were performed after data collection had completed.
Data exclusions	No data were excluded from the experiment (i.e., all runs from all participants), except from the trial regression filters (outlined in Methods).
Replication	Most analyses depend on cross-validated measure of multivariate encoding.
Randomization	Stimuli were randomized within participants, with the constraint of balanced target and distractor coherence levels across runs.
Blinding	Blinding was not necessary for within-subject analyses.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Plants

Seed stocks: N/A

Novel plant genotypes: N/A

Authentication: N/A

Magnetic resonance imaging

Experimental design

Design type: fast event-related design (main task), block design (stimulus localizers)

Design specifications: 12 blocks per subject. 6 Attend-Color: 150 trials, 6 Attend-Motion: 45 trials (interleaved). Trials were 1.5s long, with [1, 1.5, 2]s I.

Behavioral performance measures: Accuracy and reaction time were the primary behavioral measures; no performance criteria were used.

Acquisition

Imaging type(s): Functional

Field strength: 3.0

Sequence & imaging parameters: We used the following sequence parameters for our functional runs: field of view (FOV) = 211 mm x 211 mm (60 slices), voxel size = 2.4 mm, repetition time (TR) = 1.2 sec with interleaved multiband acquisitions (acceleration factor 4), echo time (TE) = 33 ms, and flip angle (FA) = 62". Slices were acquired anterior to posterior, with an auto-aligned slice orientation tilted 15° relative to the AC/PC plane. At the start of the imaging session, we collected a high-resolution structural MPRAGE with the following sequence parameters: FOV = 205 mm x 205 mm (192 slices), voxel size = 0.8 mm3, TR = 2.4 sec, TE1 = 1.86 ms, TE2 = 3.78 ms, TE3 = 5.7 ms, TE4 = 7.62, and FA = 7°. At the end of the scan, we collected a field map for susceptibility distortion correction (TR = 588ms, TE1 = 4.92 ms, TE2 = 7.38 ms, FA = 60").

Area of acquisition: whole brain scan

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software: We preprocessed our structural and functional data using fMRIprep (v20.2.6; (Esteban et al., 2019) based on the Nipype platform (Gorgolewski et al., 2011). We used FreeSurfer and ANTs to nonlinearly register structural Tiw images to the MNI152NLin6Asym template (resampling to 2mm). To preprocess functional 2w images, we applied susceptibility distortion correction using fMRIprep, co-registered our functional images to our T1w images using FreeSurfer, and slice-time corrected to the midpoint of the acquisition using AFNI. We then registered our images into MNI152NLin6Asym space using the transformation that ANTs computed for the Tiw images, resampling our functional images in a single step. For univariate analyses, we smoothed our functional images using a Gaussian kernel (8mm FWHM, as dACC responses often have a large spatial extent). For multivariate analyses, we worked in the unsmoothed template space.

Normalization: see above

Normalization template: see above

Noise and artifact removal: We included nuisance regressors to capture 1) run intercepts and 2) the average timeseries across white matter and CSF (as segmented by fMiRPrep). To reduce the influence of motion artifacts, we used robust weighted least-squares (Diedrichsen and Shadmehr, 2005; Jones et al., 2021), a procedure for optimally down-weighting noisy TRs.

Volume censoring: Analyses were masked by the Kong22 atlas for localization & parcel-based analysis.

Statistical modeling & inference

Model type and settings: We estimated contrast maps at the subject-level, which we then used for one-sample t-tests at the group-level. We controlled for family-wise error rate using threshold-free cluster enhancement (Smith and Nichols, 2009), testing whether voxels have an unlikely degree of clustering under a randomized null distribution (Implemented in PALM (Winkler et al.,

2014); 10,000 randomizations). To improve the specificity of our coverage (e.g., reducing white-matter contributions) and to facilitate our inference about functional networks (see below), we limited these analyses to voxels within the Kong2022 whole-brain parcellation (Kong et al., 2021; Schaefer et al., 2018). Surface renders were generated using surfplot (Gale et al., 2021; Vos de Wael et al., 2020), projecting from MNI space to the Human Connectome Project's fLR space (164,000 vertices).

Effect(s) tested: The primary effects of interest in our standard GLM were target coherence, distractor coherence, distractor congruence, response-coded target coherence, response-coded distractor coherence.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☒ Both

Anatomical location(s): Kong22 400 parcellation

Statistic type for inference: see above

(See Eklund et al. 2016)

Correction: All corrections were done using randomization testing (TFCE for univariate or max-stat for mulivariate)

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis: We adapted tunctions trom the pem-toolbox and ratoolbox packages for our multivariate analyses (Diedrichsen et al., 2018; Nili et al., 2014). We first fit whole-brain GLMs without spatial smoothing, separately for each scanner run. These GLMs estimated the parametric relationship between task variables and BOLD response (e.g., linearly coded target coherence), with a pattern of these parametric betas across voxels reflecting linear encoding subspace (Kriegeskorte and Diedrichsen, 2019). Within each Schaefer parcel (n=400, we spatially pre-wnitened these dera maps, reducing noise correlations between voxels that can inflate pattern similarity and reduce reliability (Walther et al., 2016). We then computed the cross-validated Pearson correlation, estimating the similarity of whitened patterns across scanner runs. We used a correlation metric to estimate the alignment between encoding subspaces, rather than distances between condition patterns, to normalize biases and scaling across stimuli (e.g., greater sensitivity to targets vs distractors) and across time (e.g., representational drift). We found convergent results when using (un-centered) cosine similarity, suggesting that our results were not trivially due to parcels univariate response, but a correlation metric had the best reliability across runs.