Hierarchical Reconfiguration of Neurocognitive Task Set Representations Mediates Cognitive Flexibility

Stephanie C. Leach[1 *], Xitong Chen[1], and Kai Hwang[1,2]

[1] Psychological and Brain Sciences, The University of Iowa, Iowa City, IA.

[2] Cognitive Control Collaborative (CCC).

* Corresponding Author

E-mail: stephanie-leach@uiowa.edu

41 pages

6 figures

250 words abstract

693 words introduction

1530 words discussion

Abstract

Cognitive control organizes contexts, stimuli, and actions into hierarchically structured representations that support flexible, goal-directed behavior. This organization allows appropriate actions to be selected in response to changing demands, contexts, and sensory inputs. Critically, different hierarchical levels place unequal demands on flexibility: sensory–motor plans must be updated rapidly to enable behavioral flexibility, whereas higher-level contextual goals are typically updated less frequently. A behavioral consequence of managing these demands is the task switch cost, the performance penalty observed when task representations are reconfigured across trials. Here, we investigated the behavioral and neural sources of switch costs using a hierarchical control task that dissociates context reconfiguration from subordinate rule switching. Functional MRI and behavioral data were collected from healthy human participants (both male and female). Behaviorally, subordinate rule switches were faster than context switches but were more strongly influenced by task-irrelevant perceptual changes, whereas context switches were relatively insensitive to such interference. To characterize the neural basis of these effects, we introduced a novel measure "neural distance", which quantifies trial-to-trial reconfiguration of multivoxel activity patterns. Across the brain, larger neural distances predicted larger RT switch costs, linking representational reorganization to behavioral performance. Importantly, representational reconfiguration differed across hierarchical levels and anatomical systems. Subordinate rule updating was modulated by perceptual input and expressed in distributed perceptual and motor networks, whereas context reconfiguration engaged the lateral middle frontal cortex and was comparatively insulated from interference. Our results reveal the structure of neural representations supporting flexible updating and interference-shielded contextual representations subserving control behavior.

Significance

Daily activities often require a mixture of both high and low flexibility demands that need to be performed concurrently. Driving requires maintaining a stable destination goal while rapidly updating motor plans (brake, accelerate, turn, etc.) in response to changing sensory information. Although people perform such tasks with ease, it remains unclear how neural and cognitive representations are structured to respond to these different demands. The present study suggests that sensory-motor plans prioritize high flexibility by allowing greater influence from sensory inputs, which can create interference across brain networks when that input is task-irrelevant. Contextual information is shielded from this interference by representing contexts as distinctly as possible in the lateral prefrontal cortex, resulting in slower but more stable context switching.

1. Introduction

Everyday tasks often consist of a mixture of high and low flexibility in goal directed behavior. For example, driving requires maintaining an overarching destination goal while continuously updating route plans and switching between accelerating and braking in response to dynamic sensory inputs. Because this overarching destination goal does not require frequent updating, flexibility demands are low. Conversely, more immediate route and sensory-motor plans require fast and accurate updating, therefore flexibility demands are high. Understanding how these demands are jointly supported requires examining the representational structures encoded in the brain that implement cognitive control (Monsell, 2003; Meiran et al., 2000; Rogers & Monsell, 1995; Badre et al., 2021).

Cognitive control depends on representations incorporating goals, contexts, stimuli, and actions. How these representations are structured and encoded shapes cognitive flexibility.

Contextual representations, often attributed to the rostral portions of mid-lateral prefrontal cortex (Badre & D'Esposito, 2007; Badre et al., 2009; Nee & Brown, 2013), are assumed to bias downstream processes to prioritize task-relevant information for contextually guided hierarchical cognitive control (Miller & Cohen, 2001; Nee & D'Esposito, 2017). Such biasing can prioritize shielding representations from interference, which in some cases can incur additional cost of flexibility (Geddert & Egner, 2024; Dreisbach & Fröber, 2018; Mayr & Grätz, 2024). Yet the same system can also support high flexibility by integrating and updating sensory-motor representations under changing contexts (Kikumoto & Mayr, 2020; Kikumoto et al., 2024) and redirecting attention toward alternative task rules or stimuli dimensions (Cellier et al., 2022). High flexibility may therefore be facilitated when control representations permit greater bottom-up influence from sensory inputs (Schumacher & Hazeltine, 2016).

The structure of control representations can be probed by examining trial-to-trial reconfiguration (Dykstra et al., 2022). Behavioral paradigms have historically focused on trial-wise transitions, based on the assumption that internal structures should constrain observable actions. One ubiquitous measure is response time "switch costs"—the additional time it takes to make a response when a task changes from one trial to the next (Monsell, 2003; Rogers & Monsell, 1995; Wylie & Allport, 2000). By systematically analyzing factors influencing switch costs, one can identify the representational properties supporting flexibility. For example, if high flexibility is implemented by increasing sensory-driven influence, then task-irrelevant sensory information should show greater influence on switch costs.

Although switch costs provide a behavioral readout of reconfiguration processes, their neural basis remains unclear. Recent work suggests information is encoded in distributed activity patterns across neuron populations, voxels, or electrodes (Ebitz & Hayden, 2021; Mur et al.,

2009). Under this population view, cognitive flexibility entails transforming one activity pattern into another. The farther the system must travel in multivariate space, the longer the transformation should take, producing a larger switch cost (Qiao et al., 2017). Distance traveled can be estimated by calculating the multivariate distance between the activity pattern on a switch trial versus the preceding trial. A small distance would require minimal energy, whereases a larger distance, reflecting costly reorganization, would predict a larger behavioral switch cost. By that logic, more concrete task-rules and sensory-motor representations requiring high flexibility should minimize representational distance. Thereby, producing smaller switch costs at the price of increased vulnerability to sensory driven feed-forward interference. In comparison, less flexible contextual representations should be encoded as distinctively as possible (i.e., larger representational distance) to minimize interference.

Given these hypotheses, our study pursued two main goals. First, we used response time switch costs to determine how distinct cognitive task representations are reconfigured during hierarchical cognitive control. Second, we quantified trial-to-trial shifts in multivoxel activity patterns, which we refer to as “neural distance”, to examine how neural representational reconfiguration is influenced by both the hierarchical structure of the task and additional factors, such as perceptual processes, and the interactions among these factors. Results confirmed that larger neural distances predicted larger RT switch costs. Moreover, representational reconfiguration differed across hierarchical levels in its sensitivity to perceptual input. Increased feedforward perceptual influence modulated subordinate rule selection but not contextual rule selection. This result suggests that contextual representations may organize lower-level rule–stimulus–response conjunctions in a way that limits interference from irrelevant perceptual

information. Finally, greater neural reconfiguration in subcortical structures was associated with greater post-error slowing.

2. **Methods**

2.1. Participants. Participants were recruited for two experiments, experiment 1 involved MRI scanning, and experiment 2 was a behavioral only experiment conducted to replicate in-scanner behavioral effects. For experiment 1, we recruited 73 healthy participants (53 female, 20 male, M_{age} = 22.31, SD_{age} = 4.45, $Range_{age}$ = 18-35 years). Out of these 73 participants, 10 were excluded for poor behavioral performance (accuracy < 75% for at least one condition), and 9 were excluded for excessive movement artifacts (mean framewise displacement > 0.4 mm across runs). After exclusion, we included data from 54 participants (39 female, 15 male, M_{age} = 22.38, SD_{age} = 4.87, $Range_{age}$ = 18-35 years). These data were also published in our previous study (Chen et al., 2024). For experiment 2, we recruited 65 healthy participants (55 female, 9 male, 1 non-binary, M_{age} = 18.63, SD_{age} = 0.71, $Range_{age}$ = 18-21 years). Out of these 65 participants, 8 were excluded for poor behavioral performance (accuracy < 75% for at least one condition). After exclusion, we included data from 57 participants (50 female, 7 male, M_{age} = 18.63, SD_{age} = 0.72, $Range_{age}$ = 18-21 years). Both experiments were approved by the University of Iowa Institutional Review Board. All participants were recruited from the University of Iowa and surrounding area to participate. All participants had normal or corrected to normal visual acuity and color vision, were right-handed, and with no history of epilepsy, psychiatric, or neurological conditions. All participants gave written informed consent. The experiments were conducted in compliance with the ethical principles expressed in the Declaration of Helsinki.

2.2. MRI data Acquisition. Imaging data were collected at the Magnetic Resonance Research Facility at the University of Iowa using a 3T GE SIGNA Premier scanner with a 48-channel head coil. Structural images were acquired using a multi-echo MPRAGE sequence (TR = 2348.52ms; TE = 2.968ms; flip angle = 8°; field of view = 256*256; 200 sagittal slices; voxel size = $1mm^3$). Functional images were acquired using an echo-planar sequence sensitive to blood oxygenated level-dependent (BOLD) contrast (Multiband acceleration factor = 2; no in-plane acceleration; TR = 1800ms; TE = 30ms; flip angle = 75°; voxel size = 1.718 mm*1.718mm *2.5 mm).

2.3. MRI Data Preprocessing. All fMRI data were preprocessed using fMRIPrep version 20.1.1 (Esteban et al., 2019) to reduce noise and transform data from subject native space to the ICBM 152 Nonlinear Asymmetrical template version 2009c for group analysis (Fonov et al., 2011). Preprocessing steps include bias field correction, skull-stripping, co-registration between functional and structural images, tissue segmentation, motion correction, and spatial normalization to the standard space. We did not perform any spatial smoothing and did not apply conventional nuisance regression. Instead, noise suppression was implemented during within the GLMsingle procedure (Prince et al., 2022). GLMsingle combines voxel-wise GLM estimation with automated denoising procedures from GLMdenoise (Kay et al., 2013), which uses cross-validation to identify and apply the optimal combination of nuisance regressors. Nuisance regressors, motion estimates, and principal components were learned to best improve cross-validation performance from whole-brain regions of interest. The framework also estimates single-trial beta weights while correcting for temporal autocorrelation and optimizing the

hemodynamic response function on a voxel-wise basis (see section 2.7). This approach provides improved sensitivity and reliability for our trial-by-trial transition analyses.

2.4. Experimental Design. In experiment 1, participants underwent an fMRI scanning session while performing a hierarchical cognitive control task. Before entering the scanner, all participants completed a tutorial and a practice session. Once participants achieved at least 80% on the practice, the experiment's scanning session began. We collected eight runs of functional scans, with each run consisting of 51 task trials. This resulted in a total of 408 task trials across all eight runs. Each functional run lasted approximately 6.5 minutes. In experiment 2, participants performed the same hierarchical cognitive control task behaviorally with the same training procedure, trial timing, block structure, and number of trials. The purpose was to replicate the behavioral effects we observed in scanner.

2.5. Behavioral Tasks. To study the representation structure for flexible hierarchical cognitive control, we developed a hierarchical cognitive control task (Chen et al., 2024). The task implements a three-level hierarchy in which texture defines context, context determines the relevant feature dimension, and the feature dimension maps onto a response rule.

2.5.1. *Task structure and trial sequence*. In this task, each trial began with a cue object presented for 500 milliseconds, followed by an image of either a face or a scene presented for 2000 milliseconds (**Figure 1A**). The inter-trial interval (ITI) was jittered between 1.5 to 10.5 seconds with an average duration of 4.25 seconds. Depending on the cue object presented before the image, participants made a yes/no judgment on one of two response rules: (1) was the image a face? (2) was the image a scene?

2.5.2. *Hierarchical rule structure*. To determine which response rule to use on a given trial, participants had to integrate three sources of information associated with three perceptual features of the cue object: texture, shape, and color (**Figure 1A**). The cue's texture determined whether the response rule was signaled by shape or color. If the texture of the cue object was filled in, the shape of the object (square or circle) would indicate which rule they needed to use (square being the face rule, "is the image a face?", and circle being the scene rule, "is the image a scene?"). Alternatively, if the texture of the cue object was hollow (i.e., only an outline), the color of the object (red or blue) would indicate which rule they needed to use (blue being the face rule and red being the scene rule). Half of the participants completed a counterbalanced version of this task with alternative yes/no response key mappings, and in which the top level of the hierarchy was flipped (i.e., hollow texture cues shape and filled texture cues color).

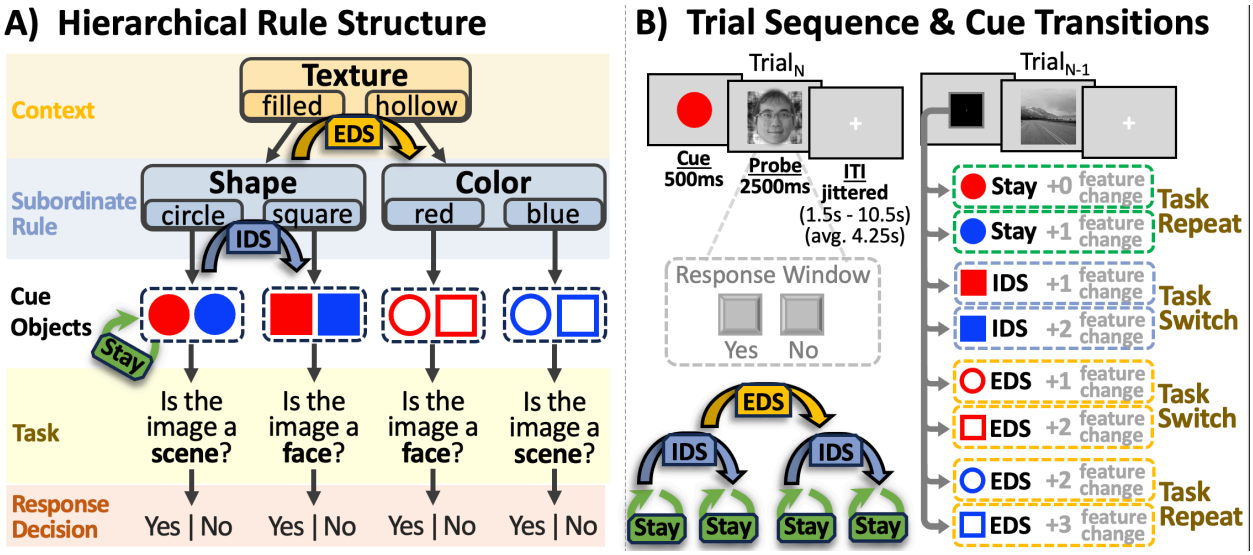

Figure 1. Hierarchical control task design. A) Depiction of the task rule hierarchy embedded within the 8 potential cue objects. B) Example trial sequence: A cue object presented for 500ms followed by an image of either a face or a scene for 2000ms and a jittered ITI (range: 1.5-10.5 seconds; average: 4.25 seconds. Example cue transitions across trials depicts how perceptual change is operationalized. Example face image is an author on this manuscript.

2.5.3. *Hierarchical rule switching*. This task design yields three types of trial-to-trial transitions that differ in the level of hierarchical reconfiguration required: Stay, intra-dimension

switch (IDS), and extra-dimension switch (EDS). On Stay trials, participants repeated the same rule as the previous trial, preserving both the texture-defined context and feature-to-rule mapping (e.g., which color or shape indicates the face task). On IDS trials, participants switched between task-relevant features (shape vs. color) within the same texture context. On EDS trials, they updated the highest-level context by switching texture cues (filled vs. hollow), thereby changing which feature signaled the response rule.

2.5.4. *Trial-to-trial transitions, perceptual, and motor factors*. In addition to hierarchical rule switching, the task design allowed us to quantify trial-to-trial perceptual and motor influences on performance. Specifically, (1) whether texture, shape, or color repeated from the previous trial, (2) how many cue object features changed (range: 0-3), (3) whether the targeted, relevant feature repeated or switched from the previous trial (e.g., if the relevant feature on trial N-1 was color red, whether it repeated or switched on the current trial), (4) whether the motor response repeated, and (5) how these factors interacted with rule switching. As described in detail below, we used linear regression to quantify the influence of each factor on response times and reconfiguration on voxel activity patterns.

2.5.5 Behavioral (Response Time) Switch Costs. To quantify behavioral costs associated with hierarchical task reconfiguration, we analyzed trial-wise response times using linear mixed-effects regression. This approach allowed us to estimate the contributions of hierarchical rule switching, perceptual feature transitions, and motor processes to performance costs while accounting for repeated measurements within subjects.

2.6. Multivariate Activity Pattern Distance. The goal of this analysis was to characterize how neural activity patterns that instantiate cognitive task representations are reconfigured from trial to trial. To this end, we quantified trial-to-trial changes in multivoxel BOLD activity patterns and examined how these representational shifts from trial-to-trial (hereafter described as neural distance; Fig. 2). Using the same analytical logic as our behavioral switch cost analysis, we fitted regression models to neural distance to determine how reconfiguration was influenced by hierarchical task structure, perceptual and motor factors, and error-related processes. This analysis included a multistep pipeline consisting of trial-wise activity estimation, computation of multivariate pattern distances, and regression-based modeling that paralleled to the behavioral switch-cost analysis.

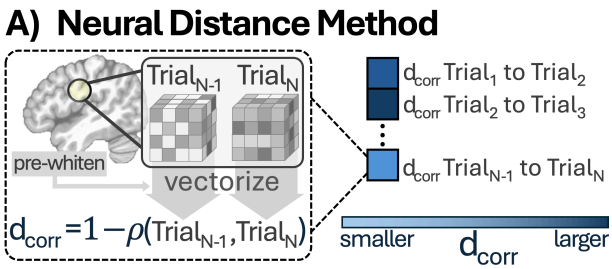

Figure 2. Neural Distance Method. A) depiction of how neural distance is calculated on adjacent trial multivoxel activity patterns. A greater change in multivoxel activity patterns is associated with a greater correlation, or neural distance.

2.6.1. *Trial-wise activity estimate*. Preprocessed fMRI time-series (outputs of fMRIPrep) were first submitted to GLMsingle (Prince et al., 2022) to estimates trial-wise beta estimates. GLMsingle employs a cross-validation framework in which voxel-wise noise regressors are derived from the data itself and evaluated on held-out runs; these regressors are then incorporated into a general linear model (GLM) to improve the reliability of single-trial β-estimates. Another key feature of GLMsingle is that every voxel will have its own best hemodynamic response function, so instead of assuming the same HRF across the whole-brain this procedure can account for difference in neurovascular coupling between regions. A further

benefit of GLMsingle estimates trial-wise beta-estimates via ridge regression. After constructing the design matrix with one regressor per trial, GLMsingle fits a penalized GLM in which a regularization parameter λ is chosen by cross-validation. This approach minimizes the inflated correlations in beta estimates between adjacent trials. At the end, this procedure yielded one amplitude estimate per trial per voxel. To ensure that variability in RT did not drive the neural-distance measures, we regressed each trial's RT from its corresponding amplitude estimate on a voxel-by-voxel basis before further analyses.

2.6.2. *Neural distance*. Using these RT-corrected trial-wise maps, we then calculated multivariate distances between trial-wise activity patterns. This was done with an 8 mm radius volumetric searchlight analysis, restricting to gray matter voxels. Within each searchlight sphere, signals were first "whitened" by applying the noise covariance matrix (equation 1) estimated across all trials and voxels in that sphere to remove shared noise correlations (equation 2) before measuring pattern shifts (Walther et al., 2016).

$$\sum = \frac{1}{T} R^T R \qquad \text{(eq 1)}$$

Where R is the model residuals matrix of size T x P. Note, T is the number of time points and P is the number of voxels. The covariance matrix, Σ, is size P x P.

$$b_k^* = b_k \sum^{-\frac{1}{2}} \qquad \text{(eq 2)}$$

Where b_k is a searchlight sphere multivoxel activity pattern of size P corresponding to one condition, k, of the possible K conditions.

This procedure removes voxel-wise correlations attributable to shared noise. We then computed the correlation distance (equations 3 and 4) between consecutive trials:

$$d_{cosine}(b_k, b_j) = 1 - \frac{\langle b_k, b_j \rangle}{\|b_k\| \|b_j\|} = 1 - \cos(\angle b_k b_j) \qquad \text{(eq 3)}$$

Where b_j is a searchlight sphere multivoxel activity pattern of size P corresponding to one condition, j, of the possible K conditions.

$$b_k = b_k - \bar{b}_k 1$$

$$b_j = b_j - \bar{b}_j 1$$

$$d_{correlation}(b_k, b_j) = 1 - \frac{\langle \tilde{b}_k, \tilde{b}_j \rangle}{\|\tilde{b}_k\| \|\tilde{b}_j\|} = 1 - \cos(\angle \tilde{b}_k, \tilde{b}_j) \qquad \text{(eq 4)}$$

Where 1 is a row vector of ones of size 1xP and $\bar{b}$ is the voxel mean.

Correlation distance was chosen over Euclidean or Mahalanobis metrics because we were primarily interested in changes in the neural representation geometry rather than overall amplitude shifts. This procedure yielded, for each voxel, a time series of neural distance values reflecting trial-to-trial representational reconfiguration.

2.7. Statistical Analyses

2.7.1 Behavioral Model overview. The primary goal of the behavioral model was to identify factors that influence response time switch costs during hierarchical cognitive control. Specifically, the model tested how performance costs varied as a function of (1) the level of hierarchical rule reconfiguration (Stay, IDS, EDS), (2) trial-to-trial changes in perceptual cue features, and (3) interactions between hierarchical switching and perceptual change. Additional regressors were included to control for known effects of motor repetition, task repetition, and error-related adjustments.

2.7.2. Behavioral preprocessing and dependent variable (z-scored response times). The dependent variable was trial-wise response time, normalized as z-scored response times (zRTs). Specifically, we first normalized each participants' RT within each experimental block and response modality (i.e., yes versus no response). When normalizing RTs we did not include error

or post-error trials when calculating the mean or standard deviation because of extensive work showing error and post-error slowing effects (Rabbitt, 1966; McClure et al., 2003). We then applied the following exclusion criteria: (1) the first trial of each run, (2) any zRTs exceeding 3 standard deviations from the mean, (3) any trials with missing responses, (4) any post-error trials that were also erroneous.

2.7.3. Primary behavioral model predictors. For the model regressors, we included the following fixed effects. First, we included hierarchical task switching (Stay, IDS, EDS) as a categorical predictor to capture the effects of abstract rule structures on performance cost. To determine how perceptual representations of the cue attributes influence switch costs beyond abstract rules, we incorporated several perceptual factors: cue texture, color, and shape, along with the number of features that changed from the previous trial (range: 0-3).

We further included an additional binary regressor indicating whether the target feature repeated or switched from the previous trial. For example, if the relevant feature on trial N-1 was color (red), the target feature was coded as "repeated" if the cue on trial N was also red, and "switched" if the cue was a different color (e.g., blue). This regressor indexed how the previous trial's task-relevant sensory representation influences current trial performance.

Critically, we included an interaction between hierarchical task switching and the number of perceptual features that changed from the previous trial. This critical interaction tested our hypotheses about flexibility prioritization. If lower-level rule representations prioritize flexibility, we expected a significant positive interaction between IDS trials and feature change count. Conversely, if contextual representations prioritize stability, we expected no significant interaction between EDS trials and feature change.

2.7.4. Behavioral model control variables and conjunction related interactions. To account for known influences on response times, we included several control variables: task-rule repetition vs. switch, yes/no response repetition vs. switch, and probe-image repetition vs. switch. We also controlled for trial-by-trial performance fluctuations by including separate regressors for post-error trials and error trials to account for post-error slowing effects.

Beyond controlling for nuisance variance, the model incorporated a set of theoretically motivated interaction terms to replicate established task-switching effects. In particular, prior work suggests that flexible behavior is supported by conjunctive representations that integrate task rules, stimuli, and responses, and that the strength of such conjunctions can be indexed indirectly through partial switch costs (Kikumoto & Mayr, 2020; Grzyb & Hubner, 2013; Kleinsorge & Heuer, 1999). Partial switch costs arise when some, but not all, components of a task representation change across trials, requiring unbinding and rebinding of the prior conjunction (Li et al., 2019). Therefore, we further included a theoretically motivated interaction term: task-rule repetition with response repetition. This interaction terms tests the canonical partial switch costs, characterized by faster responses on full repetition and full switch trials relative to partial repetition/switch trials (Grzyb & Hubner, 2013; Kleinsorge & Heuer, 1999).

Because this interaction term indexing conjunction related effects could differ as a function of hierarchical context, we ran a second mixed effects regression model where we replaced hierarchical switch type (Stay/IDS/EDS) with a binary regressor indicated in context repeated or switched. This further allowed us to include two-way interactions between (1) context repeat and task repeat and (2) context repeat and response repeat, as well as the 3-way interaction of primary interest: context repeat by task repeat by response repeat. If contextual representations organize lower-level rule–stimulus–response associations, then partial switch

costs should primarily occur within contexts and be weaker across contexts. In other words, we should observe a significant three-way interaction between context, task, and response that is driven by the existence of (1) a significant two-way interaction between task and response within context, but (2) a non-significant two-way interaction between task and response when switching across context.

2.7.5. Behavioral model random effects structure. We treated subject as a random intercept and estimated random slopes for each individual cue feature (color, shape, and texture) to capture individual differences in sensitivity to these perceptual factors. Due to the complexity of the model, we kept the random effects structure as simple as possible to avoid convergence issues.

2.7.6. Behavioral (response time) supplementary analysis. It is important to note a subtle difference between IDS and EDS trials regarding trials where two features changed. For IDS, a two-feature change always reflects a change in both a task-relevant and task-irrelevant feature. In contrast, a two-feature change on EDS trials does not always mean a change in one relevant and one irrelevant feature. Sometimes a two-feature change on EDS trials means two relevant features change and no task-irrelevant features change. Therefore, we ran an additional follow-up paired samples t-test where we ensure EDS trials where two features changed were restricted to only trials where one feature was task-relevant and the other was task-irrelevant. We then contrasted these trials with trials where only one task-relevant feature changed.

2.7.7 Regression model for neural distance. For each subject, the neural distance time series was entered as the dependent variable in a linear regression model closely paralleling the behavioral model described in Section 2.6.3. The primary predictors included hierarchical task-switching condition (Stay, IDS, EDS), perceptual feature-change variables, motor repetition,

error and post-error indicators, and their theoretically motivated interactions. Unlike the behavioral model, regressors indexing the absolute texture, color, or shape of the current trial's cue were not included because the dependent variable was a measure of dissimilarity between the current trial and previous trial, whereas the behavioral model's dependent variable was a measure of the magnitude of the current trial cost.

To further control for potential RT effects on the BOLD response (Mumford et al., 2024), which was used to calculate neural distance, we added a regressor reflecting the magnitude of trial-to-trial changes in zRT (difference in zRT across adjacent trials) and a regressor reflecting the magnitude of the current trial's zRT .

2.7.8 Group analysis of neural distance. This procedure yielded, for each subject, a whole-brain beta map of each regressor that was then entered into a one-sample t-test across participants. To control the family-wise error rate for multiple comparisons, we employed AFNI's 3dClustSim using the residuals from the GLM procedure (Cox et al., 2017). First, spatial autocorrelation of the residuals was estimated using the autocorrelation function (ACF), producing three ACF parameters that characterize the spatial smoothness of the data. These parameters were then supplied to 3dClustSim to estimate the minimum cluster size for a corrected cluster-level threshold. Significant clusters for parametric modulation are based on these calculated minimum cluster sizes of 22 voxels (cluster p-value threshold of .001 and a voxel-level p-value threshold of .005).

2.8. Behavioral Switch Cost and Neural Distance Predictions. Our framework posits that representations supporting flexibility can be readily adapted by incorporating changing sensory inputs to modify action plans. Such flexibility, however, can come at a cost of increased

susceptibility to feed-forward sensory inputs from task-irrelevant information. Representations not prioritizing flexibility are better shielded from sensory inputs, thereby minimizing interference but making reconfiguration more effortful. From this framework, we derive specific, testable predictions for both behavioral performance (zRT switch costs) and neural reconfiguration (neural distance) that we describe in detail below.

For representations that prioritize flexibility, we predict increased sensitivity to interference from task-irrelevant perceptual changes. Specifically, we expect the reconfiguration of this representation to be more susceptible to interference from task-irrelevant information due to (1) smaller representational distance and (2) greater sensory-driven influence driving representational reconfiguration. As a result, interference from irrelevant features should slow RTs and increase neural distance. Accordingly, these predictions can be tested by examining the interaction between hierarchical task switching that only involved lower-level rule reconfiguration (IDS) and the number of features (i.e., texture, color, shape) that changed across trials. Specifically, IDS trials require that the contextual texture repeats while the relevant feature switches (e.g., color). If only one feature changes, then only the task-relevant dimension has changed. If multiple features change, then both task-relevant and task-irrelevant dimensions have changed. We predict that these changes in task-irrelevant features will increase both switch costs and neural distance in brain regions encoding representations that prioritize flexibility. This pattern reflects stronger bottom-up influence from sensory inputs during subordinate rule updating during hierarchical control.

For EDS trials, which require switching the context, we predict the opposite pattern. Because contextual representations prioritize stability, we hypothesize that they are separated by larger representational distances and thus should be less influenced by trial-to-trial changes in

perceptual and motor factors. This further shields against the processing of task-irrelevant features. Consequently, the interaction between the number of cue features changing across trials and context reconfiguration (EDS) should be weaker than that for either subordinate rule reconfiguration (IDS) or no rule reconfiguration (Stay). On EDS trials, even when multiple cue features change, irrelevant features should exert weaker effect on RTs and neural distances. In parallel, regions dedicated to reconfiguring contextual representation should exhibit larger neural distance for EDS when compared to IDS.

3. Results

We analyzed both behavioral (z-scored RTs) and neural representation (neural distance) measures using a common regression framework to identify factors that influence observable behavioral costs and neural reconfiguration costs during a hierarchical cognitive control task. Our central prediction was that prioritizing cognitive flexibility would reduce switch/reconfiguration costs, albeit at the expense of increased susceptibility to interference from task-irrelevant information. More specifically, subordinate rule switches would be faster than context switches, but at the cost of being more vulnerable to interference from task-irrelevant cue feature changes, whereas context switches would incur greater reconfiguration costs but reduced interference from task-irrelevant information.

To investigate potential mechanisms underlying reduced processing of task-irrelevant information during context reconfiguration, we investigate potential interactions between context and conjunctions—integrated representations incorporating rules, stimuli, and responses, that support cognitive flexibility (Kikumoto and Mayr, 2020). More specifically, we leverage partial switch costs as an indirect measure of conjunction strength, to test whether different contexts are split into distinct representational subspaces, within which context-dependent conjunctions are

embedded. Finally, we consider how post-error processes might interact with this underlying hierarchical task representation structure to drive internally driven error correction.

3.1. Distinct hierarchical level representations interact differently with task-relevant and -irrelevant perceptual influences during reconfiguration.

Traditionally, a hierarchical representation structure has been assessed by comparing the magnitude of switch costs across different levels of hierarchical rule switching (Fig. 3A, top panel). Larger switch costs when context switches relative to subordinate-rule switches have been taken as evidence of an underlying hierarchical representation structure (Cellier et al., 2022; Chen et al., 2024). The present dataset displays these expected switch cost effects of larger zRT switch costs for context switches (EDS) as compared to both subordinate rule switches (IDS; $t(53)=8.010$, $p<.001$), and rule repeats ($t(53)=18.128$, $p<.001$), as well as subordinate rule switches (IDS) as compared to rule repeats (Stay; $t(53)=11.195$, $p<.001$; Fig. 3B). Switch costs that increase with reconfiguration of increasingly superordinate rules suggest an underlying hierarchical organization of task representations in the present dataset.

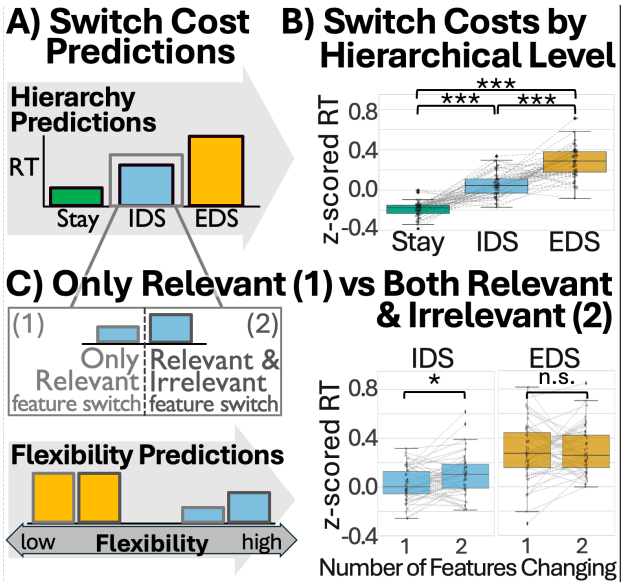

Figure 3. Response Time Switch Costs. **A**) Depiction of switch cost predictions for a hierarchical representation structure (top) and cognitive flexibility predictions for a more fine-grained switch cost analysis (bottom). **B**) Switch cost analyses comparing hierarchical rule switch conditions. Significance is based on two-tailed paired sample t-tests with zRTs. **C**) Switch cost comparisons

from the interaction between the hierarchical rule (Stay/IDS/EDS) and the number of features that changed from the last trial (perceptual change). Significance is based on post-hoc comparisons between estimated marginal means for hierarchical rule (Stay/IDS/EDS) and perceptual change.

However, focusing solely on the magnitude of switch costs across different levels of hierarchical rule switching neglects critical interactions between hierarchical task structure and the processing of task-relevant versus task-irrelevant information across trials. Additionally, previous investigations have largely overlooked trial-to-trial transition factors that arise from perceptual continuity or change. For example, whether task-relevant or task-irrelevant cue features repeat or change from one trial to the next. Concretely, a cue might switch from a solid red circle to a solid red square, in which case only the task-relevant feature (shape) changes while the task-irrelevant feature (color) repeats; alternatively, the cue might switch from a red circle to a blue square, in which case both the task-relevant and task-irrelevant features change. This is important to consider because, when cognitive flexibility is prioritized, faster switching may come at the cost of increased susceptibility to interference from task-irrelevant perceptual changes.

Therefore, we conducted more fine-grained analyses investigating how trial-to-trial changes in perceptual cue features interact with the reconfiguration of different levels of hierarchical representations. Because rule switches are indicated through changes in cue object features, we further divided switch trials (IDS and EDS) into two subgroups, those where only the task-relevant feature changed (i.e., trials where only one cue feature changed), and those where both task-relevant and task-irrelevant features changed (i.e., trials where two cue features changed). This distinction allowed us to test whether representations are shielded from

perceptual interference, as we expect for context representations, or are vulnerable to interference, as we expect for subordinate rule representations.

Consistent with our predictions (Fig. 3A-bottom), if subordinate rules prioritize flexibility by allowing greater bottom-up influence from perceptual input, then subordinate rule reconfiguration should be more susceptible to interference from changes in task-irrelevant cue features. Supporting this prediction, trial-level mixed-effects models (with random intercepts for subjects) revealed positive slopes for the number of features changing across trials (perceptual change) for both rule repeat (Stay; $\beta = 0.151$, $SE = 0.017$, $CI = [0.117, 0.184]$), and subordinate rules switch (IDS; $\beta = 0.072$, $SE = 0.023$, $CI = [0.026, 0.118]$) and a negative slope for context switch (EDS; $\beta = -0.051$, $SE = 0.023$, $CI = [-0.095, -0.007]$).

Follow-up contrasts confirmed that these positive slopes were associated with significantly slower response times when more cue features changed for both rule repeats (Stay; zero versus one; *mean difference* = 0.151, $t(19588) = 8.835$, $p < .001$), and subordinate rule switches (IDS; one versus two; *mean difference* = 0.072, $t(20227) = 3.063$, $p = .013$; Bonferroni corrected; Fig. 3C). This indicated that, as predicted, the additional task irrelevant feature change incurred an additional switch cost, i.e., slower response times, for subordinate rule switches. Moreover, for EDS trials there was no significant increase in response times when more cue features changed for context switch (EDS) trials (one versus two, *mean difference* = 0.051, $t(20888) = -2.266$, $p = .141$; Fig. 3C; one versus three, *mean difference* = 0.102, $t(20888) = -2.266$, $p = .141$, two versus three, *mean difference* = 0.051, $t(20888) = -2.266$, $p = .141$; Bonferroni corrected). This suggests that context reconfiguration is better shielded from potential feedforward perceptual interference from task-irrelevant feature changes. Note, because a two-feature change on EDS trials did not always mean both a task-relevant and task-irrelevant feature

change, we ran a supplementary post-hoc paired sample t-test where we ensured this was the case (see methods section *Behavioral (response time) supplementary analysis* for more details). Critically, this analysis also found no difference in the magnitude of zRT switch costs between one versus two features changing (paired-sample t-test $t(53) = -0.010$, $p = .797$).

Finally, we found a significant interaction between these perceptual change slopes for the different rule reconfiguration conditions (i.e., IDS versus EDS), indicating that this slope is significantly more positive for IDS trials as compared to EDS trials ($\beta = 0.123$, $t(20860) = 4.123$, $p < .001$). To summarize, these results support our hypotheses on cognitive flexibility. Specifically, the degree of perceptual change in the cue object interacts with rule reconfiguration differently depending on which hierarchical rule level is being reconfigured. Subordinate rule representations support flexibility at the cost of greater perceptual interference during reconfiguration, whereas contextual representations are less influenced by perceptual interference during reconfiguration.

To investigate potential neural underpinnings of these zRT switch cost effects, we conducted a parallel analysis using trial-to-trial changes in neural representational patterns ("neural distance") as a neural analogue of switch costs. We applied the same regression model used for the behavioral switch cost analysis to predict neural distance. We included zRT and the difference in zRT across adjacent trials as additional predictors to control for potential effects of RTs on neural dissimilarity patterns (Mumford et al., 2024).

Based on prior work implicating the middle frontal gyrus (MFG) in contextual encoding and hierarchical cognitive control (Badre & D'Esposito, 2007; Badre & Nee, 2018; Nee & D'Esposito, 2016), we predicted that context reconfiguration (EDS) would elicit a larger neural distance in MFG than subordinate rule reconfiguration (IDS). In parallel with behavioral

findings, we further predicted that EDS reconfiguration would be relatively insensitive to perceptual interference. In contrast, we predicted that IDS reconfiguration would show an interaction with perceptual interference (Fig. 4A). Specifically, we predicted that neural distance in MFG during EDS trials would be unaffected by the number of perceptual cue features changing across trials. In contrast, we predicted that premotor and visual attention regions would show larger neural distances on IDS trials when a greater number of perceptual features changed, reflecting both the encoding of subordinate rules in premotor regions (Nee & Brown, 2013; Badre & Frank, 2012) and increased processing of task-irrelevant perceptual information during subordinate rule reconfiguration.

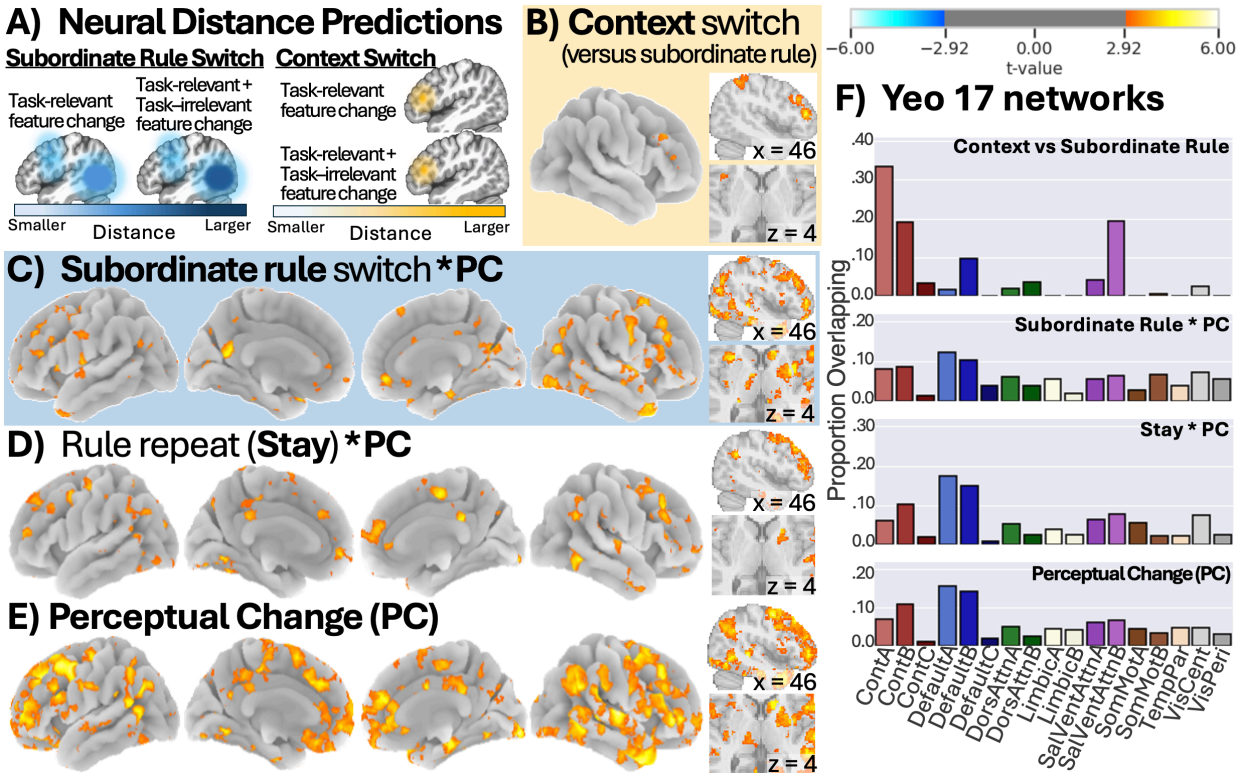

Figure 4. Neural Distance. **A**) Depiction of expected effects for context reconfiguration versus subordinate rule reconfiguration. **B**) Observed main effect of context reconfiguration contrasted with subordinate rule reconfiguration—significant clusters in MFG and IFS, regions belonging to the control and salient/ventral attention networks. Note, there were no significant clusters for the main effect of subordinate rule contrasted with stay (rule repeat) trials. **C-D**) Interactions between perceptual change (i.e., number of cue object features changing across trials; range 0-3) and subordinate rule reconfiguration (**C**) or rule repeat (Stay) trials (**D**). **E**) Main effect of

perceptual change. **F**) Bar plots depicting overlap between significant clusters from **B-E** and the Yeo 17 networks.

In line with these predictions, we identified two significant clusters in lateral frontal regions showing greater neural distance for EDS compared to IDS (IFS and caudal MFG; Fig. 4B). In contrast, widespread clusters showed significantly greater neural distance for the overall main effect of the number of changing cue features (Fig. 4E) and its interactions with both IDS trials (Fig. 4C) and Stay trials (Fig. 4D). Within these regions, the neural distance was greater when both task-relevant and task-irrelevant features changed compared to when only the task-relevant feature changed, consistent with increased susceptibility to perceptual interference during subordinate rule processing. These clusters were localized in a widely distributed fashion, such as those associated with the dorsal and ventral attention networks, the visual networks, as well as the default mode networks (Fig. 4F).

Regarding the control variables, after including the primary measures of interest (hierarchical switch type and perceptual change), we found no significant response time switch costs associated with task switch ($\beta = 0.018$, $t(20880) = 0.506$, $p = .613$), response switch ($\beta = 0.019$, $t(20880) = 0.683$, $p = .495$), or probe switch ($\beta = 0.027$, $t(20880) = 1.071$, $p = .284$). On the other hand, we did find a significant switch cost associated with previous target feature switches ($\beta = 0.063$, $t(20900) = 3.040$, $p = .002$). For the neural distance regression, we found clusters associated with a significantly larger neural distance for response, probe, and previous target feature switches. For response switching, we found significantly greater neural distance in left motor cortex and visual cortex (Fig. S1A), a pattern that is expected given that our task required responses to be made with the right hand. For probe switches, we found significantly greater neural distance in ventral occipital and temporal regions (Fig. S1B), such as fusiform face

area (FFA) and parahippocampal place area (PPA), as expected given that our task used face and place images. Finally, for target feature switches, we found significantly greater neural distance in distributed clusters, including caudal PFC, primary visual cortex, right parahippocampus to inferior and middle temporal gyrus, and the right putamen (Fig. S1C). We limit further discussion on these control variables as they are not of primary interest.

Taken together, these results suggest that the slowing of response times during IDS trials with multiple feature changes reflects larger shifts in representational patterns across distributed regions. These larger neural distances likely index greater sensitivity to task-irrelevant feature changes, consistent with increased bottom-up influence during subordinate rule reconfiguration. By contrast, modulation of neural distance in MFG suggests that context reconfiguration elicits large representational shifts irrespective of the number of perceptual features changing across trials, consistent with reduced susceptibility to perceptual feedforward interference.

3.2. Conjunction representations supporting cognitive flexibility are context dependent.

Prior work suggests cognitive flexibility, i.e., fast and efficient switching between rules/tasks, is supported by the creation of conjunctive representations (Kikumoto and Mayr, 2020). These conjunctive representations consist of rules, stimuli, and responses. A common measure of conjunctions has been partial switch costs—greater switch costs when only a subset of conjunction components change across trials, as compared to trials where all components switch or repeat (Grzyb & Hubner, 2013; Kleinsorge & Heuer, 1999). When all components of a conjunction repeat, then the previous trial's conjunction can be re-instantiated. However, when some components repeat and some switch, the previous conjunction needs to first be unbound so that a new conjunction can be created with the components that repeat across trials. Furthermore, the conjunction components that repeat across trials can cause increased interference when

sensory-motor plans differ between these two trials. Thus, this unbinding and rebinding in addition to increased interference that needs to be resolved cause the observed "cost'. When all components of a conjunction switch, the unbinding process can be skipped and the lack of shared components with the previous trial means no additional interference needs to be overcome. However, a new conjunction still needs to be constructed, which is why response times are still slower overall than trials where the previous conjunction can be re-instated and utilized.

Previous work on conjunctive representations has largely examined tasks with a single level of rule structure (Kikumoto and Mayr, 2020; Rangel et al., 2023). As such, it remains an open question as to whether conjunctions consist of only the lower-level rule, or if they include higher-order context-level information that our task was designed to probe. Recent evidence suggests that conjunctive representations may be organized into context-specific subspaces, enabling more efficient readout of task-relevant information (Kikumoto et al., 2024). Given that in our task, context indicates if color or shape is the currently relevant rule, we predicted that conjunction-related partial switch costs would be observed within each context, but not when switching across contexts. Under this account, context serves to index distinct conjunctive subspaces, such that switching context both circumvents the need to unbind the previous conjunction and reduces interference from the previous conjunction.

To measure this, we first ran a modified version of the above RT mixed effects model where the hierarchical rule switch (Stay/IDS/EDS) was replaced with a variable indicating if context switched or repeated (2 levels). This allowed us to run a three-way interaction between context, task, and response, while controlling for all the same additional variables as before. As expected, we found a significant three-way interaction ($\beta = 0.128$, $t(20930) = 2.142$, $p = .032$). Post-hoc comparisons revealed that the two-way interaction between task and response, which

reflects partial switch costs, was only significant when context repeated (β = -0.124, *t(20916)* = -2.300, *p* = .022; Fig. 5A), but not when context switched (β = 0.004, *t(20926)* = 0.065, *p* = .949; Fig. 5B). To replicate this result, we ran a follow-up behavioral experiment with the same experimental paradigm. As expected, we found a significant three-way interaction (β = 0.251, *t(21690)* = 4.173, *p* < .001). Post-hoc comparisons revealed that the two-way interaction between task and response, which reflects partial switch costs, was only significant when context repeated (β = -0.247, *t(21704)* = -4.891, *p* < .001; Fig. 5C), but not when context switched (β = 0.004, *t(21687)* = 0.056, *p* = .955; Fig. 5D).

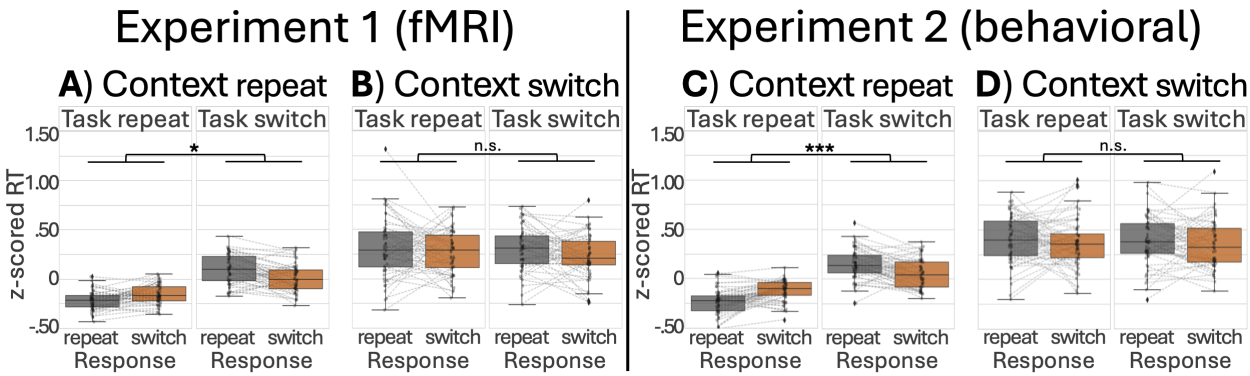

Figure 5. Partial Switch Costs and Context. **A**) Bar plots showing expected partial switch cost effects when context repeats. **B**) Bar plots showing lack of partial switch costs when context switches. Significance based on post-hoc comparisons between estimated marginal means for 3way interaction between context, task, and response repeat/switch.

The existence of a significant two-way interaction between task and response within context matches prior findings of partial switch costs (Kikumoto and Mayr, 2020; Rangel et al., 2023). In contrast, the lack of a significant two-way interaction when switching contexts suggests that partial switch costs are weakened during context reconfiguration. These results lend further support to the idea that cognitive flexibility is supported by creating context-specific conjunctive representations, which enable faster subordinate rule switching within context and decreased interference when switching across context.

To address the neural underpinnings of partial switch costs indicative of potential context-specific conjunctive representations, we conducted a parallel analysis with a very similar regression model to predict neural distance. Key differences were that cue color, shape, and texture were removed (similar to the regression predicting neural distance reported in 3.1) and regressors indicating the zRT and zRT difference between adjacent trials were included. However, we found no significant clusters associated with a three-way interaction between context, task, and response repeat/switch. This suggests that, although the observed partial switch cost effect on context repeat trials, but not context switch trials, is robust for response times, its neural underpinnings likely arise from a combination of distributed neural representations rather than from anatomically focal, restricted sources.

3.3. Distinct hierarchical level interactions with post-error correction.

The above results further implicate error detection and correction processes as operating over the underlying hierarchical structure of task representations. Error detection and correction processes must utilize the underlying task representation structure to properly attribute the source of the error and re-establish the appropriate task representation to guide subsequent behavior. Although the initial cause of an error has multiple possibilities (e.g., instantiating the wrong context, erroneous button presses, lapses in attention), post-error correction should be more uniform in the sense that it re-engages the underlying task representation structure. Therefore, more effective post-error correction should be accompanied by greater reconfiguration of multivoxel activity patterns.

To investigate interactions between the underlying hierarchical task representation structure and error processes, we included error and post-error trials in the regressions predicting

zRTs and neural distance. The average error rate (Fig. 6B-left) was .034 (SE = .004) for Stay trials, .051 (SE = .005) for IDS trials, and .065 (SE = .006) for EDS trials. Post-error trial proportions (Fig. 6B-right) were highly similar between task switching conditions (Stay M=.048, SE=.004; IDS M=.046, SE=.005; EDS M=.053, SE=.005). Consistent with existing literature (Rabbitt, 1966; McClure et al., 2003), the mixed effects regression predicting zRTs showed significantly slower zRTs on error trials ($t(20,870) = 10.504$, $p < .001$), and post-error slowing ($t(20,820) = 2.359$, $p = .018$). Note, the experimental paradigm reported here provided no feedback. Therefore, the observed post-error slowing of RTs and greater neural distance following error trials suggest that participants engaged in internal error monitoring in the absence of external feedback. To ensure we were looking at error correction on post-error trials, we restricted analyses to post-error trials where participants performed accurately following an error.

The neural distance results showed no significant clusters associated with errors on the current trial. However, post-error trials were associated with significant increases in neural distance in distributed clusters, indicating substantial representational reconfiguration following errors (Fig. 6A). Critically, we observed cortical clusters in mPFC, parahippocampus, as well as subcortical clusters in the basal ganglia and mediodorsal thalamus.

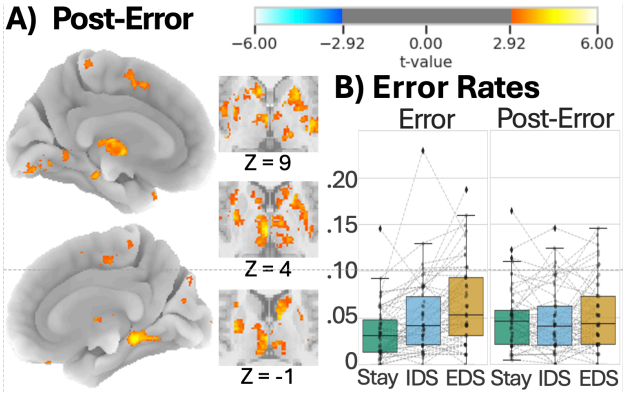

Figure 6. Errors and Post-error Correction. A) Significant clusters with a larger neural distance on post-error trials B) Error and Post-error trial rates by hierarchical rule condition (Stay/IDS/EDS).

4. Discussion

In this study, we investigated the behavioral (response time) and neural correlates of switch costs associated with hierarchical cognitive control. We were primarily interested in how the reconfiguration of distinct cognitive representations would interact with (1) the underlying hierarchical task representation structure and (2) other systems/processes necessary for goal-directed behavior, such as perceptual or error processes. To address these questions, we combined a hierarchical task design with trial-by-trial behavioral switch cost analyses and a novel multivoxel pattern–based measure of "neural distance", allowing us to characterize both performance costs and representational reconfiguration during hierarchical cognitive control. In the sections that follow, we interpret these findings in the context of existing theories of cognitive control and representational flexibility.

The behavioral phenomenon of performance costs on trials requiring a task switch is well-established (Monsell, 2003; Rogers & Monsell, 1995; Wylie & Allport, 2000; Allport et al., 1994). However, relatively fewer studies have investigated the neural underpinning of behavioral switch costs, particularly how they relate to the internal structure of task representations. A previous study found that greater dissimilarity in multivoxel activity patterns on task switch trials predicted greater response time task switch costs (Qiao et al., 2017). Nevertheless, given the multifaceted nature of task representations (Schumacher & Hazeltine, 2016; Botvinick, 2008), several critical questions remain unresolved: whether representational reconfiguration operates similarly across different hierarchical levels, how abstract contextual representations interact or are interfered with by changes in perceptual input, and whether perceptual and error-related processes shape reconfiguration in ways that support cognitive flexibility. The present study

addresses these gaps by extending a novel neural distance analyses to a hierarchical task structure that specifically investigates contextual updating and subordinate rule reconfiguration, allowing us to directly test how different representational levels vary in their susceptibility to perceptual interference.

A complementary line of work suggests that stimulus-response associations are bound into an event file (Hommel, 1998, 2004), or conjunction (Kikumoto & Mayr, 2020). When this integrated sensory-motor representation can be re-used across trials, performance is faster and more accurate than when this integrated representation needs to be updated. Additionally, performance is slower and less accurate when there is partial overlap between components across adjacent trials—commonly referred to as partial switch costs (Grzyb & Hubner, 2013; Kleinsorge & Heuer, 1999). Despite this, most previous hierarchical control studies have not systematically investigated whether task-relevant and irrelevant perceptual features repeat or switch across trials, nor have they examined how such feature-level continuity interacts with hierarchical task structure. As a result, how different hierarchical levels are differentially affected by perceptual interference has remained largely unexplored. By considering these factors when investigating the behavioral and neural correlates of switch costs, the present study reveals novel findings regarding interactions between the reconfiguration of distinct hierarchical levels and perceptual interference.

Specifically, we found a clear distinction between context reconfiguration and subordinate rule reconfiguration in the degree of susceptibility to feedforward perceptual interference. Subordinate rule reconfiguration showed the greatest amount of feedforward perceptual influence on reconfiguration processes. When task-irrelevant features changed across trials, we observed slower RTs and a greater neural distance in distributed regions, notably,

premotor and visual attention regions. Presumably, this increased dissimilarity in trial-to-trial multivoxel activity patterns when task-irrelevant features changed reflected greater processing of this irrelevant perceptual event. In contrast, the lack of modulation of neural distance in MFG (i.e., equally large neural distance regardless of whether task-irrelevant cue features changed) suggested that context reconfiguration was better shielded from potential feedforward perceptual interference. Given prior work suggesting MFG encodes high dimensional, abstract, task rule structures (Badre et al., 2021; Badre & Nee, 2018), this reduced susceptibility to interference is critical for successful goal-directed behavior.

One explanation for why context reconfiguration appears to be better shielded from feedforward perceptual interference is that contextual information may organize more lower-level sensory-motor representations into distinct latent representational spaces. Recent work by Kikumoto and colleagues (2024) suggests that the control system might encode conjunctive representations (integrated representations of rules, stimuli, and responses) in distinct latent subspaces. More specifically, these distinct representational subspaces, indexed by a contextual cue, can contain context-specific conjunctions that drive cognitive flexibility. Context-specific conjunctions would allow for a smaller representational distance within context, meaning greater flexibility, i.e., faster switching, while also shielding the current goal-relevant conjunctions from interference across contexts. This greater flexibility within context likely drives the observed partial switch costs when context repeats. The greater representational distance that must be traveled when switching contextual subspaces could explain the overall slower response times and lack of partial switch costs when switching context. Presumably, the latter would be due to the lack of interference from the previous context-specific conjunction. Moreover, the lack of a robust 3-way interaction between context, task, and response repeat/switch predicting neural

distance suggests that the neural underpinnings of partial switch costs arise form distributed neural assemblies and not an anatomically focal, or restricted, source.

Organizing task representations into context-specific latent spaces may confer additional benefits. By separating contexts, our control system can more optimally learn subordinate rules and their unique stimulus-response associations that are specific to each context. This would allow for greater context-dependent flexibility while reducing interference across context. Such representational organization may also support adaptive error processing. Related to this idea, we observed significant post-error slowing on post-error trials, accompanied by increased neural distance in the mediodorsal thalamus. It is possible that this increased neural distance reflects context-related reconfiguration following error detection, whereby the mediodorsal thalamus contributes to updating or re-engaging the appropriate contextual representation to guide subsequent behavior. This interpretation is consistent with our prior work demonstrating that the mediodorsal thalamus encodes contextual information and exhibits signals related to context and latent state updating during hierarchical cognitive control (Chen et al., 2024; Leach et al., 2025), suggesting a role for this region in resetting or stabilizing task representations after errors.

The notion that context-dependent cognitive flexibility involves distinct representational subspaces also points to an interesting avenue for future research with the arousal system. Hierarchical contexts can be understood as internal control states that govern which sets of task rules and stimulus–response associations are currently active. Shifting between such contexts therefore requires coordinated reconfiguration of internal states, rather than simple updating of stimulus–response mappings. From this perspective, the neuromodulatory systems that facilitate state transitions may play a critical role in hierarchical cognitive control. For example, Bouret and Sara (2005) suggest the locus coeruleus (LC) norepinephrine (NE) system acts as a network

reset system, making it easier to shift internal states. Although previous has largely focused on how this impacts latent, or hidden, state shifting (Nassar, 2024), this same mechanism could support reconfiguration of contextual representations during hierarchical control. Specifically, NE release may make it easier to shift, or reconfigure, context by disrupting recurrent activity maintaining the current context in MFG. If two distinct activity patterns encode each context and the current context is maintained through recurrent connections, then disrupting this recurrent activity would make it easier for an external switch signal to shift to the activity pattern encoding the alternative context. Another explanation is that post-error slowing is context specific because ongoing associative learning and error correction needs to be context specific. In other words, errors must be properly attributed to the correct context for optimal goal-directed behavior. This latter point relates to the literature on latent, or hidden, states, which require predictive inference specific to distinct latent states (Nassar et al., 2019). The LC/NE system has been proposed to help with latent state shifting at contextual change points to aid predictive learning (Nassar, 2024). Although the present task utilizes cue objects that explicitly indicate context, it could engage similar mechanism, including the need to disrupt recurrent PFC activity to aid context switching and the need to properly assign errors to the appropriate context when engaging error correction systems. Therefore, future work on hierarchical task control should incorporate estimates of arousal system activity. For example, how the arousal system might relate to context reconfiguration, even in the absence of salient error events.

In conclusion, we report three major findings regarding switch costs and hierarchical cognitive control. First, trial-to-trial reconfiguration of neural activity patterns encoding distinct hierarchical levels drive observable switch costs. Greater reconfiguration of neural activity patterns predicted larger observable switch costs. Second, the reconfiguration of distinct

hierarchical representation levels is differentially impacted by feedforward perceptual information, as is evident through both response time and neural distance measure. More specifically, subordinate rule representations associated with more concrete sensory-motor plans prioritized greater flexibility at the cost of greater interference from feedforward perceptual information. In other words, showing slower response times and greater neural activity pattern reconfiguration when task irrelevant cue features changed across trials. To counter this increased interference, contextual representations prioritized shielding information from feedforward perceptual interference at the cost of reduced flexibility. In other words, showing overall slower response times and greater neural distance that is not modulated by task-irrelevant cue feature changes. This likely reflects a balance of cognitive flexibility that maximizes fast and efficient switching of concrete sensory-motor plans while also shielding internally maintained contextual rules from interference. Finally, response time switch cost results tentatively suggest this shielding is achieved by partitioning context into distinct representational, or latent, spaces.

References

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences, 7*(3), 134-140. doi:10.1016/s1364-6613(03)00028-7

Meiran, N., Chorev, Z., & Sapir, A. (2000). Component processes in task switching. *Cogn Psychol, 41*(3), 211-253. doi:10.1006/cogp.2000.0736

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*(2), 207-231. doi:10.1037/0096-3445.124.2.207

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. (2007). Functional Magnetic Resonance Imaging Evidence for a Hierarchical Organization of the Prefrontal Cortex. *Journal of Cognitive Neuroscience, 19*(12), 2082-2099.

Badre, D., Hoffman, J., Cooney, J. W., & D'Esposito, M. (2009). Hierarchical cognitive control deficits following damage to the human frontal lobe. *Nat Neurosci, 12*(4), 515-522. doi:10.1038/nn.2277

Nee, D. E., & Brown, J. W. (2013). Dissociable frontal-striatal and frontal-parietal networks involved in updating hierarchical contexts in working memory. *Cereb Cortex, 23*(9), 2146-2158. doi:10.1093/cercor/bhs194

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review Neuroscience, 24*.

Nee, D. E., & D'Esposito, M. (2017). Causal evidence for lateral prefrontal cortex dynamics supporting cognitive control. *Elife, 6*. doi:10.7554/eLife.28040

Geddert, R., & Egner, T. (2024). Contextual control demands determine whether stability and flexibility trade off against each other. *Atten Percept Psychophys, 86*(7), 2529-2551. doi:10.3758/s13414-024-02955-x

Dreisbach, G., & Fröber, K. (2018). On How to Be Flexible (or Not): Modulation of the Stability-Flexibility Balance. *Current Directions in Psychological Science, 28*(1), 3-9. doi:10.1177/0963721418800030

Mayr, U., & Grätz, D. (2024). Does cognitive control have a general stability/flexibility tradeoff problem? *Current Opinion in Behavioral Sciences, 57*. doi:10.1016/j.cobeha.2024.101389

Kikumoto, A., & Mayr, U. (2020). Conjunctive representations that integrate stimuli, responses, and rules are critical for action selection. *Proceedings of the National Academy of Sciences, 117*(19), 10603-10608. doi:10.1073/pnas.1922166117

Kikumoto, A., Bhandari, A., Shibata, K., & Badre, D. (2024). A transient high-dimensional geometry affords stable conjunctive subspaces for efficient action selection. *Nat Commun, 15*(1), 8513. doi:10.1038/s41467-024-52777-6

Cellier, D., Petersen, I. T., & Hwang, K. (2022). Dynamics of Hierarchical Task Representations. *J Neurosci, 42*(38), 7276-7284. doi:10.1523/JNEUROSCI.0233-22.2022

Schumacher, E. H., & Hazeltine, E. (2016). Hierarchical task representation: Task files and response selection. *Current Directions in Psychological Science, 25*(6), 449-454. doi:10.1177/0963721416665085

Dykstra, T., Smith, D. M., Schumacher, E. H., & Hazeltine, E. (2022). Measuring task structure with transitional response times: Task representations are more than task sets. *Psychon Bull Rev, 29*(5), 1812-1820. doi:10.3758/s13423-021-02035-3

Wylie, G., & Allport, A. (2000). Task switching and the measurement of switch costs. *Psychological Research, 63*, 21.

Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. *Neuron, 109*(19), 3055-3068. doi:10.1016/j.neuron.2021.07.011

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI--an introductory guide. *Soc Cogn Affect Neurosci, 4*(1), 101-109. doi:10.1093/scan/nsn044

Qiao, L., Zhang, L., Chen, A., & Egner, T. (2017). Dynamic Trial-by-Trial Recoding of Task-Set Representations in the Frontoparietal Cortex Mediates Behavioral Flexibility. *J Neurosci, 37*(45), 11037-11050. doi:10.1523/JNEUROSCI.0935-17.2017

Chen, X., Leach, S. C., Hollis, J., Cellier, D., & Hwang, K. (2024). The thalamus encodes and updates context representations during hierarchical cognitive control. *PLoS Biol, 22*(12), e3002937. doi:10.1371/journal.pbio.3002937

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., . . . Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat Methods, 16*(1), 111-116. doi:10.1038/s41592-018-0235-4

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., & Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage, 54*. doi:10.1016/j.neuroimage.2010.07.033314

Prince, J. S., Charest, I., Kurzawski, J. W., Pyles, J. A., Tarr, M. J., & Kay, K. N. (2022). Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *Elife*. doi:10.7554/eLife.77599

Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F., & Wandell, B. A. (2013). GLMdenoise: a fast, automated technique for denoising task-based fMRI data. *Front Neurosci, 7*, 247. doi:10.3389/fnins.2013.00247

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

Rabbitt, P. A. (1966). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology, 71*(2).

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron, 38*.

Grzyb, K. R., & Hubner, R. (2013). Excessive response-repetition costs under task switching How response inhibition amplifies response conflict. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 13.

Kleinsorge, T., & Heuer, H. (1999). Hierarchical switching in a multi-dimensional task space. *Psychological Research, 62*, 12.

Li, J., Cao, B., Han, J., Xie, L., & Li, F. (2019). Not inertia but reconfiguration: Asymmetric switch cost in a hierarchical task. *Brain Res, 1720*, 146291. doi:10.1016/j.brainres.2019.06.010

Mumford, J. A., Bissett, P. G., Jones, H. M., Shim, S., Rios, J. A. H., & Poldrack, R. A. (2024). The response time paradox in functional magnetic resonance imaging analyses. *Nat Hum Behav, 8*(2), 349-360. doi:10.1038/s41562-023-01760-0

Cox, R. W., Chen, G., Glen, D. R., Reynolds, R. C., & Taylor, P. A. (2017). fmri clustering and false-positive rates. *PNAS, 114*(17), 2. doi:10.1073/pnas.1614961114Fig

Badre, D., & Nee, D. E. (2018). Frontal Cortex and the Hierarchical Control of Behavior. *Trends in Cognitive Science, 22*(2), 170-188. doi:10.1016/j.tics.2017.11.005

Nee, D. E., & D'Esposito, M. (2016). The hierarchical organization of the lateral prefrontal cortex. *Elife, 5*, e12112. doi:10.7554/eLife.12112

Badre, D., & Frank, M. J. (2012). Mechanisms of hierarchical reinforcement learning in cortico-striatal circuits 2: evidence from fMRI. *Cereb Cortex, 22*(3), 527-536. doi:10.1093/cercor/bhr117

Allport, A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set Exploring the dynamic control of tasks.

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends Cogn Sci, 12*(5), 201-208. doi:10.1016/j.tics.2008.02.009

Hommel, B. (1998). Event Files: Evidence for Automatic Integration of Stimulus-Response Episodes. *Visual Cognition, 5*(1-2), 183-216. doi:10.1080/713756773

Hommel, B. (2004). Event files: feature binding in and across perception and action. *Trends Cogn Sci, 8*(11), 494-500. doi:10.1016/j.tics.2004.08.007

Nassar, M. R. (2024). Toward a computational role for locus coeruleus/norepinephrine arousal systems. *Curr Opin Behav Sci, 59*. doi:10.1016/j.cobeha.2024.101407

Nassar, M. R., Bruckner, R., & Frank, M. J. (2019). Statistical context dictates the relationship between feedback-related EEG signals and learning. *Elife, 8*. doi:10.7554/eLife.46975

Acknowledgments

Author contributions: SCL wrote the first draft of the paper, SCL and KH edited the paper, KH and SCL designed research, SCL and XC performed research, SCL and KH analyzed data, SCL wrote the paper.

Authors declare that they have no competing interest.

All data and code will be made available upon publication.

Research reported here was supported by a National Institutes of Mental Health grant R01MH122613 (KH) and the Iowa Neuroscience Institute. This work was conducted on an MRI instrument funded by 1S10OD025025-01

Figure 1. Hierarchical control task design. A) Depiction of the task rule hierarchy embedded within the 8 potential cue objects. B) Example trial sequence: A cue object presented for 500ms followed by an image of either a face or a scene for 2000ms and a jittered ITI (range: 1.5-10.5 seconds; average: 4.25 seconds. Example cue transitions across trials depicts how perceptual change is operationalized. Example face image is an author on this manuscript.

Figure 2. Neural Distance Method. A) depiction of how neural distance is calculated on adjacent trial multivoxel activity patterns. A greater change in multivoxel activity patterns is associated with a greater correlation, or neural distance.

Figure 3. Response Time Switch Costs. **A**) Depiction of switch cost predictions for a hierarchical representation structure (top) and cognitive flexibility predictions for a more fine-grained switch cost analysis (bottom). **B**) Switch cost analyses comparing hierarchical rule switch conditions. Significance is based on two-tailed paired sample t-tests with zRTs. **C**) Switch cost comparisons from the interaction between the hierarchical rule (Stay/IDS/EDS) and the number of features that changed from the last trial (perceptual change). Significance is based on post-hoc comparisons between estimated marginal means for hierarchical rule (Stay/IDS/EDS) and perceptual change.

Figure 4. Neural Distance. **A**) Depiction of expected effects for context reconfiguration versus subordinate rule reconfiguration. **B**) Observed main effect of context reconfiguration contrasted with subordinate rule reconfiguration—significant clusters in MFG and IFS, regions belonging to the control and salient/ventral attention networks. Note, there were no significant clusters for the main effect of subordinate rule contrasted with stay (rule repeat) trials. **C-D**) Interactions between perceptual change (i.e., number of cue object features changing across trials; range 0-3) and subordinate rule reconfiguration (**C**) or rule repeat (Stay) trials (**D**). **E**) Main effect of perceptual change. **F**) Bar plots depicting overlap between significant clusters from **B-E** and the Yeo 17 networks.

Figure 5. Partial Switch Costs and Context. **A**) Bar plots showing expected partial switch cost effects when context repeats. **B**) Bar plots showing lack of partial switch costs when context switches. Significance based on post-hoc comparisons between estimated marginal means for 3way interaction between context, task, and response repeat/switch.

Figure 6. Errors and Post-error Correction. A) Significant clusters with a larger neural distance on post-error trials B) Error and Post-error trial rates by hierarchical rule condition (Stay/IDS/EDS).