# Orthogonal neural encoding of targets and distractors supports multivariate cognitive control

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

*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*

* *Corresponding author:* hritz@princeton.edu

# Abstract

People can overcome a wide array of mental challenges by coordinating their neural information processing to align with their goals. Recent behavioral work has shown that people can independently control their attention across multiple features during perceptual decision-making, but the structure of the neural representations that enables this multivariate control remains mysterious. We hypothesized that the brain solves this complex coordination problem by orthogonalizing feature-specific representations of task demands and attentional priority, allowing the brain to independently monitor and adjust multiple streams of stimulus information. To test this hypothesis, we measured fMRI activity while participants performed a task designed to tag processing and control over feature-specific information that is task-relevant (targets) versus task-irrelevant (distractors). We then characterized the geometry of these neural representations using a novel multivariate analysis (Encoding Geometry Analysis), estimating where the encoding of different task features is correlated versus orthogonal. We identified feature-specific representations of task demands and attentional priority in the dorsal anterior cingulate cortex (dACC) and intraparietal sulcus (IPS), respectively, consistent with differential roles for these regions in monitoring versus directing information processing. Representations of attentional priority in IPS were fully mediated by the control requirements of the task, associated with behavioral performance, and depended on connectivity with nodes in the frontoparietal control network, suggesting that these representations serve a fundamental role in supporting attentional control. Together, these findings provide evidence for a neural geometry that can enable coordinated control over multiple sources of information.

Keywords: cognitive control, attention, decision-making, fMRI

# Introduction

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 goal pursuit (Badre et al., 2021; Danielmeier and Ullsperger, 2011; Egner, 2008; Ritz et al., 2022a). For example, someone may respond to a mistake by becoming more cautious, enhancing task-relevant processing, or suppressing task-irrelevant processing (Danielmeier and Ullsperger, 2011), and previous work has shown that people simultaneously deploy multiple such strategies at the same time in response to different task demands (Danielmeier et al., 2011; Fischer et al., 2018; Leng et al., 2021; Ritz and Shenhav, 2021). While the breadth of these control adjustments and the conditions under which they occur have become increasingly clear, how we achieve this level of coordination remains largely mysterious. In particular, it is unclear how people monitor and direct simultaneous control signals over multiple parallel streams of information. Here, we seek to fill this gap by combining recent developments in experimentally 'tagging' information processing streams during cognitive control (Flesch et al., 2022; Kayser et al., 2010b; Ritz and Shenhav, 2021) with emerging analytic methods for quantifying representational geometry (Bernardi et al., 2020; Ebitz et al., 2020; Libby and Buschman, 2021). We hypothesized this multivariate control depends on independent neural representations, in the form of orthogonal encoding subspaces, that track multiple sources of difficulty and selectively adjusts the attentional priority across multiple task features.

Previous work has proposed that independent neural representations play an important role in cognitive control, but have largely examined how these representations minimize interference between tasks. When tasks are in conflict, the brain uses orthogonal task representations to minimize cross-talk (Flesch et al., 2022; Kaufman et al., 2014; Mante et al., 2013; Minxha et al., 2020; Pagan et al., 2022; Panichello and Buschman, 2021; Salinas, 2004), consistent with the optimal strategy in artificial neural networks (Flesch et al., 2022; Mante et al., 2013; Musslick et al., 2020). A compelling possibility is that the cognitive control system uses a similar representational format to coordinate multiple control signals *within* a task as well (Ebitz et al., 2020; Libby and Buschman, 2021; Rust and Cohen, 2022). A large body of work has shown that cognitive control networks encode multiple task parameters (Flesch et al., 2022; Freund et al., 2021; Jackson et al., 2017, 2021; Kayser et al., 2010b; Vermeylen et al., 2020; Woolgar et al., 2011, 2015b, 2015a), and 'global' measures of cognitive control like overall difficulty or effort (Freund et al., 2021; Kragel et al., 2018; Smith et al., 2019; Vermeylen et al., 2019). However, little is known about whether different control parameters are encoded independently from one another, which would allow the brain to simultaneously coordinate multiple forms of goal-directed task processing.

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 decision

and control processes (Ritz and Shenhav, 2021). Building on prior work (Danielmeier et al., 2011; Kayser et al., 2010b; Mante et al., 2013), this task incorporates elements of perceptual decision-making (discrimination of a target feature) and inhibitory control (overcoming a salient and prepotent distractor). We have shown that we can separately tag target and distractor processing in participants' performance on this task, and that target and distractor processing are independently controlled (Ritz and Shenhav, 2021). In conjunction with this process-tagging approach, our second innovation was to develop a novel multivariate fMRI analysis for measuring relationships between neural feature representations (*encoding geometry*). By combining the strengths of multivariate encoding analyses and representation similarity analyses into a method we refer to as 'Encoding Geometry Analysis' (EGA), we can characterize when and where the brain has independent representations of how targets and distractors contribute to task difficulty, and how these different features are prioritized by top-down attention.

In brief, we found that key nodes within the cognitive control network use orthogonal representations of target and distractor information to support cognitive control. In the dorsal anterior cingulate cortex (dACC), encoding of target and distractor difficulty was spatially separated, arranged along a rostrocaudal gradient. In the intraparietal sulcus (IPS), encoding of target and distractor stimulus strength was spatially overlapping, but with orthogonal encoding profiles. These regional distinctions are consistent with hypothesized roles in the planning and implementation of (multivariate) attentional policies (Gottlieb et al., 2020; Shenhav et al., 2013). Furthermore, we found that task representations depended on task automaticity, task performance, and frontoparietal connectivity, consistent with these representations playing a critical role in cognitive control. Together, these results suggest that cognitive control uses representational formats that allow the brain to control multiple forms of information processing.

# Results

## Task overview

Human participants performed the Parametric Attentional Control Task (PACT; (Ritz and Shenhav, 2021) during fMRI. On each trial, participants responded to an array of colored moving dots (colored random dot kinematogram; Figure 1a). In the critical condition (Attend-Color), participants respond with a keypress based on which of two colors were in the majority. In alternating scanner runs, participants instead responded based on motion (Attend-Motion), which was designed to be less control-demanding due to the (Simon-like) congruence between motion direction and response direction (Danielmeier et al., 2011; Ritz and Shenhav, 2021). Across trials, we independently and parametrically manipulated target and distractor information across five levels of target coherence (e.g., % of dots in the majority color) and distractor congruence (e.g., % of dots moving either in the congruent or incongruent direction relative to the correct color response; Figure 1b). This task allowed us to 'tag' participants' sensitivity to each

dimension by measuring behavioral and neural responses to independently manipulated target and distractor features.

## Behavior

Participants had overall good performance on the task, with a high level of accuracy (median Accuracy = 89%, IQR = [84% - 92%]), and a low rate of missed responses (median lapse rate = 2%, IQR = [0% - 5%]). We used mixed effects regressions to characterize how target coherence and distractor congruence influenced participants' accuracy and log-transformed correct reaction times. Replicating previous behavioral findings using this task, participants were sensitive to both target and distractor information (Ritz and Shenhav, 2021). When target coherence was weaker, participants responded slower ($t_{(27.6)} = 16.1$, $p = 1.60 \times 10^{-15}$) and less accurately ($t_{(28)} = -8.90$, $p = 1.19 \times 10^{-9}$; Figure 1c). When distractors were more incongruent, participants also responded slower ($t_{(28.8)} = 5.09$, $p = 2.15 \times 10^{-5}$) and less accurately ($t_{(28)} = -4.66$, $p = 6.99 \times 10^{-5}$; Figure 1d). Also replicating prior findings with this task, interactions between targets and distractors were not significant for reaction time ($t_{(28.2)} = 0.143$, $p = .887$) and had a weak influence on accuracy ($t_{(28)} = 2.36$, $p = .0257$), with model omitting target-distractor interactions providing a better complexity-penalized fit (RT $\Delta$AIC = 17.7, Accuracy $\Delta$AIC = 1.38).

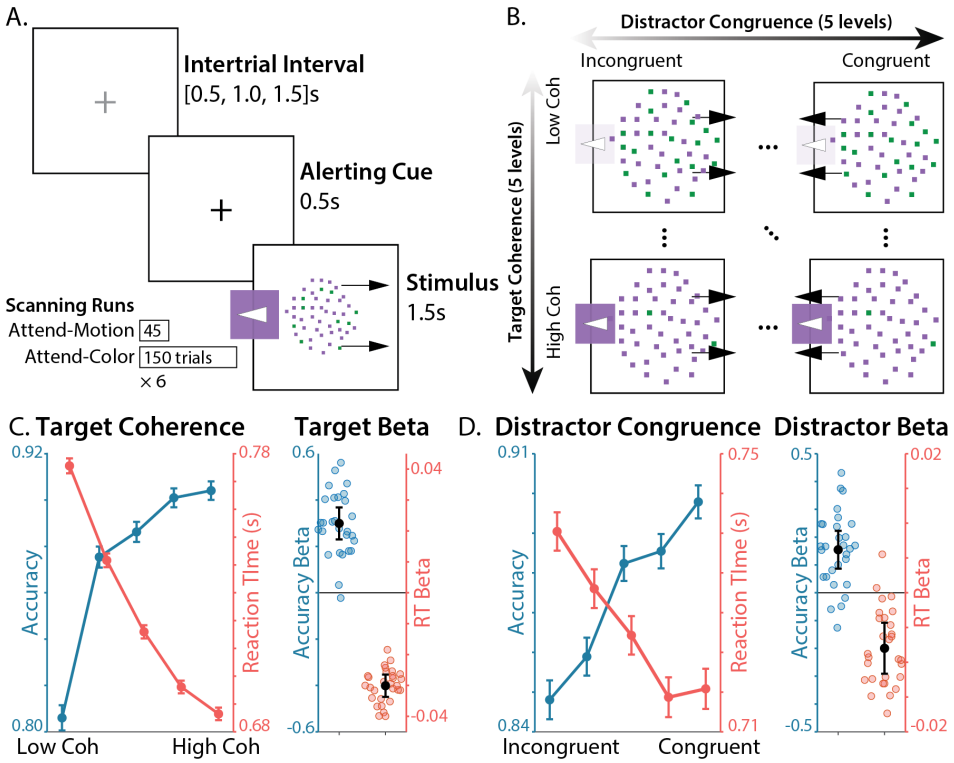


**Figure 1.** *Task and Behavior.* **A)** Participants responded to a color-motion random dot kinematogram (RDK) with a button press. Participants either responded to the left/right motion direction of the RDK (Attend-Motion runs) or based on the majority color (Attend-Color runs; critical condition). **B)** We parametrically and independently manipulated target coherence (% of dots in the majority color) and distractor congruence (signed motion coherence relative to the target response). **C)** Participants were faster and more accurate when the target was more coherent. **D)**

Participants were faster and more accurate when the distractor was more congruent with the target. Error bars on line plots reflect within-participant SEM, error bars on regression fixed-effect betas reflect 95% CI.

## Distinct coding of target- and distractor-related control demands in dACC

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). Our task allowed us to test whether these two sources of increasing control demand are tracked within common regions of 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 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 dACC. In caudal dACC, we found significant clusters encoding the parametric effect of target difficulty (Figure 2a; negative effect of target coherence in green), and in more rostral dACC we found clusters encoding parametric distractor incongruence (negative effect of distractor congruence in blue). These analyses control for omission errors, and additionally controlling for commission errors produced the same whole-brain pattern at a reduced threshold (see Supplementary Figure 1, see also Figure 4e for convergent multivariate analyses). As additional evidence of their dissociable encoding with dACC, we further found that the spatial patterns of target and distractor regression weights were uncorrelated within dACC ($t_{(28.0)} = 1.32$, $p = .197$, logBF = -0.363).

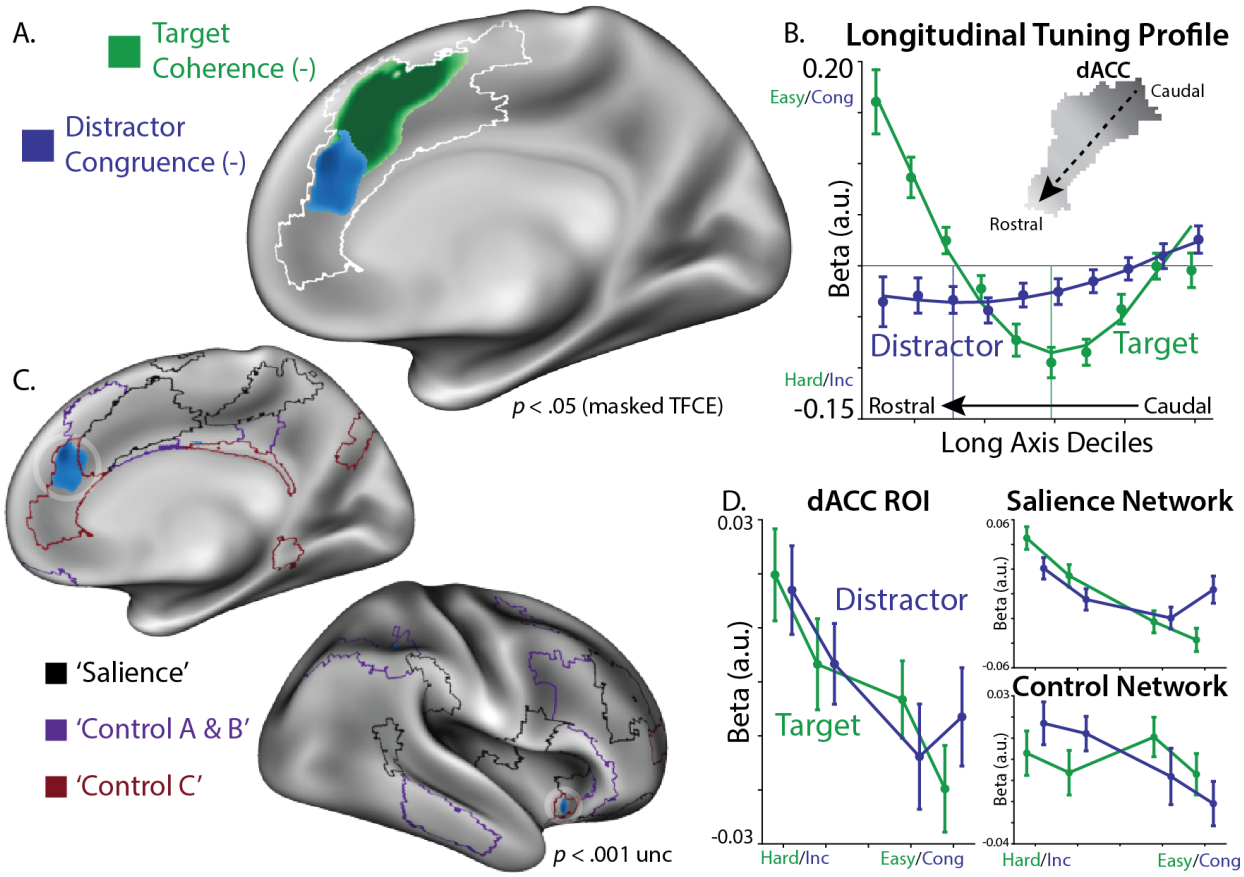


**Figure 2.** *Distinct coding of target and distractor difficulty in dACC.* **A)** We looked for linear target coherence and distractor congruence signals within an a priori dACC mask (white outline; overlapping Kong22 parcels and medial 'cognitive control' Neurosynth mask). We found that voxels in the most caudal dACC reflected target difficulty (green), more rostral voxels reflected distractor incongruence (blue). Statistical tests are corrected using non-parametric threshold-free cluster enhancement. **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. Fit lines are the quantized predictions from a second-order polynomial regression. We used these regression betas to estimate the minima for target and distractor tuning (i.e., location of strongest difficulty effects), finding that the target difficulty peak (vertical green line) was more caudal than the distractor incongruence peak (vertical blue line). **C)** Plotting the uncorrected whole-brain response, distractor incongruence responses (blue) were strongest within the 'Control C' sub-network (red), both in dACC and anterior insula. **D)** BOLD responses across levels of target coherence and distractor congruence, plotted within the whole dACC ROI (left), or the salience network and control network parcels within the dACC ROI (right).

To further quantify how feature encoding changed along the longitudinal axis of dACC, we used principal component analysis to extract the axis position of dACC voxels (see 'dACC longitudinal axis analyses' in Methods), and then regressed target and distractor beta weights onto these axis scores. We found that targets had stronger difficulty coding in more caudal voxels ($t_{(27.9)} = 3.74$, $p = .000840$), with a quadratic trend ($t_{(26.5)} = 4.48$, $p = .000129$). In line with previous work on both perceptual and value-based decision-making (Clairis and Pessiglione, 2020; Fleming et al., 2018; Shenhav et al., 2016a, 2016b), 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 ROI (Supplementary Figure 2). In contrast to targets, distractors had stronger incongruence coding in more rostral voxel ($t_{(28.0)}$ = -3.26, $p$ = .00294), 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, $p$ = .0159). Target and distractor minima were uncorrelated across subjects ($r_{(27)}$ = .0282, $p$ = .880, logBF = -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 dACC occurred at the boundary between two well-characterized functional networks (Kong et al., 2021; Schaefer et al., 2018; Yeo et al., 2011). Whereas distractor-related demands were more strongly encoded rostrally in the Control Network (particularly within regions of dACC and insula corresponding to the 'Control C' Sub-Network; (Kong et al., 2021)), target-related demands were more strongly encoded caudally within the 'Salience' Network (Figure 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 ($\Delta$BIC > 1675).

## Orthogonal encoding of target and distractor coherence in intraparietal sulcus

We found that dACC appeared to dissociably encode target and distractor difficulty, consistent with a role in monitoring different task demands and/or specifying different control signals (Shenhav et al., 2013). To identify neural mechanisms that potentially execute control towards these different task features (i.e., that enable the prioritization attention to targets versus distractors), we next tested for regions that encode the strength of target and distractor information. In particular, we sought to examine where in the brain these targets and distractors shared a common neural code (e.g., as a global index of spatial salience) and where these features are encoded distinctly (e.g., as separate targets of control).

An initial whole-brain univariate analysis showed that overlapping regions throughout occipital, parietal, and prefrontal cortices track the overall strength (i.e., unsigned coherence) of both target and distractor information (Figure 3a; conjunction in orange). 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, $p$ = .048; Accuracy: $t_{(28)}$ = -0.845, $p$ = .406), suggesting that these neural responses are unlikely to reflect task difficulty per se.

While these activations point towards widespread and coarsely overlapping encoding of the salience of these two features, they lack information about whether those features are encoded similarly or differently 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 Encoding Geometry Analysis (EGA). Whereas pattern similarity analyses typically quantify relationships between representations of specific stimuli or responses (e.g., whether they could be classified, (Kriegeskorte and Diedrichsen, 2019)), EGA characterizes relationships between encoding subspaces (patterns of contrast weights) across different task features, consistent with recent analyses trends in systems neuroscience (Bernardi et al., 2020; Cohen and Maunsell, 2010; Ebitz et al., 2020; Flesch et al., 2022; Kimmel et al., 2020; Libby and Buschman, 2021). A stronger correlation between encoding subspaces (either positive or negative) indicates that features are similarly encoded (e.g., confusable by a decoder; Figure 3b), whereas weak correlation indicate that these representations are orthogonal (and thus distinguishable by a decoder; (Kriegeskorte and Diedrichsen, 2019)). Unlike standard pattern similarity, the sign of these relationships is interpretable in EGA, reflecting how features are coded relative to one another. We estimated this encoding alignment within each parcel, correlating unsmoothed and spatially pre-whitened patterns of parametric regression betas across scanner runs to minimize spatiotemporal autocorrelation (Diedrichsen and Kriegeskorte, 2017; Nili et al., 2014; Walther et al., 2016).

Focusing on regions that encoded both target and distractor information (parcels where both group-level $p < .001$), EGA revealed clear dissociations between regions that represent these features in alignment versus orthogonally. Within visual cortex and the superior parietal lobule (SPL), target and distractor representations demonstrated significant negative correlations (Figure 3C, blue), suggesting aligned encoding. In contrast, early visual cortex and intraparietal sulcus (IPS) demonstrated target-distractor correlations near zero (Figure 3C, black), suggesting orthogonal encoding.

To bolster our interpretation of the latter findings as reflecting orthogonal (i.e., uncorrelated) representations rather than merely small but non-significant correlations, we employed Bayesian t-tests at the group level to estimate the relative likelihood that these encoding dimensions were orthogonal or correlated. Consistent with our previous analyses, we found strong evidence for correlation (positive log bayes factors) in more medial regions of occipital and posterior parietal cortex (e.g., SPL), and strong evidence for orthogonality (negative log bayes factors) in more lateral regions of occipital and posterior parietal cortex (e.g., IPS; Figure 3D). Additional analyses buttressed this account, demonstrating that coherence orthogonality in IPS is not due to differences in encoding reliability, as a similar topography was observed with disattenuated correlations (normalizing correlations by their reliability; see Supplementary Figure 3), nor due to the choice of Bayes factor priors (see Supplementary Figure 4).

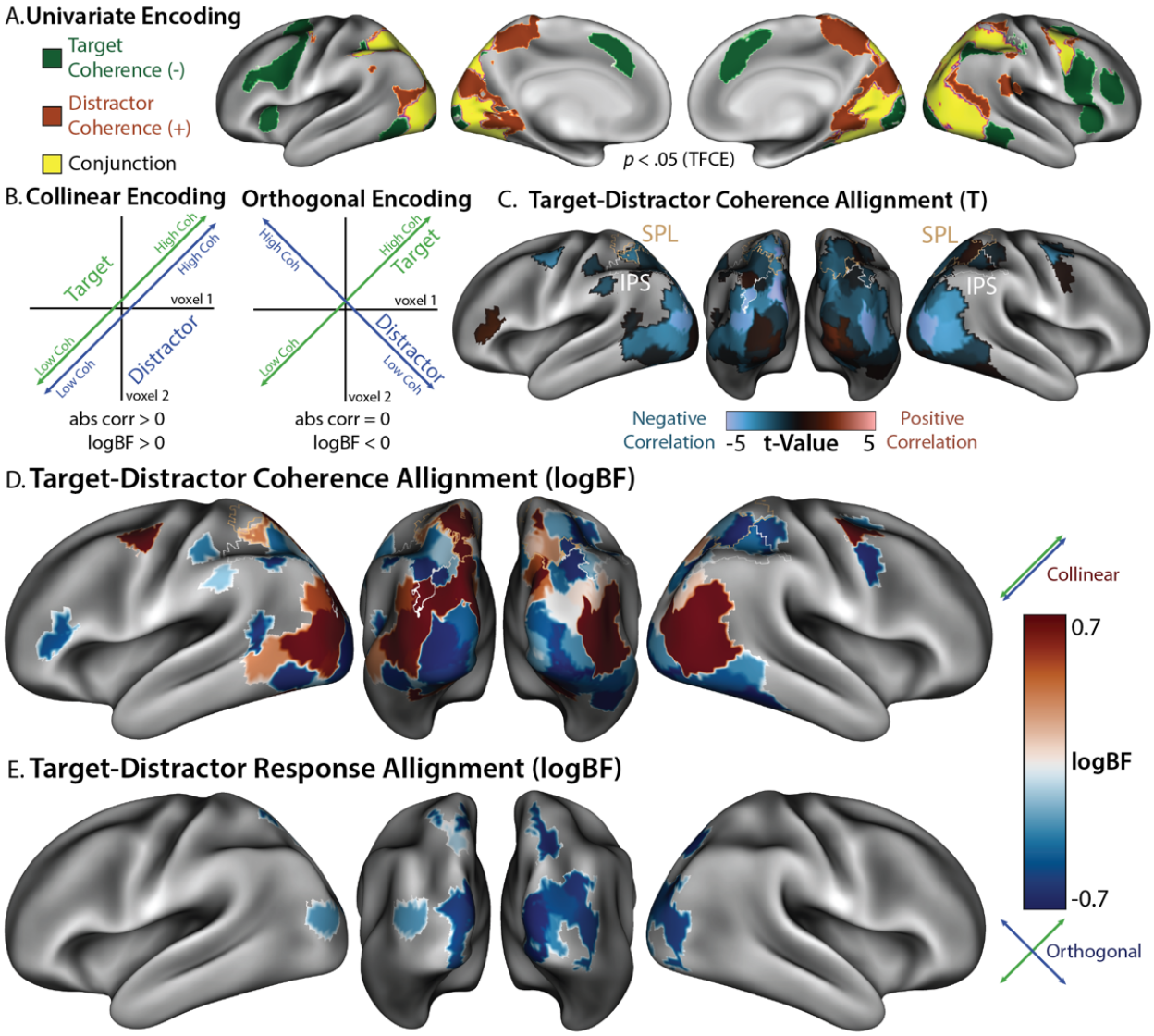


**Figure 3.** *Encoding Geometry Analysis (EGA) dissociates target and distractor encoding.* **A)** Parametric univariate responses to weak target coherence (green), strong distractor coherence (orange), and their conjunction (yellow). Statistical tests are corrected for multiple comparisons using non-parametric threshold-free cluster enhancement (TFCE). **B)** We quantified encoding alignment using encoding geometry analyses (EGA), correlating beta maps of parametric target and distractor coherence effects (cross-validated). Positive or negative correlations (i.e., log bayes factors > 0) reflect correlated representation (e.g., allow for cross-decoding). Correlations near zero (i.e., log bayes factors < 0) reflect orthogonal representations. **C)** Encoding alignment within parcels in which target and distractor encoding was jointly reliable (both $p < .001$ uncorrected). Representations were negatively correlated within Superior Parietal Lobule (SPL in gold; Kong22 labels), and uncorrelated within Intraparietal Sulcus (IPS in white; Kong22 labels). **D)** Bayesian analyses provide explicit evidence for orthogonality within IPS (i.e., negative BF; theoretical minima: -0.71). **E)** Coherence coded in terms of response (i.e., supporting a left vs right choice). Target and distractor response encoding overlapped in visual cortex and SPL and was represented orthogonally.

These results focus on the salience of information available at the time of stimulus presentation, for instance demonstrating that SPL exhibits aligned representations of target and distractor salience. Past decision-making research has separately demonstrated that SPL tracks the amount

of information stimuli provide in support of a given response (e.g., responding left vs. right; (Hunt et al., 2012; Kayser et al., 2010a, 2010b). We found that this was also true for our task. In addition to encoding the salience (unsigned coherence) of targets and distractors, SPL and visual cortex also tracked the decision evidence (response-signed coherence) provided by those same features (Figure 3e). EGA revealed that response features were represented orthogonally, in parcels with correlated coherence representations (compare Figure 3d and 3e), consistent with previous observations of multiple decision-related signals in SPL (Hunt et al., 2012).

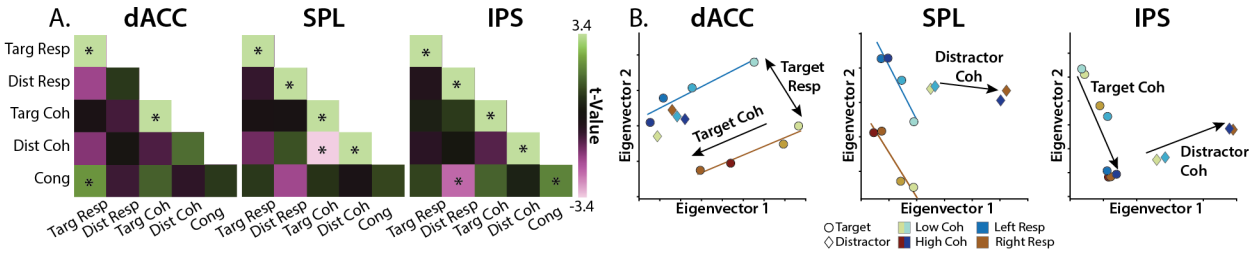


**Figure 4.** *Region-specific feature encoding.* **A)** Similarity matrices for dACC, SPL, and IPS, correlating feature response, feature coherence, and feature congruence. Encoding strength on diagonal (right-tailed *p*-value), encoding alignment on off-diagonal (two-tailed *p*-value). **B)** Classical MDS embedding of target (circle) and distractor (diamond) representations at different levels of response evidence. Colors denote responses, hues denote coherence.

We complemented our whole-brain analyses with ROI analyses in areas exhibiting reliable encoding of key variables (dACC, SPL, and IPS). Consistent with our analyses above, we found that target and distractor coherence encoding was aligned in SPL, but not in IPS (Figure 4A, compare to Figure 3d). We again found that target and distractor response evidence (signed coherence) were encoded in SPL. Directly comparing these regions, we found stronger encoding of target responses in SPL, stronger encoding of target coherence in IPS, and stronger alignment between target-distractor coherence alignment in SPL. Unlike our univariate results, we did not find congruence encoding in dACC (though this was found in IPS). Instead, dACC showed multivariate encoding of target difficulty and response.

| **Task Feature** | **Contrast (SPL - IPS)** |
|---|---|
| Target Response | $t_{(28)} = 3.89$, $p = .000562$ |
| Distractor Response | $t_{(28)} = 0.896$, $p = .378$ |
| Target-Distractor Response Alignment | $t_{(28)} = -0.145$, $p = .886$ |
| Target Coherence | $t_{(28)} = -3.89$, $p = 9.36 \times 10^{-9}$ |
| Distractor Coherence | $t_{(28)} = 1.40$, $p = .170$ |
| Target-Distractor Coherence Alignment | $t_{(28)} = -2.99$, $p = .00580$ |

**Table 1.** *Feature encoding contrasted across parietal cortex.* Differences in encoding (within-predictor reliability) and alignment (between-predictor correlation) between SPL and IPS.

Though not evident in the parcel-level analysis, our ROI analyses revealed that response evidence was also encoded in IPS. An interesting explanation for the discrepancy between these analyses emerged when examining the relationship between the target- and distractor-related response encoding across these parcels (e.g., testing whether a parcel with stronger target response encoding also had stronger distractor response encoding). Across SPL parcels, target and distractor response encoding was significantly positively correlated (participant-bootstrapped correlation, 95% CI [.015, .56]), despite orthogonal representations in this ROI. In contrast, target and distractor encoding was not significantly correlated across parcels in IPS (95% CI [-.41, .070]; SPL-IPS difference: 95% CI [.10, .85]), helping to explain the discrepancy between parcel-level and ROI-level encoding.

To further characterize how feature coherence and response evidence are encoded across these regions, we performed multidimensional scaling over their representations (Figure 4b; (Diedrichsen and Kriegeskorte, 2017; Kriegeskorte et al., 2006)). Briefly, this method allows us to visualize – in a non-parametric manner – the relationships between representations of different feature levels (e.g., levels of target salience), by estimating each feature level separately within a GLM and then using singular value decomposition to project these patterns into a 2D space (see Methods for additional details). We find that coherence and response axes naturally emerge in the top two principal components in this analysis within dACC, SPL, and IPS. Coherence axes (light to dark shading) are parallel between left (blue) and right (brown) responses, suggesting a response-independent encoding. In these components, response encoding appeared to be binary, in contrast to parametric coherence encoding (we found similar whole-brain encoding maps for binary-coded responses; see Supplementary Figure 5). Critically, whereas coherence encoding axes within SPL was aligned between targets (circles) and distractors (diamonds; confirming correlated encoding), in IPS these representations form perpendicular lines (confirming orthogonal encoding).

## Task demands dissociate coherence and response encoding

Our findings thus far demonstrate two sets of dissociations within and across brain regions. In dACC, we find that distinct regions encode the control demands related to discriminating targets (caudal dACC) versus overcoming distractor incongruence (rostral dACC). In posterior parietal cortex, we find that overlapping regions track the overall salience 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 was 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 encoding and control, we examined how the encoding of feature coherence differed between matched task that placed stronger or weaker demands on cognitive control. So far, our analyses have focused on conditions in which participants needed to respond to the color feature while

ignoring the motion feature (Attend-Color task), but on alternating scanner runs participants instead responded to the motion dimension and ignored the color 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-Color, as the left/right motion directions are compatible with the left/right response directions (i.e., Simon facilitation; (Ritz and Shenhav, 2021)). Comparing these tasks allows us to disambiguate bottom-up attentional salience from the top-down contributions to attentional priority (Woolgar et al., 2015b, 2015a, 2011).

Consistent with previous work (Ritz and Shenhav, 2021), performance on the Attend-Motion task was better overall (mean RT: 565ms vs 725ms, sign-rank $p = 2.56 \times 10^{-6}$; mean Accuracy: 93.7% vs 87.5%, sign-rank $p = .000318$). Unlike the Attend-Color task, performance was not impaired by distractor incongruence (i.e., color distractors; RT: $t_{(28)} = -1.39$, $p = .176$; Accuracy: $t_{(28)} = 0.674$, $p = .506$). To investigate these task-dependent feature representations, we fit a GLM that included both tasks. To control for performance differences across tasks, we only analyzed accurate trials and included trial-wise RT as a nuisance covariate, concatenating RT across tasks.

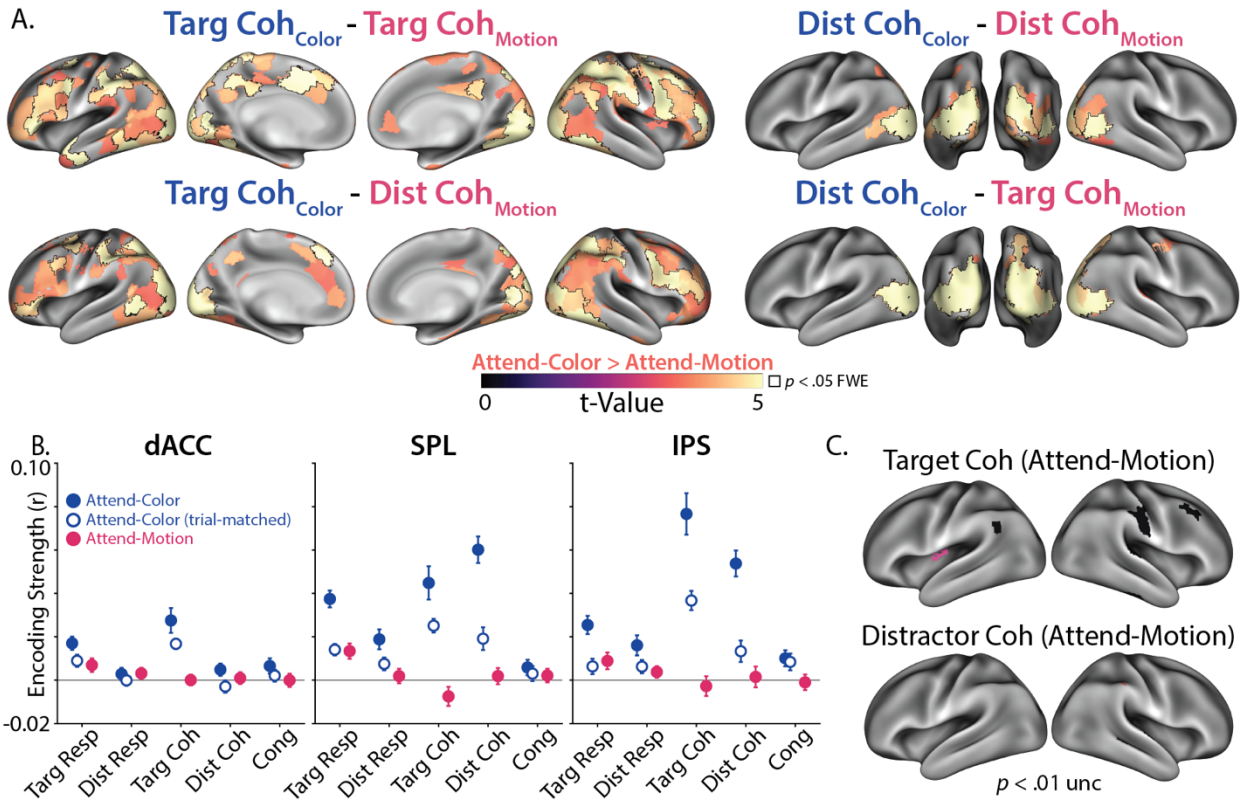


**Figure 5.** *Task-dependent encoding strength.* **A)** Across cortex, feature coherence encoding was stronger during Attend-Color than Attend-Motion, matched for the same number of trials. Attend-Color had stronger encoding for target coherence (top left), distractor coherence (top right), color coherence (bottom left) and motion coherence (bottom right). Parcels are thresholded at $p < .001$ (uncorrected), outlined parcels are significant at $p < .05$ FWE (max-statistic randomization test across all parcels). **B)** Target and distractor coherence information was encoded

more strongly during Attend-Color than Attend-Motion in dACC, SPL and IPS. Attend-Color encoding plotted from the whole sample (blue fill) and a trial-matched sample (first 45 trials of each run; white fill) In Attend-Motion runs, only target response was significantly encoded (magenta). **C)** Target and distractor coherence was not reliably encoded during the Attend-Motion task (liberally thresholded at $p < .01$ uncorrected).

In contrast to the widespread encoding of both motion and color coherence that we observed during the Attend-Color task (Figure 3), encoding was consistently stronger during the Attend-Color relative to the Attend-Motion task (Figure 5A), consistent with a role for cognitive control. Coherence encoding was stronger during Attend-Color whether classifying according to goal-relevance (target vs. distractor) or the features themselves (motion vs. color), and was present both whole-brain (Figure 5A) and within task-relevant ROIs (Figure 5B). Notably, robust coherence encoding was absent when participants were performing the Attend-Motion task (Figure 5C).

In contrast to these stark task-related differences in coherence encoding, we found that neural encoding of target response information (response-signed color coherence in the Attend-Color task and response-signed motion coherence in the Attend-Motion task) was preserved across tasks, including within dACC, SPL, and IPS (Figure 5B). Consistent with previous experiments examining context-dependent decision-making (Aoi et al., 2020; Flesch et al., 2022; Kayser et al., 2010b; Mante et al., 2013; Pagan et al., 2022; Takagi et al., 2021), we found stronger target response encoding relative to distractor response encoding, in our case in the response-encoding SPL (Attend-Color: $t_{(28)} = 4.26$, one-tailed $p = 0.0001$; Attend-Motion: $t_{(28)} = 2.37$, one-tailed $p = 0.0124$). We also found that target response encoding during Attend-Motion was aligned with Attend-Color, both for *motion* response encoding ('stimulus axis'; SPL: one-tailed $p = .0236$, IPS: one-tailed $p = .0166$) and *target* response encoding ('decision axis'; SPL: one-tailed $p = 1.29 \times 10^{-6}$; IPS: one-tailed $p = .0005$), again in agreement with these previous experiments. Whereas our experiment replicates previous observations of the neural representations supporting contextual decision-making, we now extended these findings to understand how attentional priority signals are encoded in response to the asymmetrical response inference that is characteristic of cognitive control (Miller and Cohen, 2001).

## Aligned encoding dimensions for feature coherence and task performance

We next explored whether the encoding of feature coherence, seemingly in the service of cognitive control, was related to how well participants performed the task. We tested this question by determining whether feature coherence representations were aligned with representations of behavior (i.e., alignment between stimulus and behavioral subspaces (Stringer et al., 2019)). Specifically, we included trial-level reaction time and accuracy in our first-level GLMs, and tested how target and distractor encoding was aligned with task performance encoding.

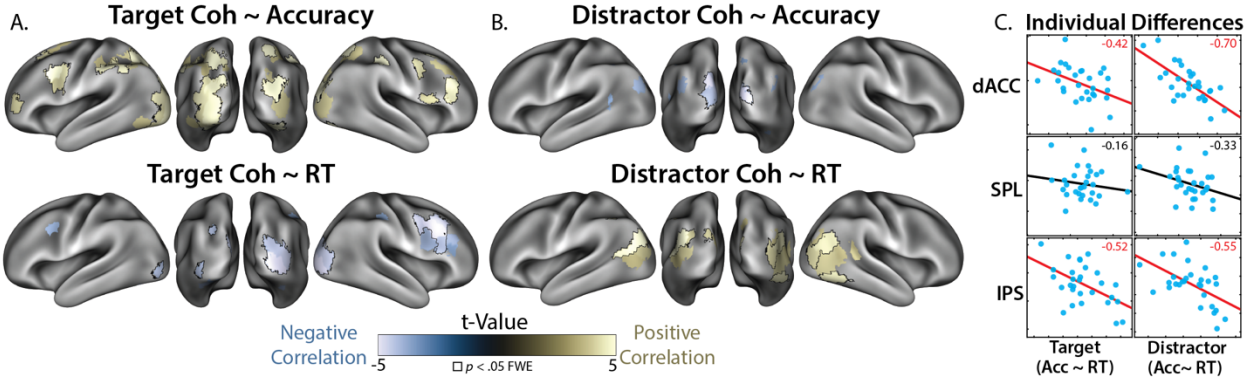


**Figure 6.** *Alignment between feature and performance encoding.* **A)** Alignment between encoding of target coherence and performance (top row: Accuracy, bottom row: RT). **B)** Alignment between encoding of distractor coherence and performance (top row: Accuracy, bottom row: RT). Across A and B, parcels are thresholded at $p <$ .001 (uncorrected, in jointly reliable parcels), and outlined parcels are significant at $p < .05$ FWE (max-statistic randomization test across jointly reliable parcels). **C)** Individual differences in feature-RT alignment correlated with feature-accuracy alignment across regions (correlation values in top right; $p < .05$ in red).

We found that the encoding of target and distractor coherence was aligned with performance across frontoparietal and visual regions (Figure 6a-b). If a regions' encoding of target coherence reflects how sensitive the participant was to target information on that trial, 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 visual and frontoparietal cortex: target encoding was aligned with better performance (faster RTs and higher accuracy; Figure 6a), whereas distractor encoding was aligned with worse performance (slower RTs and lower accuracy; Figure 6b).

Finally, we examined whether performance-coherence alignment reflected individual differences in participants' task performance in our main task-related ROIs (see Figures 3-4). In particular, we tested whether the alignment between features and behavior reflects specific relationships with speed or accuracy, or whether they reflected overall increases in evidence accumulation (e.g., faster responding and higher accuracy). Within each ROI, we correlated feature-RT alignment with feature-accuracy alignment across subjects. We found that in dACC and IPS, participants showed the negative correlation between performance alignment measures predicted by an increase in processing speed (Figure 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*s > .10), again consistent with feature-specific processing. These findings suggest that feature coherence are related to processing efficiency, again supporting the importance of coherence representations in adaptive control.

## Coherence encoding aligns with prefrontal connectivity through IPS

Cortical encoding of target and distractor coherence depended on task demands and was aligned with performance across prefrontal and posterior cortex. Since this widespread encoding of task information likely reflects distributed network involvement in cognitive control (Corbetta and Shulman, 2002; Goldman-Rakic, 1988; Miller and Cohen, 2001), we sought to understand how frontal and parietal systems interact during task performance. We focused our analyses on IPS and lateral PFC (lPFC), linking the core parietal site of orthogonal coherence encoding (IPS) to an prefrontal site previous work suggests provides top-down feedback during cognitive control (Goldman-Rakic, 1988; Kastner and Ungerleider, 2000; Suzuki and Gottlieb, 2013; Yantis and Serences, 2003). Previous work has found that IPS attentional biases lower-level stimulus encoding in visual cortices (Kay and Yeatman, 2017; Saalmann et al., 2007), and that IPS mediates directed connectivity between lPFC and visual cortex during perceptual decision-making (Kayser et al., 2010b). Here, we extended these experiments to test how IPS mediates the relationship between prefrontal feedback and stimulus encoding.

To investigate these cortical interactions, we developed a novel multivariate connectivity analysis to test whether coherence encoding was aligned with prefrontal connectivity, and whether this coherence-connectivity relationship was mediated by parietal cortex. We first estimated the voxel-averaged residual timeseries in lPFC (SPM12's eigenvariate), and then included this residual timeseries alongside task predictors in a whole-brain regression analysis (Figure 7A). Next, we used EGA to test whether there was alignment between patterns encoding lPFC functional connectivity (i.e., betas from the residual timeseries predictor) and patterns encoding target and distractor coherence. Finally, we compared regression estimates between a model that included lPFC only ('solo' model) to a model that included both lPFC and IPS ('both' model). Comparing the strength of lPFC-coherence alignment with and without IPS is a test of whether parietal cortex mediates lPFC-coherence alignment (MacKinnon et al., 2007).

We found that lPFC connectivity patterns were aligned with coherence-encoding patterns in visual cortex (Figure 7B). 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 (MacKinnon et al., 2007).

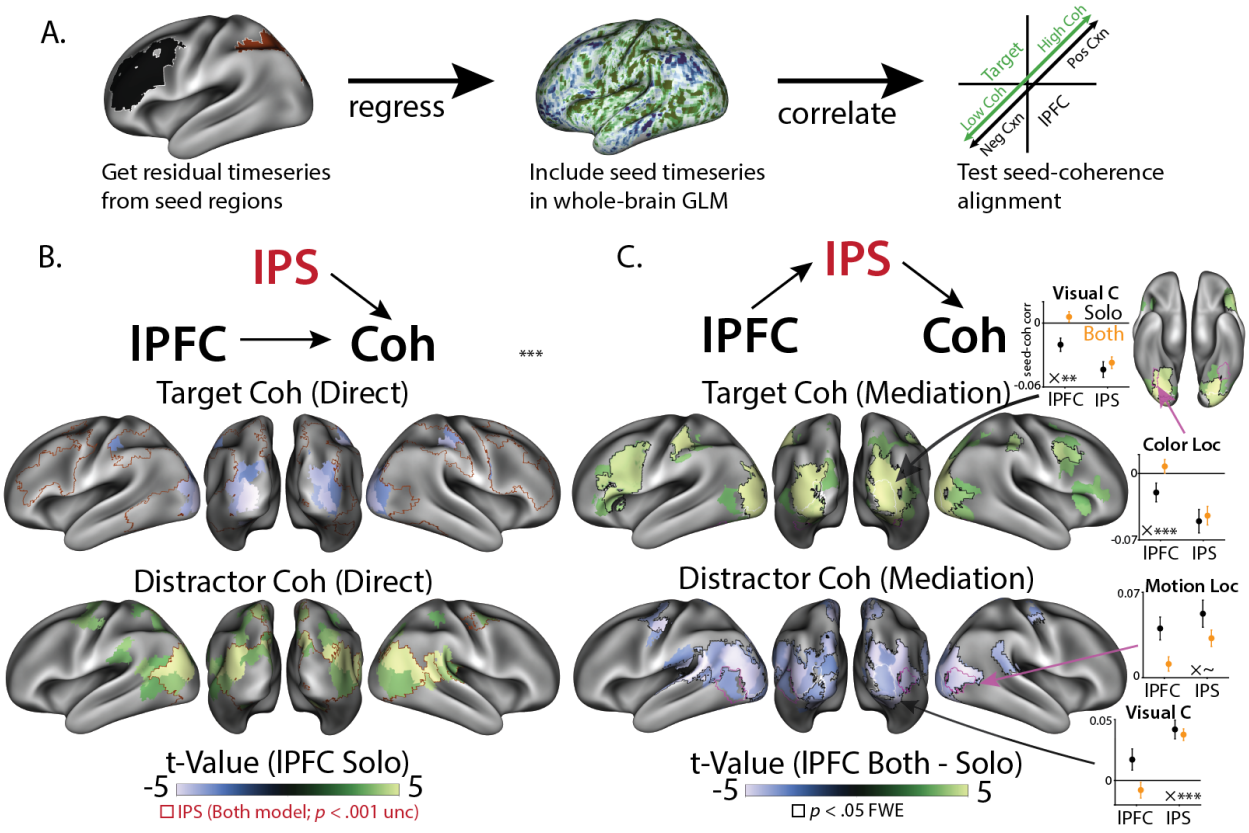


**Figure 7.** *IPS mediates alignment between lPFC and feature encoding.* **A)** We estimated connectivity encoding by getting aggregated residual timeseries from our seed regions (eigenvariate; top), including these timeseries in our GLM (middle), and then testing the alignment between connectivity and feature patterns. **B)** Connectivity patterns from lPFC (colormap) and IPS (red outline) were aligned with target and distractor coherence patterns ($p < .001$ uncorrected, in jointly reliable parcels). IPS effects are outlined to show overlap, with all effects in a consistent direction to lPFC. **C)** lPFC-feature alignment contrasted between lPFC-only model ('Solo') and lPFC + IPS model ('Both'). Including IPS reduced the alignment between lPFC and feature encoding (compare the sign of the main effect in B to the contrast in C). Parcels are thresholded at $p < .001$ (uncorrected, jointly reliable parcels), and outlined parcels are significant at $p < .05$ FWE (max-statistic randomization test across jointly reliable parcels). Insets graphs: seed-coherence alignment in Solo models (black) and Both model (orange) across visual cortex. 'Visual C' is defined by our parcellation (Kong et al., 2021), whereas Color and Motion localizers are parcels near the peak response identified during feature localizer runs (see Methods). In all areas, lPFC alignment was more affected by IPS than IPS alignment was affected by lPFC (~ $p < .10$, * $p < .01$, *** $p < .001$).

Our critical test was whether IPS mediated the relationship between lPFC connectivity and coherence encoding. We found that this mediation was strongest in early visual cortex, where the alignment between lPFC and feature coherence was reduced in a model that included IPS relative to a model without IPS (Figure 7C). The negatively correlated target-lPFC relationship became more positive when IPS was included (top), and the positively correlated distractor-PFC relationship became more negative when IPS was included (bottom). Critically, we found that IPS reduced prefrontal-coherence alignment in early visual cortex more than lPFC reduced parietal-coherence alignment (Figure 7C inset; Supplementary Figure 6), consistent with frontal-

to-parietal directed connectivity in previous research (Kayser et al., 2010b; Suzuki and Gottlieb, 2013). The opposite relationship, lPFC mediation of IPS connectivity, appeared in higher-level visual cortex for distractor coherence (Supplementary Figure 6), though these effects were not reliable in explicit contrasts and may reflect projections from both regions.

# Discussion

In classical Cybernetics, the study of control in animals and machines, the 'Law of Requisite Variety' states that an effective controller should be at least as complex as the system it aims to control (Ashby, 1961). In this experiment, we explored whether neural control systems follow this law by having representations with the same dimensionality as the processes they regulate (Badre et al., 2021). Consistent with behavioral evidence that participants can independently control their sensitivity to task-relevant and task-conflicting information (Ritz and Shenhav, 2021), we found that key nodes of canonical cognitive control networks use orthogonal neural representations of targets and distractors. Within dACC, orthogonal representations of target and distractor difficulty arose from coarse spatial encoding along a rostrocaudal axis. Within IPS, orthogonal representations of target and distractor coherence were present in finer-grained spatial patterns. Feature coherence representations were distinct from choice-related signals, and depended on task automaticity, task performance, and frontoparietal connectivity. Together, these results reveal a neural mechanism for how cognitive control prioritizes multiple streams of information during decision-making.

Neurocomputational theories have proposed that dACC is involved in planning control across multiple levels of abstraction (Holroyd and McClure, 2015; Holroyd and Yeung, 2011; Shenhav et al., 2013; Vassena et al., 2017). Past work has found that control abstraction is hierarchically organized along dACC's rostrocaudal axis, with more caudal dACC involved in lower-level action control, and more rostral dACC involved in higher-level strategy control (Shenhav et al., 2018; Taren et al., 2011; Venkatraman et al., 2009; Zarr and Brown, 2016), an organization that may reflect a more general hierarchy of abstraction within PFC (Badre and D'Esposito, 2009; Badre and Nee, 2018; Koechlin and Summerfield, 2007; Taren et al., 2011). Consistent with this account, we found that caudal dACC tracked the coherence of the target and distractor dimensions, especially within the 'salience' 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 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 (Badre and D'Esposito, 2009), the latter of which may require additional disengagement from distractor-dependent attentional capture. Across levels of hierarchy, dACC demonstrated independent representations of target and distractor information (see also Figure 4a), a requirement for coordinating control across multiple feature processing streams.

Interestingly, despite finding robust univariate encoding of distractor congruence in dACC, we did not find corresponding encoding of congruence in our multivariate analyses within this region. It is possible that this discrepancy reflects differences in the functional form or the spatial smoothness of the underlying signals (which these two methods would be differentially sensitive to), and/or that the size of our congruence effect simply wasn't large enough to be reliably detected by our multivariate analyses (thus appearing instead as a weak correlation). Whatever the case may be, the univariate effects are at the very least consistent with a much wider literature documenting congruence effects within this region.

Whereas dACC encoded the difficulty of task features (e.g., distractor incongruence), in parietal cortex we found overlapping representations of feature strength (e.g., distractor coherence). In 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 (Esterman et al., 2009; Greenberg et al., 2010; Molenberghs et al., 2007; Serences et al., 2004; Serences and Yantis, 2007; Yantis et al., 2002). In contrast, IPS had orthogonal representations of feature coherence, consistent with selective prioritization of task-relevant information (Adam and Serences, 2021; Greenberg et al., 2010; Jackson et al., 2017; Kay and Yeatman, 2017; Molenberghs et al., 2007; Serences and Yantis, 2007; Suzuki and Gottlieb, 2013; Woolgar et al., 2015b, 2015a, 2011; Yantis et al., 2002). Our previous work has demonstrated behavioral evidence for independent control over target and distractor attentional priority in this task (Ritz and Shenhav, 2021), with different task variables selectively enhancing target or distractor sensitivity (see also (Egner, 2008; Soutschek et al., 2015)). Orthogonal feature representation in IPS may offer a mechanism for this feature-selective control, consistent with theoretical accounts of IPS implementing a priority map that combines stimulus- or value-dependent salience with goal-dependent feedback from PFC (Bisley and Goldberg, 2010; Corbetta and Shulman, 2002; Gottlieb et al., 2020; Yantis and Serences, 2003).

We further explored whether these coherence representations depended on cognitive control by comparing Attend-Color and Attend-Motion tasks, which have been shown to differ dramatically in their control demands (Ritz and Shenhav, 2021). As in previous work, task performance was much better in Attend-Motion runs than Attend-Color runs, and participants were not sensitive to color distractors. Consistent with previous work on context-dependent decision-making, response-coded feature information had similarly strong encoding across tasks, with generalizable encoding dimensions for choice and motion directions (Flesch et al., 2022; Mante et al., 2013; Takagi et al., 2021). In contrast to these decision representations, we found that coherence representations disappeared in the easier Attend-Motion task. This observation is consistent with previous experiments finding that feature decoding is stronger for more difficult tasks (Rust and Cohen, 2022; Woolgar et al., 2011, 2015b, 2015a) or when people are incentivized to use cognitive control (Etzel et al., 2016; Hall-McMaster et al., 2019). Moreover,

stimuli and responses were matched across tasks, helping to rule out alternative accounts of coherence encoding based on 'bottom-up' stimulus salience, decision-making, or eye movements. Instead, difficulty-dependent coherence encoding may reflect the involvement of an attention control system that can separately regulate target and distractor processing. Supporting this account, difficulty-dependent coherence representations were aligned with performance representations, with individual differences in feature-performance alignment consistent with adjustments to processing fluency.

Classic models of prefrontal involvement in cognitive control (Desimone and Duncan, 1995; Kastner and Ungerleider, 2000; Miller and Cohen, 2001) propose that prefrontal cortex biases information processing in sensory regions depending on task goals. In line with this macro-scale organization, we found that coherence encoding in visual cortex was related to functional connectivity with the frontoparietal control network. In particular, coherence encoding in visual cortex was aligned with patterns of functional connectivity to lateral prefrontal cortex, and this feature-connectivity relationship was mediated by IPS. The results of this novel multivariate connectivity analysis are consistent with previous research supporting a role for IPS in top-down control of visual encoding (Kay and Yeatman, 2017; Lauritzen et al., 2009; Saalmann et al., 2007), as well as a granger-causal PFC-IPS-visual pathway during distractor decision-making (Kayser et al., 2010b). Here, we demonstrate stable 'communication subspaces' between visual cortex and PFC (Semedo et al., 2019; Srinath et al., 2021), which can plausibly communicate feedforward information about coherence or feedback adjustments to feature gain. Critically, our findings are consistent with IPS, a critical site for orthogonal feature representations in our experiment, playing a central role in linking prefrontal cortex with early perceptual processing.

This experiment provides 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 (Pylyshyn and Storm, 1988; Vul et al., 2009), including parametric relationships between attentional load and IPS activity (Culham et al., 2001, 1998; Howe et al., 2009; Jovicich et al., 2001; Ritz et al., 2022b), little is known about how the brain coordinates multiple control signals (Ritz et al., 2022a). Future experiments should further elaborate on this frontoparietal control circuit, interrogating how incentives influence different task representations (Etzel et al., 2016; Hall-McMaster et al., 2019; Parro et al., 2018; Peck et al., 2009; Wisniewski et al., 2015), and how neural and behavioral indices of control causally depend on perturbations of neural activity (Jackson et al., 2021). Future experiments should also use temporally-resolved neural recording technologies like (i)EEG or (OP-)MEG to better understand the within-trial dynamics of multivariate control (Ritz and Shenhav, 2021; Weichart et al., 2020). 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: M = 21.2, SD = 3.4) participated in this experiment. All participants had self-reported normal color vision and no history of neurological disorders. Two participants missed one Attend-Color 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. Participants provided informed consent, in accordance with Brown University's institutional review board.

## Task

The main task closely followed our previously reported behavioral experiment (Ritz and Shenhav, 2021). On each trial, participants saw a random dot kinematogram (RDK) against a black background. This RDK consisted of colored dots that moved left or right, and participants responded to the stimulus with button presses using their left or right thumbs.

In Attend-Color blocks (six blocks of 150 trials), participants responded depending on which color was in the majority. Two colors were mapped to each response (four colors total), and dots were a mixture of one color from each possible response. Dots colors 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), participants responded based on the dot motion instead of the dot color. Dot motion consisted of a mixture between dots moving coherently (either left or right) and dots moving in a random direction. Attend-Motion blocks were shorter because they acted to reinforce motion sensitivity and provide a test of stimulus-dependent effects.

Critically, dots always had color and motion, and we varied the strength of color coherence (% of dots in the majority) and motion coherence (% of dots moving coherently) across trials. Our previous experiments have found that in Attend-Color blocks, participants are still influenced by motion information, introducing a response conflict when color and motion are associated with different responses (Ritz and Shenhav, 2021). Target coherence (e.g., color coherence during Attend-Color) was linearly spaced between 65% and 95% with 5 levels, and distractor congruence (signed coherence relative to the target response) was linearly spaced between -95% and 95% with 5 levels. In order to increase the salience of the motion dimension relative to the color dimension, the display was large (~10 degrees of visual angle) and dots moved quickly (~10 degrees of visual angle per second).

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 reaction time and visual stimulation (the fixation cross changed from white to gray to indicate the response). The inter-trial interval was uniformly sampled from 1.0, 1.5, or 2.0 seconds. This ITI was relatively short in order to maximize the behavioral effect, and because efficiency simulations showed that it increased power to detect parametric effects of target and distractor coherence (e.g., relative to a more standard 5 second ITI). The fixation cross changed from gray to white for the last 0.5 seconds before the stimulus to provide an alerting cue.

## Procedure

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

Following the twelve blocks of the scanner task, participants underwent localizers for color and motion, based on the tasks used in our previous experiments (Shenhav et al., 2018). Both localizers were block designs, alternating between 16 seconds of feature present and 16 seconds of feature absent for seven cycles. For the color localizer, participants saw an aperture the same size as the task, either filled with colored squares that were resampled every second during stimulus-on ('Mondrian stimulus'), or luminance-matched gray squares that were similarly resampled during stimulus-off. For the motion localizer, participants saw white dots that were moving with full coherence in a different direction every second during stimulus-on, or still dots for stimulus-off. No responses were required during the localizers.

## MRI sequence

We scanned 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 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 × 205 mm (192 slices), voxel size = 0.8 $mm^3$, 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°).

## fMRI preprocessing

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 T1w images to the MNI152NLin6Asym template (resampling to 2mm). 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 (8mm FWHM, 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 (v7771) for our univariate general linear model (GLM) analyses. Due to high trial-to-trial collinearity from to our short ITIs, we performed all analyses across trials, rather than extracting single-trial estimates. Our regression models used whole trials as events (i.e., a 1.5 second boxcar aligned to the stimulus onset). We parametrically modulated these events with standardized trial-level predictors (e.g., linear-coded target coherence, or contrast-coded errors), and then convolved these predictors with SPM's canonical HRF, concatenating our voxel timeseries across runs. 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.

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 fsLR space (164,000 vertices).

## 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 cingulate cortex and preSMA. This mask was constructed by 1) getting a meta-analytic mask of cingulate responses co-occurring with 'cognitive control' (Neurosynth uniformity test; (Yarkoni et al., 2011), and taking any parcels from the whole-brain Schaefer parcellation (400 parcels; (Kong et al., 2021; Schaefer et al., 2018) that had a 50 voxel overlap with this meta-analytic mask. We used this parcellation because it provided more selective gray matter coverage than the Neurosynth mask alone and it allowed us to categorize voxels membership in putative functional networks.

To characterize the spatial organization within 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 voxel's 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 subjects and accommodate within-subject correlations between voxels. Our model predicted gradient score from the linear and quadratic expansions of the target and distractor betas (gradientScore ~ 1 + target + target$^2$ + distractor + distractor$^2$ + (1 + target + target$^2$ + distractor + distractor$^2$ | subject)). To characterize the network-dependent organization of target and distractor encoding, we complexity-penalized fits between models that either 1) predicted target or distractor betas from linear and quadratic expansions of gradient scores, or 2) predicted target/distractor betas 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 (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-whitened these beta 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. Note that this analysis approach is related to 'Parallelism Scores' (Bernardi et al., 2020), but here we use parametric encoding models and emphasize not only deviations from parallel/orthogonal, but also the signed alignment between features (e.g., Figures 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 correlations before analysis. We used a Bayesian t-test to test for orthogonality (bayesFactor toolbox in MATLAB, based on (Rouder et al., 2012)). The Bayes factor 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 logBF = -0.705 (i.e., the Bayes factor when $t_{(28)} = 0$).

Our measure of encoding strength was whether encoding subspaces were reliable across blocks (i.e., 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 provides 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 timeseries (Kriegeskorte and Diedrichsen, 2019). 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; see Supplementary Figure 7).

When looking at alignment between two subspaces across parcels, we first selected parcels that significantly encoded both factors ('jointly reliable parcels', both $p < .001$ uncorrected). This selection process acts as a thresholded version of classical correlation disattenuation (Spearman, 1987; Thornton and Mitchell, 2017), and we confirmed through simulations this selection procedure does not increase type 1 error rate. We corrected for multiple comparisions using non-parametric max-statistic randomization tests across parcels (Nichols and Holmes, 2002). These randomization tests determine the likelihood of an observed effects 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 max and min group-level effect sizes

across all parcels, estimating the strongest parcel-wise effect we'd expect if there wasn't a 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 (assessed using colintest in MATLAB; (Belsley et al., 1980), but we wanted to confirm that predictor correlations wouldn't bias our estimates of encoding alignment. We simulated data from a pattern component model (Diedrichsen et al., 2018) 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.

## Multivariate Connectivity Analysis

To estimate what information is plausibly communicated between cortical areas, we measured the alignment between multivariate connectivity patterns (i.e., the 'communication subspace' with a seed region, (Semedo et al., 2019)) and local feature encoding patterns. First, we residualized our Performance GLM (see Table 2) from a seed region's timeseries, and then extracted the variance-weighted average timecourse (i.e., the leading eigenvariate from SPM12's volume of interest function). We then re-estimated our Performance GLM, including the block-specific seed timeseries as a covariate, and performed EGA between seed and coherence patterns. We found convergent results when we residualized a quadratic expansion of our Performance GLM from our seed region, helping to confirm that connectivity alignment wasn't 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 color- and motion-sensitive cortex by finding the bilateral Kong22 parcels that roughly covered the area of strongest block-level contrast during our localizer runs.

To estimate the mediation of lPFC connectivity by IPS, we compared models in which just lPFC or just IPS were used for EGA against a model where both seeds were included as covariates in the same model (MacKinnon et al., 2007). Our test of mediation was the group-level difference in lPFC seed-coherence alignment before and after including IPS. While these analyses are inherently cross-sectional (i.e., lPFC and IPS are measured at the same time), we supplemented these analyses by showing that the mediating effect of IPS on lPFC was much larger than the mediating effect of lPFC on IPS (see Figure 7c; Supplementary Figure 6).

| **Model Name** | **Trial selection** | **Predictors (z-scored)** |
|---|---|---|
| Feature (univariate) | No omission errors; block-concatenated | target coherence, distractor coherence, distractor congruence; response-coded target coherence, response-coded distractor coherence; omission errors (run-concatenated) |
| Feature (multivariate) | No errors; block-separated | target coherence, distractor coherence, distractor congruence; response-coded target coherence, response-coded distractor coherence; errors (run-concatenated) |
| Performance (multivariate) | No omission errors; block-separated | target coherence, distractor coherence, response-coded target coherence, response-coded distractor coherence, reaction time, accuracy; omission errors (run-concatenated) |
| Between-Task (multivariate) | No errors; block-separated | target coherence, distractor coherence, distractor congruence; response-coded target coherence, response-coded distractor coherence; errors (run-concatenated); reaction time (run-concatenated) |

**Table 2.** *fMRI models*. First-level general linear models used for univariate and multivariate fMRI analyses.

**Acknowledgements:** This work was supported NIH grants S10OD025181 and R01MH124849 (A.S.), NSF CAREER Award 2046111 (A.S.), as well as the C.V. Starr Postdoctoral Fellowship (H.R.). We are grateful to Joonhwa Kim for her assistance in data collection, and to Michael J. Frank, Matthew N. Nassar, Jonathan Cohen, Michael Esterman, Romy Frömer, Jörn Diedrichsen, Apoorva Bhandari, Debbie Yee, Sam Nastase, and the Shenhav Lab for helpful discussions.

**Conflicts of Interest:** None

**Data Availability:** Data and analysis scripts will be made available upon publication.

# Supplementary Figures

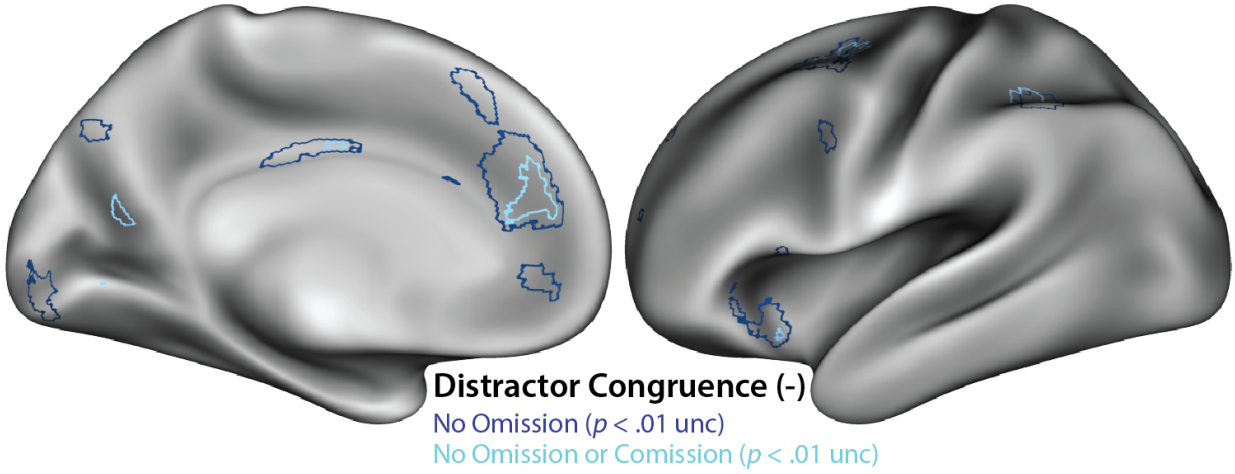


**Figure S1.** *Error control analysis.* Distractor congruence effect when controlling for different types of errors. Our primary analysis only analyzed trials without omission errors (navy), here plotted at a liberal uncorrected threshold. When we analyze trials without omission errors and commission 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.

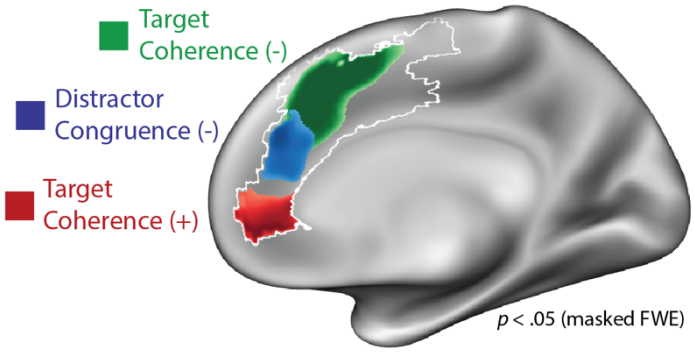


**Figure S2.** *Target ease.* Parametric effects of target coherence and distractor congruence, showing the rostral effect of target ease (positive relationship with target coherence) in red.

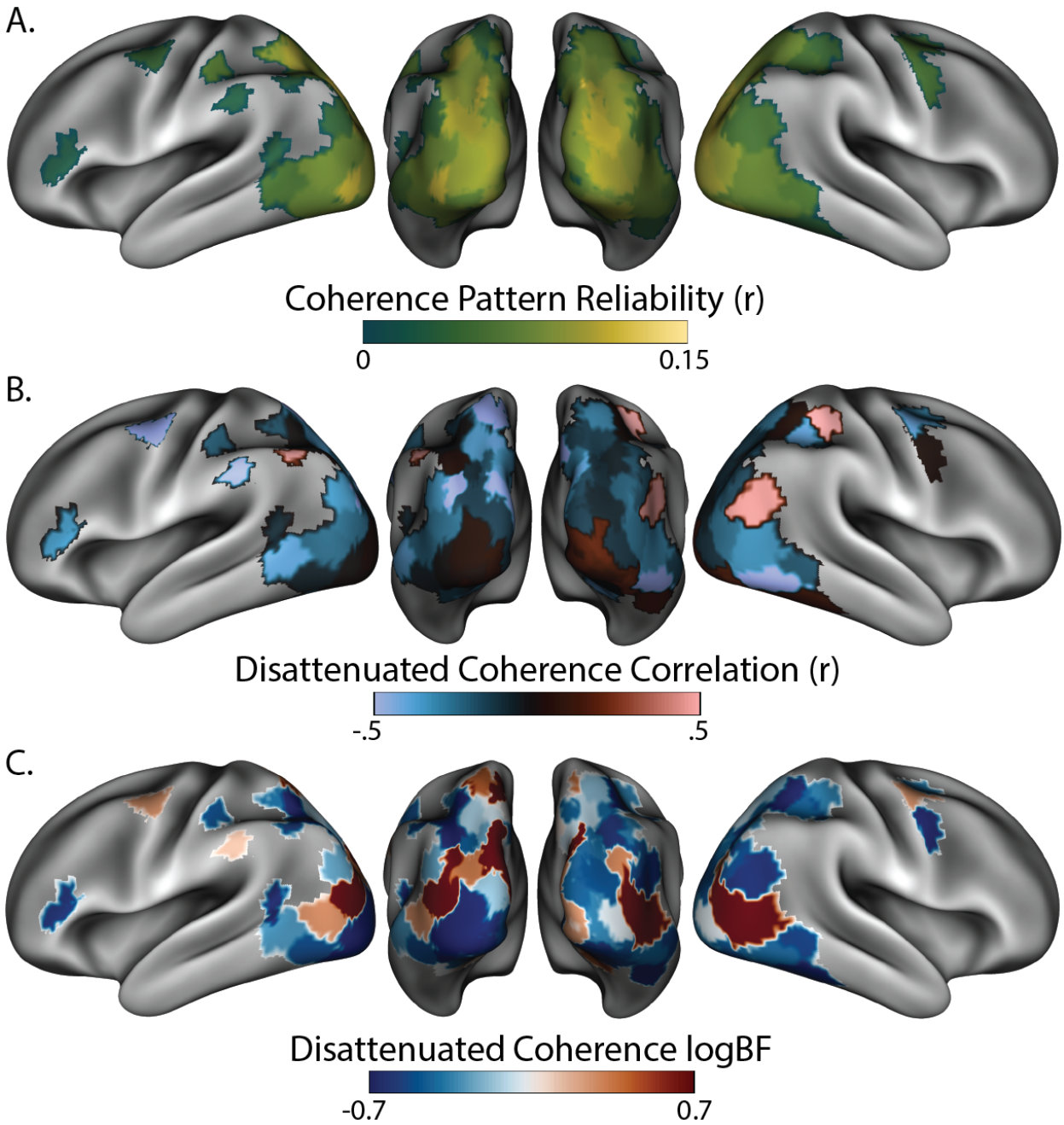


**Figure S3.** *Reliability control analysis.* **A)** Geometric mean of target and distractor coherence reliability ($\sqrt{r_{targ} \times r_{dist}}$), plotted in the reliability-thresholded parcels used in Figure 4. Reliability provides the theoretical upper bound on correlation strength. Median across participants, excluding participants with non-positive reliability. **B)** Target-distractor correlations, normalized by target-distractor reliability (i.e., disattenuated correlations) **C)** Log bayes factors for disattenuated target-distractor correlations. Compare to Figure 4C.

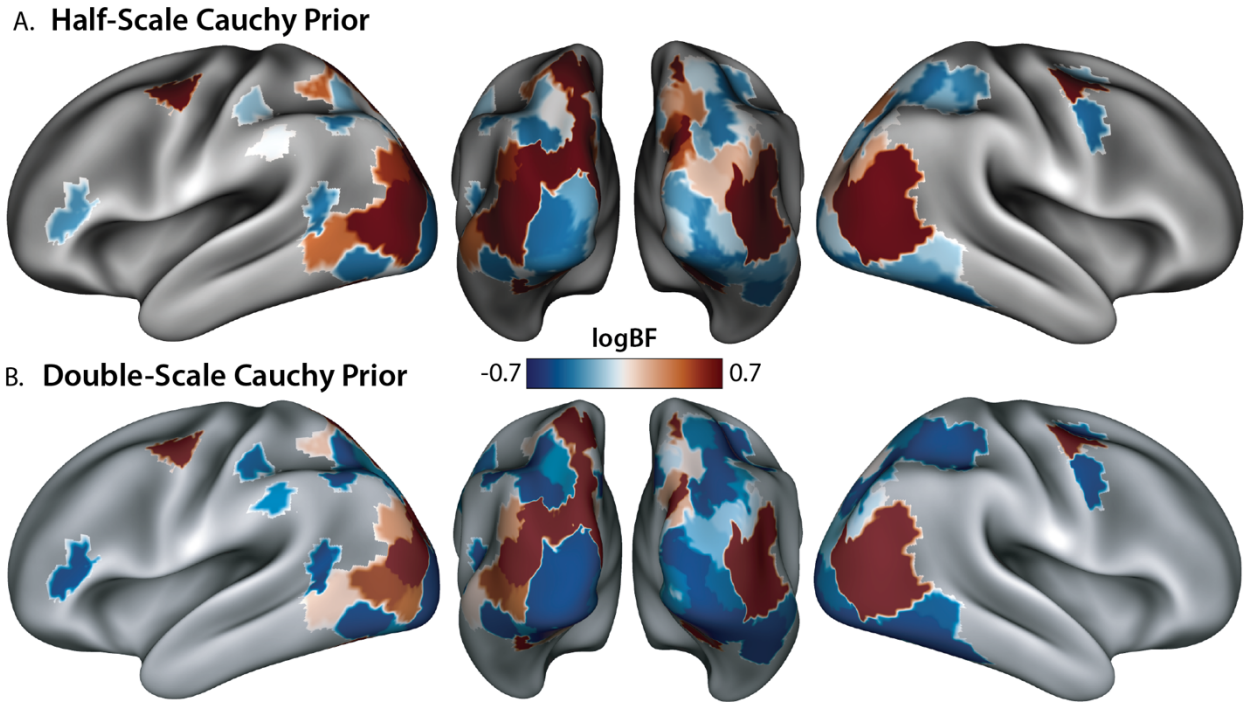


**Figure S4.** *Bayes factor prior control analysis.* **A)** Log bayes factors for target-distractor coherence alignment using a narrower prior (one-half the default Cauchy scale = 0.35). Minimum logBF is -0.46 at $t_{(28)} = 0$. **B)** Same log bayes factor using a wider prior (double the default Cauchy scale = 1.41). Minimum logBF = -0.99 at $t_{(28)} = 0$. Across different prior parameterizations, note the similarity to Figure 4C.

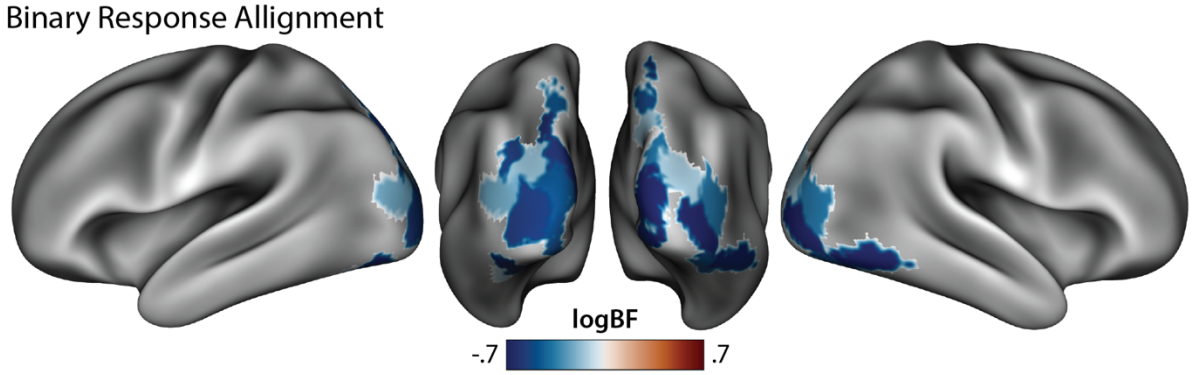


**Figure S5.** *Binary response encoding control analysis.* Target-distractor response encoding alignment using binary responses rather than coherence-modulated responses. Note the similarity to Figure 4D.

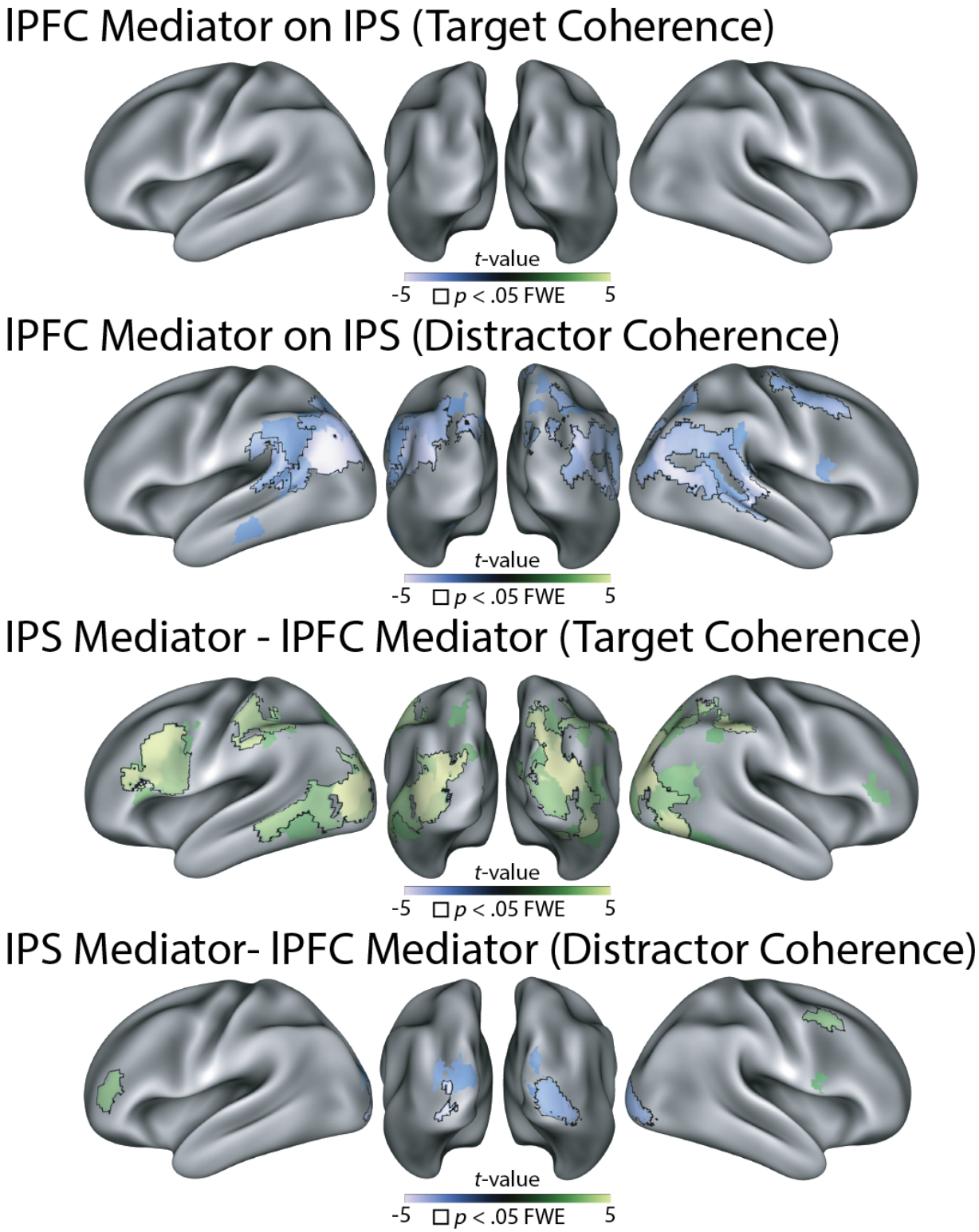


**Figure S6.** *lPFC mediation.* IPS→lPFC→Coherence mediation for target coherence (first row) and distractor coherence (second row; compare to Figure 7c). Contrast between IPS-mediation and lPFC-mediation for target coherence (third row) and distractor coherence (fourth row).

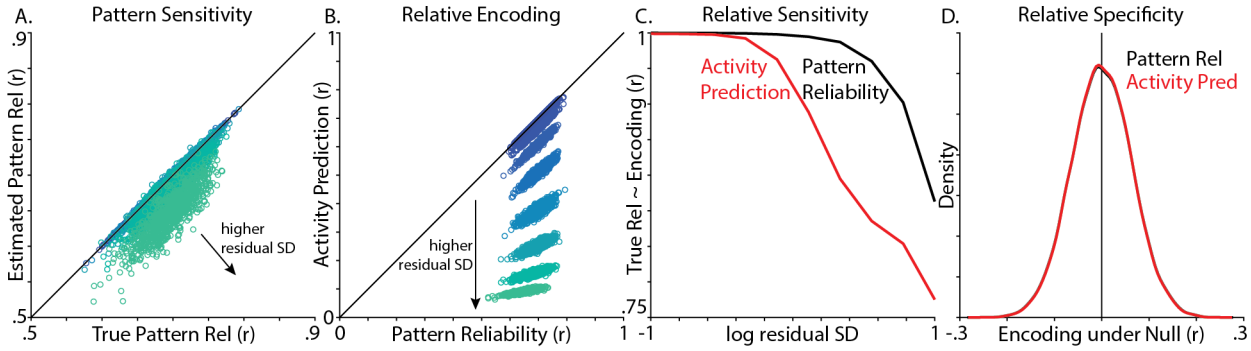


**Figure S7.** *Encoding Geometry Analysis (EGA) validation.* We validated how well we could recover the similarity between linear Gaussian models (training: $Y = XB + \Sigma$, test: $Y' = X'B' + \Sigma$). $Y$ is the $[1000 \times 250]$ activity timeseries, $X$ is the $[1000 \times 1]$ design matrix, $B$ is the $[1 \times 250]$ encoding profile, and $\Sigma$ 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 ($B' = B + \mathcal{N}(0,1)$), based on noisy activity timeseries ($Y = XB + \mathcal{N}(0, \sigma_Y)$; $Y' = X'B' + \mathcal{N}(0, \sigma_Y)$). The first method was *pattern reliability* (i.e., our EGA method), correlating the encoding profile estimated during training ($\hat{B} = X^{\dagger}Y$, $\dagger$ indicates pseudoinverse) with the encoding profile estimated during test ($\widehat{B'} = X'^{\dagger}Y'$). The second method was *activity prediction* (i.e., the traditional encoding approach), correlating the ground-truth test activity ($Y'$) with the predicted test activity ($\widehat{Y'} = X'\hat{B}$). To simulate the high measurement noise inherent to fMRI, we compared these methods under different levels of residual SD ($\sigma_Y$). **A)** Estimated pattern reliability tracked the true pattern reliability, across the full range of residual SD. **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 (i.e., 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 ($B'_{null} = \mathcal{N}(0,1)$).

# References

Adam KCS, Serences JT. 2021. History modulates early sensory processing of salient distractors. *J Neurosci*. doi:10.1523/JNEUROSCI.3099-20.2021

Aoi MC, Mante V, Pillow JW. 2020. Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. *Nat Neurosci*. doi:10.1038/s41593-020-0696-5

Ashby WR. 1961. An introduction to cybernetics. Chapman & Hall Ltd.

Badre D, Bhandari A, Keglovits H, Kikumoto A. 2021. The dimensionality of neural representations for control. *Curr Opin Behav Sci* **38**:20–28. doi:10.1016/j.cobeha.2020.07.002

Badre D, D'Esposito M. 2009. Is the rostro-caudal axis of the frontal lobe hierarchical? *Nat Rev Neurosci* **10**:659–669. doi:10.1038/nrn2667

Badre D, Nee DE. 2018. Frontal Cortex and the Hierarchical Control of Behavior. *Trends Cogn Sci* **22**:170–188. doi:10.1016/j.tics.2017.11.005

Belsley DA, Kuh E, Welsch RE. 1980. Wiley Series in Probability and Statistics. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* 293–300.

Bernardi S, Benna MK, Rigotti M, Munuera J, Fusi S, Daniel Salzman C. 2020. The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell* **0**. doi:10.1016/j.cell.2020.09.031

Bisley JW, Goldberg ME. 2010. Attention, intention, and priority in the parietal lobe. *Annu Rev Neurosci* **33**:1–21. doi:10.1146/annurev-neuro-060909-152823

Clairis N, Pessiglione M. 2020. Value, confidence, deliberation: a functional partition of the medial prefrontal cortex demonstrated across rating and choice tasks. *bioRxiv*. doi:10.1101/2020.09.17.301291

Cohen MR, Maunsell JHR. 2010. A neuronal population measure of attention predicts behavioral performance on individual trials. *J Neurosci* **30**:15241–15253. doi:10.1523/JNEUROSCI.2171-10.2010

Corbetta M, Shulman GL. 2002. Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci* **3**:201–215. doi:10.1038/nrn755

Culham JC, Brandt SA, Cavanagh P, Kanwisher NG, Dale AM, Tootell RB. 1998. Cortical fMRI activation produced by attentive tracking of moving targets. *J Neurophysiol* **80**:2657–2670. doi:10.1152/jn.1998.80.5.2657

Culham JC, Cavanagh P, Kanwisher NG. 2001. Attention response functions: characterizing brain areas using fMRI activation during parametric variations of attentional load. *Neuron* **32**:737–745. doi:10.1016/s0896-6273(01)00499-8

Danielmeier C, Eichele T, Forstmann BU, Tittgemeyer M, Ullsperger M. 2011. Posterior medial frontal cortex activity predicts post-error adaptations in task-related visual and motor areas. *J Neurosci* **31**:1780–1789. doi:10.1523/JNEUROSCI.4299-10.2011

Danielmeier C, Ullsperger M. 2011. Post-error adjustments. *Front Psychol* **2**:233. doi:10.3389/fpsyg.2011.00233

Desimone R, Duncan J. 1995. Neural mechanisms of selective visual attention. *Annu Rev Neurosci* **18**:193–222. doi:10.1146/annurev.ne.18.030195.001205

Diedrichsen J, Kriegeskorte N. 2017. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Comput Biol* **13**:e1005508. doi:10.1371/journal.pcbi.1005508

Diedrichsen J, Shadmehr R. 2005. Detecting and adjusting for artifacts in fMRI time series data. *Neuroimage* **27**:624–634. doi:10.1016/j.neuroimage.2005.04.039

Diedrichsen J, Yokoi A, Arbuckle SA. 2018. Pattern component modeling: A flexible approach for understanding the representational structure of brain activity patterns. *Neuroimage* **180**:119–133. doi:10.1016/j.neuroimage.2017.08.051

Ebitz BR, Smith EH, Horga G, Schevon CA, Yates MJ, McKhann GM, Botvinick MM, Sheth SA, Hayden BY. 2020. Human dorsal anterior cingulate neurons signal conflict by amplifying task-relevant information. *bioRxiv*. doi:10.1101/2020.03.14.991745

Egner T. 2008. Multiple conflict-driven control mechanisms in the human brain. *Trends Cogn Sci* **12**:374–380. doi:10.1016/j.tics.2008.07.001

Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, Kent JD, Goncalves M, DuPre E, Snyder M, Oya H, Ghosh SS, Wright J, Durnez J, Poldrack RA, Gorgolewski KJ. 2019. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat Methods* **16**:111–116. doi:10.1038/s41592-018-0235-4

Esterman M, Chiu Y-C, Tamber-Rosenau BJ, Yantis S. 2009. Decoding cognitive control in human parietal cortex. *Proc Natl Acad Sci U S A* **106**:17974–17979. doi:10.1073/pnas.0903593106

Etzel JA, Cole MW, Zacks JM, Kay KN, Braver TS. 2016. Reward Motivation Enhances Task Coding in Frontoparietal Cortex. *Cereb Cortex* **26**:1647–1659. doi:10.1093/cercor/bhu327

Fischer AG, Nigbur R, Klein TA, Danielmeier C, Ullsperger M. 2018. Cortical beta power reflects decision dynamics and uncovers multiple facets of post-error adaptation. *Nat Commun* **9**:5038. doi:10.1038/s41467-018-07456-8

Fleming SM, van der Putten EJ, Daw ND. 2018. Neural mediators of changes of mind about perceptual decisions. *Nat Neurosci* **21**:617–624. doi:10.1038/s41593-018-0104-6

Flesch T, Juechems K, Dumbalska T, Saxe A, Summerfield C. 2022. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron* **0**. doi:10.1016/j.neuron.2022.01.005

Freund MC, Bugg JM, Braver TS. 2021. A Representational Similarity Analysis of Cognitive Control during Color-Word Stroop. *J Neurosci* **41**:7388–7402. doi:10.1523/JNEUROSCI.2956-20.2021

Gale DJ, Vos de Wael. R, Benkarim O, Bernhardt B. 2021. Surfplot: Publication-ready brain surface figures. doi:10.5281/zenodo.5567926

Goldman-Rakic PS. 1988. Topography of cognition: parallel distributed networks in primate association cortex. *Annu Rev Neurosci* **11**:137–156. doi:10.1146/annurev.ne.11.030188.001033

Gorgolewski K, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML, Ghosh SS. 2011. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front Neuroinform* **5**:13. doi:10.3389/fninf.2011.00013

Gottlieb J, Cohanpour M, Li Y, Singletary N, Zabeh E. 2020. Curiosity, information demand and attentional priority. *Current Opinion in Behavioral Sciences* **35**:83–91. doi:10.1016/j.cobeha.2020.07.016

Greenberg AS, Esterman M, Wilson D, Serences JT, Yantis S. 2010. Control of spatial and feature-based attention in frontoparietal cortex. *J Neurosci* **30**:14330–14339. doi:10.1523/JNEUROSCI.4248-09.2010

Hall-McMaster S, Muhle-Karbe PS, Myers NE, Stokes MG. 2019. Reward Boosts Neural Coding of Task Rules to Optimize Cognitive Flexibility. *J Neurosci* **39**:8549–8561. doi:10.1523/JNEUROSCI.0631-19.2019

Holroyd CB, McClure SM. 2015. Hierarchical control over effortful behavior by rodent medial frontal cortex: A computational model. *Psychol Rev* **122**:54–83. doi:10.1037/a0038339

Holroyd CB, Yeung N. 2011. An Integrative Theory of Anterior Cingulate Cortex Function: Option Selection in Hierarchical Reinforcement Learning. *Neural Basis of Motivational and Cognitive Control*. doi:10.7551/mitpress/9780262016438.003.0018

Howe PD, Horowitz TS, Morocz IA, Wolfe J, Livingstone MS. 2009. Using fMRI to distinguish components of the multiple object tracking task. *J Vis* **9**:10.1-11. doi:10.1167/9.4.10

Hunt LT, Kolling N, Soltani A, Woolrich MW, Rushworth MFS, Behrens TEJ. 2012. Mechanisms underlying cortical activity during value-guided choice. *Nat Neurosci* **15**:470–6, S1-3. doi:10.1038/nn.3017

Jackson J, Rich AN, Williams MA, Woolgar A. 2017. Feature-selective Attention in Frontoparietal Cortex: Multivoxel Codes Adjust to Prioritize Task-relevant Information. *J Cogn Neurosci* **29**:310–321. doi:10.1162/jocn_a_01039

Jackson JB, Feredoes E, Rich AN, Lindner M, Woolgar A. 2021. Concurrent neuroimaging and neurostimulation reveals a causal role for dlPFC in coding of task-relevant information. *Commun Biol* **4**:588. doi:10.1038/s42003-021-02109-x

Jones MS, Zhu Z, Bajracharya A, Luor A, Peelle JE. 2021. A multi-dataset evaluation of frame censoring for task-based fMRI. *bioRxiv*. doi:10.1101/2021.10.12.464075

Jovicich J, Peters RJ, Koch C, Braun J, Chang L, Ernst T. 2001. Brain areas specific for attentional load in a motion-tracking task. *J Cogn Neurosci* **13**:1048–1058.

Kastner S, Ungerleider LG. 2000. Mechanisms of visual attention in the human cortex. *Annu Rev Neurosci* **23**:315–341. doi:10.1146/annurev.neuro.23.1.315

Kaufman MT, Churchland MM, Ryu SI, Shenoy KV. 2014. Cortical activity in the null space: permitting preparation without movement. *Nat Neurosci* **17**:440–448. doi:10.1038/nn.3643

Kay KN, Yeatman JD. 2017. Bottom-up and top-down computations in word- and face-selective cortex. *Elife* **6**. doi:10.7554/eLife.22341

Kayser AS, Buchsbaum BR, Erickson DT, D'Esposito M. 2010a. The functional anatomy of a perceptual decision in the human brain. *J Neurophysiol* **103**:1179–1194. doi:10.1152/jn.00364.2009

Kayser AS, Erickson DT, Buchsbaum BR, D'Esposito M. 2010b. Neural representations of relevant and irrelevant features in perceptual decision making. *J Neurosci* **30**:15778–15789. doi:10.1523/JNEUROSCI.3163-10.2010

Kimmel DL, Elsayed GF, Cunningham JP, Newsome WT. 2020. Value and choice as separable and stable representations in orbitofrontal cortex. *Nat Commun* **11**:3466. doi:10.1038/s41467-020-17058-y

Koechlin E, Summerfield C. 2007. An information theoretical approach to prefrontal executive function. *Trends Cogn Sci* **11**:229–235. doi:10.1016/j.tics.2007.04.005

Kong R, Yang Q, Gordon E, Xue A, Yan X, Orban C, Zuo X-N, Spreng N, Ge T, Holmes A, Eickhoff S, Yeo BTT. 2021. Individual-Specific Areal-Level Parcellations Improve Functional Connectivity Prediction of Behavior. *Cereb Cortex* **31**:4477–4500. doi:10.1093/cercor/bhab101

Kragel PA, Kano M, Van Oudenhove L, Ly HG, Dupont P, Rubio A, Delon-Martin C, Bonaz BL, Manuck SB, Gianaros PJ, Ceko M, Reynolds Losin EA, Woo C-W, Nichols TE, Wager TD. 2018. Generalizable representations of pain, cognitive control, and negative emotion in medial frontal cortex. *Nat Neurosci* **21**:283–289. doi:10.1038/s41593-017-0051-7

Kriegeskorte N, Diedrichsen J. 2019. Peeling the Onion of Brain Representations. *Annu Rev Neurosci* **42**:407–432. doi:10.1146/annurev-neuro-080317-061906

Kriegeskorte N, Goebel R, Bandettini P. 2006. Information-based functional brain mapping. *Proc Natl Acad Sci U S A* **103**:3863–3868. doi:10.1073/pnas.0600244103

Lauritzen TZ, D'Esposito M, Heeger DJ, Silver MA. 2009. Top–down flow of visual spatial attention signals from parietal to occipital cortex. *J Vis* **9**:18–18. doi:10.1167/9.13.18

Leng X, Yee D, Ritz H, Shenhav A. 2021. Dissociable influences of reward and punishment on adaptive cognitive control. *PLoS Comput Biol* **17**:e1009737. doi:10.1371/journal.pcbi.1009737

Libby A, Buschman TJ. 2021. Rotational dynamics reduce interference between sensory and memory representations. *Nat Neurosci* 1–12. doi:10.1038/s41593-021-00821-9

MacKinnon DP, Fairchild AJ, Fritz MS. 2007. Mediation analysis. *Annu Rev Psychol* **58**:593–614. doi:10.1146/annurev.psych.58.110405.085542

Mante V, Sussillo D, Shenoy KV, Newsome WT. 2013. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**:78–84. doi:10.1038/nature12742

Miller EK, Cohen JD. 2001. An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* **24**:167–202. doi:10.1146/annurev.neuro.24.1.167

Minxha J, Adolphs R, Fusi S, Mamelak AN, Rutishauser U. 2020. Flexible recruitment of memory-based choice representations by the human medial frontal cortex. *Science* **368**. doi:10.1126/science.aba3313

Molenberghs P, Mesulam MM, Peeters R, Vandenberghe RRC. 2007. Remapping attentional priorities: differential contribution of superior parietal lobule and intraparietal sulcus. *Cereb Cortex* **17**:2703–2712. doi:10.1093/cercor/bhl179

Musslick S, Saxe A, Hoskin AN, Reichman D, Cohen JD. 2020. On the Rational Boundedness of Cognitive Control: Shared Versus Separated Representations. doi:10.31234/osf.io/jkhdf

Nichols TE, Holmes AP. 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp* **15**:1–25. doi:10.1002/hbm.1058

Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N. 2014. A toolbox for representational similarity analysis. *PLoS Comput Biol* **10**:e1003553. doi:10.1371/journal.pcbi.1003553

Pagan M, Tang VD, Aoi MC, Pillow JW, Mante V, Sussillo D, Brody CD. 2022. A new theoretical framework jointly explains behavioral and neural variability across subjects performing flexible decision-making. *bioRxiv*. doi:10.1101/2022.11.28.518207

Panichello MF, Buschman TJ. 2021. Shared mechanisms underlie the control of working memory and attention. *Nature* 1–5. doi:10.1038/s41586-021-03390-w

Parro C, Dixon ML, Christoff K. 2018. The neural basis of motivational influences on cognitive control. *Hum Brain Mapp* **39**:5097–5111. doi:10.1002/hbm.24348

Peck CJ, Jangraw DC, Suzuki M, Efem R, Gottlieb J. 2009. Reward modulates attention independently of action value in posterior parietal cortex. *J Neurosci* **29**:11182–11191. doi:10.1523/JNEUROSCI.1929-09.2009

Pylyshyn ZW, Storm RW. 1988. Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spat Vis* **3**:179–197. doi:10.1163/156856888x00122

Ritz H, Leng X, Shenhav A. 2022a. Cognitive Control as a Multivariate Optimization Problem. *J Cogn Neurosci* **34**:569–591. doi:10.1162/jocn_a_01822

Ritz H, Shenhav A. 2021. Humans reconfigure target and distractor processing to address distinct task demands. *bioRxiv* 2021.09.08.459546. doi:10.1101/2021.09.08.459546

Ritz H, Wild CJ, Johnsrude IS. 2022b. Parametric Cognitive Load Reveals Hidden Costs in the Neural Processing of Perfectly Intelligible Degraded Speech. *J Neurosci* **42**:4619–4628. doi:10.1523/JNEUROSCI.1777-21.2022

Rouder JN, Morey RD, Speckman PL, Province JM. 2012. Default Bayes factors for ANOVA designs. *J Math Psychol* **56**:356–374. doi:10.1016/j.jmp.2012.08.001

Rust NC, Cohen MR. 2022. Priority coding in the visual system. *Nat Rev Neurosci* 1–13. doi:10.1038/s41583-022-00582-9

Saalmann YB, Pigarev IN, Vidyasagar TR. 2007. Neural mechanisms of visual attention: how top-down feedback highlights relevant locations. *Science* **316**:1612–1615. doi:10.1126/science.1139140

Salinas E. 2004. Fast remapping of sensory stimuli onto motor actions on the basis of contextual modulation. *J Neurosci* **24**:1113–1118. doi:10.1523/JNEUROSCI.4569-03.2004

Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo X-N, Holmes AJ, Eickhoff SB, Yeo BTT. 2018. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cereb Cortex* **28**:3095–3114. doi:10.1093/cercor/bhx179

Semedo JD, Zandvakili A, Machens CK, Yu BM, Kohn A. 2019. Cortical Areas Interact through a Communication Subspace. *Neuron* **102**:249-259.e4. doi:10.1016/j.neuron.2019.01.026

Serences JT, Schwarzbach J, Courtney SM, Golay X, Yantis S. 2004. Control of object-based attention in human cortex. *Cereb Cortex* **14**:1346–1357. doi:10.1093/cercor/bhh095

Serences JT, Yantis S. 2007. Spatially selective representations of voluntary and stimulus-driven attentional priority in human occipital, parietal, and frontal cortex. *Cereb Cortex* **17**:284–293. doi:10.1093/cercor/bhj146

Shenhav A, Botvinick MM, Cohen JD. 2013. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron* **79**:217–240. doi:10.1016/j.neuron.2013.07.007

Shenhav A, Straccia MA, Botvinick MM, Cohen JD. 2016a. Dorsal anterior cingulate and ventromedial prefrontal cortex have inverse roles in both foraging and economic choice. *Cogn Affect Behav Neurosci*. doi:10.3758/s13415-016-0458-8

Shenhav A, Straccia MA, Botvinick MM, Cohen JD. 2016b. Dorsal anterior cingulate and ventromedial prefrontal cortex have inverse roles in both foraging and economic choice. *bioRxiv*. doi:10.1101/046276

Shenhav A, Straccia MA, Musslick S, Cohen JD, Botvinick MM. 2018. Dissociable neural mechanisms track evidence accumulation for selection of attention versus action. *Nat Commun* **9**:2485. doi:10.1038/s41467-018-04841-1

Smith EH, Horga G, Yates MJ, Mikell CB, Banks GP, Pathak YJ, Schevon CA, McKhann GM, Hayden BY, Botvinick MM, Sheth SA. 2019. Widespread temporal coding of cognitive control in the human prefrontal cortex. *Nat Neurosci* **66**:83. doi:10.1038/s41593-019-0494-0

Smith SM, Nichols TE. 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage* **44**:83–98. doi:10.1016/j.neuroimage.2008.03.061

Soutschek A, Stelzel C, Paschke L, Walter H, Schubert T. 2015. Dissociable effects of motivation and expectancy on conflict processing: an fMRI study. *J Cogn Neurosci* **27**:409–423. doi:10.1162/jocn_a_00712

Spearman C. 1987. The Proof and Measurement of Association between Two Things. *Am J Psychol* **100**:441–471. doi:10.2307/1422689

Srinath R, Ruff DA, Cohen MR. 2021. Attention improves information flow between neuronal populations without changing the communication subspace. *bioRxiv*. doi:10.1101/2021.03.31.437940

Stringer C, Pachitariu M, Steinmetz N, Reddy CB, Carandini M, Harris KD. 2019. Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364**:255. doi:10.1126/science.aav7893

Suzuki M, Gottlieb J. 2013. Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nat Neurosci* **16**:98–104. doi:10.1038/nn.3282

Takagi Y, Hunt LT, Woolrich MW, Behrens TE, Klein-Flügge MC. 2021. Adapting non-invasive human recordings along multiple task-axes shows unfolding of spontaneous and over-trained choice. *Elife* **10**. doi:10.7554/eLife.60988

Taren AA, Venkatraman V, Huettel SA. 2011. A parallel functional topography between medial and lateral prefrontal cortex: evidence and implications for cognitive control. *J Neurosci* **31**:5026–5031. doi:10.1523/JNEUROSCI.5762-10.2011

Thornton MA, Mitchell JP. 2017. Consistent Neural Activity Patterns Represent Personally Familiar People. *J Cogn Neurosci* **29**:1583–1594. doi:10.1162/jocn_a_01151

Vassena E, Deraeve J, Alexander WH. 2017. 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. doi:10.1162/jocn_a_01160

Venkatraman V, Rosati AG, Taren AA, Huettel SA. 2009. Resolving response, decision, and strategic control: evidence for a functional topography in dorsomedial prefrontal cortex. *J Neurosci* **29**:13158–13164. doi:10.1523/JNEUROSCI.2708-09.2009

Vermeylen L, Wisniewski D, González-García C, Hoofs V, Notebaert W, Braem S. 2020. Shared Neural Representations of Cognitive Conflict and Negative Affect in the Medial Frontal Cortex. *J Neurosci* **40**:8715–8725. doi:10.1523/JNEUROSCI.1744-20.2020

Vermeylen L, Wisniewski D, González-García C, Hoofs V, Notebaert W, Braem S. 2019. Shared Neural Representations of Cognitive Conflict and Negative Affect in the Dorsal Anterior Cingulate Cortex. *bioRxiv*. doi:10.1101/824839

Vos de Wael R, Benkarim O, Paquola C, Lariviere S, Royer J, Tavakol S, Xu T, Hong S-J, Langs G, Valk S, Misic B, Milham M, Margulies D, Smallwood J, Bernhardt BC. 2020. BrainSpace: a toolbox for the analysis of macroscale gradients in neuroimaging and connectomics datasets. *Commun Biol* **3**:103. doi:10.1038/s42003-020-0794-7

Vul E, Alvarez G, Tenenbaum J, Black M. 2009. Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model In: Bengio Y, Schuurmans D, Lafferty J, Williams C, Culotta A, editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.

Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N, Diedrichsen J. 2016. Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage* **137**:188–200. doi:10.1016/j.neuroimage.2015.12.012

Weichart ER, Turner BM, Sederberg PB. 2020. A model of dynamic, within-trial conflict resolution for decision making. *Psychol Rev*. doi:10.1037/rev0000191

Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE. 2014. Permutation inference for the general linear model. *Neuroimage* **92**:381–397. doi:10.1016/j.neuroimage.2014.01.060

Wisniewski D, Reverberi C, Momennejad I, Kahnt T, Haynes J-D. 2015. The Role of the Parietal Cortex in the Representation of Task–Reward Associations. *J Neurosci* **35**:12355–12365. doi:10.1523/JNEUROSCI.4882-14.2015

Woolgar A, Afshar S, Williams MA, Rich AN. 2015a. Flexible Coding of Task Rules in Frontoparietal Cortex: An Adaptive System for Flexible Cognitive Control. *J Cogn Neurosci* **27**:1895–1911. doi:10.1162/jocn_a_00827

Woolgar A, Hampshire A, Thompson R, Duncan J. 2011. Adaptive coding of task-relevant information in human frontoparietal cortex. *J Neurosci* **31**:14592–14599. doi:10.1523/JNEUROSCI.2616-11.2011

Woolgar A, Williams MA, Rich AN. 2015b. Attention enhances multi-voxel representation of novel objects in frontal, parietal and visual cortices. *Neuroimage* **109**:429–437. doi:10.1016/j.neuroimage.2014.12.083

Yantis S, Schwarzbach J, Serences JT, Carlson RL, Steinmetz MA, Pekar JJ, Courtney SM. 2002. Transient neural activity in human parietal cortex during spatial attention shifts. *Nat Neurosci* **5**:995–1002. doi:10.1038/nn921

Yantis S, Serences JT. 2003. Cortical mechanisms of space-based and object-based attentional control. *Curr Opin Neurobiol* **13**:187–193. doi:10.1016/s0959-4388(03)00033-3

Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD. 2011. Large-scale automated synthesis of human functional neuroimaging data. *Nat Methods* **8**:665–670. doi:10.1038/nmeth.1635

Yeo BTT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, Roffman JL, Smoller JW, Zöllei L, Polimeni JR, Fischl B, Liu H, Buckner RL. 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J Neurophysiol* **106**:1125–1165. doi:10.1152/jn.00338.2011

Zarr N, Brown JW. 2016. Hierarchical error representation in medial prefrontal cortex. *Neuroimage* **124**:238–247. doi:10.1016/j.neuroimage.2015.08.063