Neuroscience

Conflicts are represented in a cognitive space to reconcile domain-general and domain-specific cognitive control

Guochun Yang, Haiyan Wu, Qi Li, Xun Liu, Zhongzheng Fu, Jiefeng Jiang

CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing 100101, China • Department of Psychology, University of Chinese Academy of Sciences, Beijing 100101, China • Department of Psychological and Brain Sciences, University of Iowa, Iowa City, IA 52242, USA • Cognitive Control Collaborative, University of Iowa, Iowa City, IA 52242, USA • Centre for Cognitive and Brain Sciences and Department of Psychology, University of Macau, Taipa, Macau 999078, China • Beijing Key Laboratory of Learning and Cognition, School of Psychology, Capital Normal University, Beijing 100048, China • Department of Neurosurgery, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA • Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2
September 12, 2023 (this version)

Reviewed preprint version 1
May 9, 2023

Sent for peer review
February 21, 2023

Posted to bioRxiv
February 14, 2023

Abstract

Cognitive control resolves conflicts between task-relevant and -irrelevant information to enable goal-directed behavior. As conflicts can arise from different sources (e.g., sensory input, internal representations), how a finite set of cognitive control processes can effectively address potentially infinite conflicts remains a major challenge. Based on the cognitive space theory, different conflicts can be parameterized and represented as distinct points in a (low-dimensional) cognitive space, which can then be resolved by a limited set of cognitive control processes working along the dimensions. It leads to a hypothesis that conflicts similar in their sources are also represented similarly in the cognitive space. We designed a task with five types of conflicts that could be conceptually parameterized. Both human performance and fMRI activity patterns in the right dorsolateral prefrontal (dlPFC) support that different types of conflicts are organized based on their similarity, thus suggesting cognitive space as a principle for representing conflicts.

eLife assessment

Yang et al. investigate whether distinct sources of conflict are represented in a common cognitive space. The study uses an interesting task that mixes different sources of difficulty and reports that the brain appears to represent these sources as a mixture on a continuum in prefrontal areas. While the findings could be **valuable** to theory in this area, there are concerns with the analysis, design and results, that raise uncertainty regarding the main conclusion of a shared cognitive space. Thus, the evidence reported here ranges from **solid** to **incomplete**.

Introduction

Cognitive control enables humans to behave purposefully by modulating neural processing to resolve conflicts between task-relevant and task-irrelevant information. For example, when naming the color of the word "BLUE" printed in red ink, we are likely to be distracted by the word meaning, because reading a word is highly automatic in daily life. To keep our attention on the color, we need to mobilize the cognitive control processes to resolve the conflict between the color and word by boosting/suppressing the processing of color/word meaning. As task-relevant and task-irrelevant information can come from different sources, the sources of conflicts and how they should be resolved can vary greatly (Kornblum et al., 1990). For example, the conflict may occur between items of sensory information, such as between a red light and a police officer signaling cars to pass. Alternatively, conflict may occur between sensory and motor information, such as when a voice on the left asks you to turn right. The large variety of conflict sources implies that there may be unlimited number of conflict conditions. A key unsolved question in cognitive control is how our brain efficiently resolves these different types of conflicts.

A first step to addressing this question is to examine the commonalities and/or dissociations across different types of conflicts that can be categorized into different *domains*. Examples of the domains of conflicts include experimental paradigm (Freitas et al., 2007; Magen & Cohen, 2007), sensory modality (Hazeltine et al., 2011; Yang et al., 2017), or conflict type regarding the dimensional overlap of conflict processes (Jiang & Egner, 2014; Liu et al., 2004).

Two solutions to resolving a wide range of conflict types are proposed. They differ based on whether the same cognitive control mechanisms are applied across domains. On the one hand, the *domain-general* cognitive control theories posit that the frontoparietal cortex adaptively encodes task information and can thus flexibly implement control strategies for different types of conflicts. This is supported by the generalizable control adjustment (i.e., encountering a conflict trial from one type can facilitate conflict resolution of another type) (Freitas et al., 2007; Kan et al., 2013) and similar neural patterns (Peterson et al., 2002; Wu et al., 2020) across distinct conflict tasks. A broader domain-general view holds that the frontoparietal brain regions/networks are widely involved in multiple control demands well beyond the conflict domain (Assem et al., 2020; Cole et al., 2013), which explains the remarkable flexibility in human behaviors. However, since domain-general processes are by definition likely shared by different tasks, when we need to handle multiple task demands at the same time, the efficiency of both tasks would be impaired due to resource competition or interference (Musslick & Cohen, 2021). Therefore, the domain-general processes is evolutionarily less advantageous for humans to deal with the diverse situations requiring high efficiency (Cosmides & Tooby, 1994). On the other hand, the *domain-specific* theories argue that different types of conflicts are handled by distinct cognitive control processes (e.g., where and how information processing should be modulated)(Egner, 2008; Kim et al., 2012). However, according to the domain-specific view, the potentially unlimited conflict situations require a large variety of preexisting control processes, which is biologically implausible (Abrahamse et al., 2016).

To reconcile the two theories, researchers recently proposed that cognitive control might be a mixture of domain-general and domain-specific processes. For instance, Freitas and Clark (2015) found that trial-by-trial adjustment of control can generalize across two conflict domains to different degrees, leading to domain-general (strong generalization) or domain-specific (weak or no generalization) conclusions depending on the task settings of the consecutive conflicts. Similarly, different brain networks may show domain-generality (i.e., representing multiple conflicts) or domain-specificity (i.e., representing individual conflicts separately)(Jiang & Egner, 2014; Li et al., 2017). Even within the same brain area (e.g., medial frontal cortex), Fu et al. (2022) found that the neural population activity can be factorized into orthogonal dimensions

encoding both domain-general and domain-specific conflict information, which can be selectively read out by downstream brain regions. While the mixture view provides an explanation for the contradictory findings (Braem et al., 2014), it suffers the same criticism as domain-specific cognitive control theories, as it still requires unlimited cognitive control processes to fully cover all possible conflicts.

A key to reconciling domain-general and domain-specific cognitive control is to organize the nearly infinite possible types of conflicts using a system with limited, dissociable dimensions. A construct with a similar function is the *cognitive space* (Bellmund et al., 2018), which extends the idea of cognitive map (Behrens et al., 2018) to the representation of abstract information. Critically, the cognitive space view holds that the representations of different abstract information are organized continuously and the representational geometry in the cognitive space are determined by the similarity among the represented information (Bellmund et al., 2018).

In the human brain, it has been shown that abstract (Behrens et al., 2018; Schuck et al., 2016) and social (Park et al., 2020) information can be represented in a cognitive space. For example, social hierarchies with two independent scores (e.g., popularity and competence) can be represented in a 2D cognitive space (one dimension for each score), such that each social item can be located by its score in the two dimensions (Park et al., 2020). In the field of cognitive control, recent studies have begun to conceptualize different control states within a cognitive space (Badre et al., 2021). For example, Fu et al.(2022) mapped different conflict conditions to locations in a low/high dimensional cognitive space to demonstrate the domain-general/domain-specific problems; Grahek et al.(2022) used a cognitive space model of cognitive control settings to explain behavioral changes in the speed-accuracy tradeoff. However, the cognitive spaces proposed in these studies were only applicable to a limited number of control states involved in their designs. Therefore, it remains unclear whether there is a cognitive space that can explain an unlimited number of control states, similar to that of the spatial location (Bellmund et al., 2018) and non-spatial knowledge (Behrens et al., 2018). A challenge to answering this question lies in how to construct control states with continuous levels of similarity.

Our recent work (Yang et al., 2021) showed that it is possible to manipulate continuous conflict similarity by using a mixture of two independent conflict types with varying ratios, which can be used to further examine the behavioral and neural evidence for the cognitive space view. It is also unclear how the cognitive space of cognitive control is encoded in the brain, although that of spatial locations and non-spatial abstract knowledge has been relatively well investigated in the medial temporal lobe, medial prefrontal and orbitofrontal system (Behrens et al., 2018; Bellmund et al., 2018). Recent research has suggested that the abstract task structure could be encoded and implemented by the frontoparietal network (Vaidya & Badre, 2022; Vaidya et al., 2021), but whether a similar neural system encodes the cognitive space of cognitive control remains untested.

The present study aimed to test the geometry of cognitive space in conflict representation. Specifically, we hypothesize that different types of conflicts are represented as points in a cognitive space. Importantly, the distance between the points, which reflects the geometry of the cognitive space, scales with the difference in the sources of the conflicts being represented by the points. The dimensions in the cognitive space of conflicts can be the aforementioned *domains*, in which domain-specific cognitive control processes are defined. For a specific type of conflict, its location in the cognitive space can be parameterized using a limited number of coordinates, which reflect how much control is needed for each of the domain-specific cognitive control processes. The cognitive space can also represent different types of conflicts with low dimensionality (Badre et al., 2021; MacDowell et al., 2022). Different domains can be represented conjunctively in a single cognitive space to achieve domain-general cognitive control, as conflicts from different sources can be resolved using the same set of cognitive control processes. We further hypothesize

that the cognitive space representing different types of conflicts may be located in the frontoparietal network due to its essential roles in conflict resolution (Freund, Bugg, et al., 2021; Fu et al., 2022) and abstract task representation (Vaidya & Badre, 2022).

In this study, we adjusted the paradigm from our previous study (Yang et al., 2021) by including transitions of trials from five different conflict types, which enabled us to test if these conflict types are organized in a cognitive space (**Fig. 1A**). Specifically, on each trial, an arrow, pointing either upwards or downwards, was presented on one of the 10 possible locations on the screen. Participants were required to respond to the pointing direction of the arrow (up or down) by pressing either the left or right key. Importantly, conflicts from two sources can occur in this task. On one hand, the vertical location of the arrow can be incongruent with the direction (e.g., an up-pointing arrow on the lower half of the screen), resulting spatial Stroop conflict (Liu et al., 2004; Lu & Proctor, 1995). On the other hand, the horizontal location of the arrow can be incongruent with the response key (e.g., an arrow requiring left response presented on the right side of the screen), thus causing Simon conflict (Lu & Proctor, 1995; Simon & Small, 1969). As the arrow location rotates from the horizontal axis to the vertical axis, spatial Stroop conflict increases, and Simon conflict decreases. Therefore, the 10 possible locations of the arrow give rise to five conflict types with unique blend of spatial Stroop and Simon conflicts (Yang et al., 2021). As the increase in spatial Stroop conflict is perfectly correlated with the decrease in Simon conflict, we can use a 1D cognitive space to represent all five conflict types.

One way to parameterize (i.e., defining a coordinate system) the cognitive space is to encode each conflict type by the angle of the axis connecting its two possible stimulus locations (**Fig. 1B**). Within this cognitive space, the similarity between two conflict types can be quantified as the cosine value of their angular difference (**Fig. 1C**). The rationale behind defining conflict similarity based on combinations of different conflict sources, such as spatial-Stroop and Simon, stems from the evidence that these sources undergo independent processing (Egner, 2008; Li et al., 2014; Liu et al., 2010; Wang et al., 2014). Identifying these distinct sources is critical in efficiently resolving potentially infinite conflicts. If the conflict types are organized as a cognitive space in the brain, the similarity between conflict types in the cognitive space should be reflected in both the behavior and similarity in the neural representations of conflict types. Our data from two experiments using this experimental design support both predictions: using behavioral data, we found that the influence of congruency (i.e., whether the task-relevant and task-irrelevant information indicate the same response) from the previous trial to the next trial increases with the conflict similarity between the two trials. Using fMRI data, we found that more similar conflicts showed higher multivariate pattern similarity in the right dorsolateral prefrontal cortex (dlPFC).

Results

Conflict type similarity modulated behavioral congruency sequence effect (CSE)

Experiment 1

We conducted a behavioral experiment (n = 33, 18 females) to examine how CSEs across different conflict types are influenced by their similarity. First, we validated the experimental design by testing the congruency effects. All five conflict types showed robust congruency effects such that the incongruent trials were slower and less accurate than the congruent trials (Note S1; Fig. S1 A/B). To test the influence of similarity between conflict types on behavior, we examined the CSE in consecutive trials. Specifically, the CSE was quantified as the interaction between previous and current trial congruency and can reflect how (in)congruency on the previous trial influences cognitive control on the current trial (Egner, 2007; Schmidt & Weissman, 2014). It has been

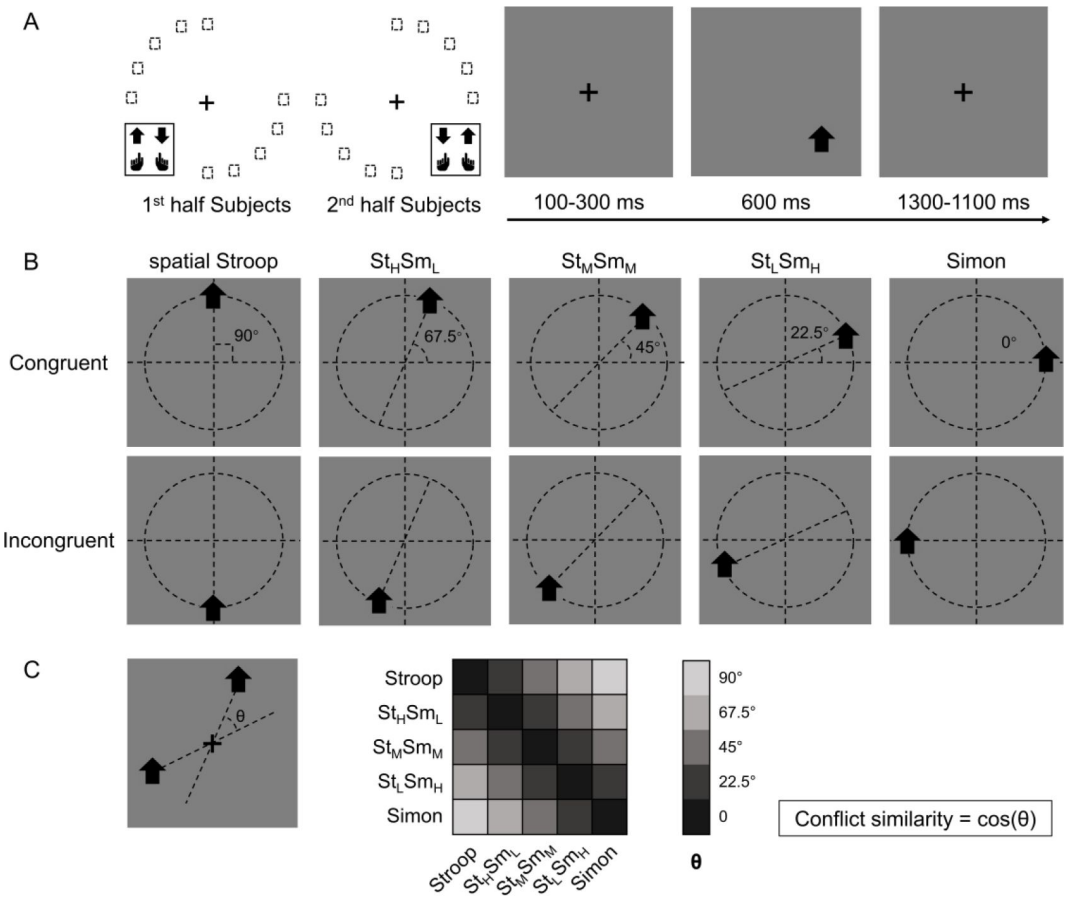


Fig. 1.

Experimental design.

(A) The left panel shows the orthogonal stimulus-response mappings of the two participant groups. In each group the stimuli were only displayed at two quadrants of the circular locations. One group were asked to respond with the left button to the upward arrow and with the right button to the downward arrow presented in the to-left and bottom-right quadrants, and the other group vice versa. The right panel shows the time course of one example trial. The stimuli were displayed for 600 ms, preceded and followed by fixation crosses that lasted for 1400 ms in total. (B) Examples of the five types of conflicts, each containing congruent and incongruent conditions. The arrows were presented at locations along five orientations with isometric polar angles, in which the vertical location introduces the spatial Stroop conflict, and the horizontal location introduces the Simon conflict. Dashed lines are shown only to indicate the location of arrows and were not shown in the experiments. (C) The definition of the angular difference between two conflict types and the conflict similarity. The angle θ is determined by the acute angle between two lines that cross the stimuli and the central fixation. Therefore, stimuli of the same conflict type form the smallest angle of 0, and stimuli between Stroop and Simon form the largest angle of 90°, and others are in between. Conflict similarity is defined by the cosine value of θ. H = high; L = low; M = medium.

shown that the CSE diminishes if the two consecutive trials have different conflict types (Akcay & Hazeltine, 2011; Egner et al., 2007; Torres-Quesada et al., 2013). Similarly, we tested whether the size of CSE increases as a function of conflict similarity between consecutive trials. To this end, we organized trials based on a 5 (previous trial conflict type) × 5 (current trial conflict type) × 2 (previous trial congruency) × 2 (current trial congruency) factorial desi gn, with the first two and the last two factors capturing between-trial conflict similarity and the CSE, respectively. The cells in the 5 × 5 matrix were mapped to different similarity levels according to the angular difference between the two conflict types (**Fig. 1C**). As shown in **Fig. 2**, the CSE, measured in both reaction time (RT) and error rate (ER), scaled with conflict similarity.

To test the modulation of conflict similarity on the CSE, we constructed a linear mixed effect model to predict RT/ER in each cell of the factorial design using a predictor encoding the interaction between the CSE and conflict similarity (see Methods). The results showed a significant effect of conflict similarity (RT: $\beta = 0.10 \pm 0.01$, $t(1978) = 15.82$, $p < .001$, $\eta^2 = .120$; ER: $\beta = 0.15 \pm 0.02$, $t(1978) = 7.84$, $p < .001$, $\eta^2 = .085$, Fig. S2B/E). In other words, the CSE increased with the conflict similarity between two consecutive trials. The conflict similarity modulation effect remained significant after regressing out the influence of physical proximity between the stimuli of consecutive trials (Note S2). As a control analysis, we also compared this approach to a two-stage analysis that first calculated the CSE for each previous × current trial conflict type condition and then tested the modulation of conflict similarity on the CSEs (Yang et al., 2021). The two-stage analysis also showed a strong effect of conflict similarity (RT: $\beta = 0.58 \pm 0.04$, $t(493) = 14.74$, $p < .001$, $\eta^2 = .388$; ER: $\beta = 0.36 \pm 0.05$, $t(493) = 7.01$, $p < .001$, $\eta_p^2 = .320$, Fig. S2A/D). Importantly, individual modulation effects of conflict similarity were positively correlated between the two approaches (RT: $r = 0.48$; ER: $r = 0.86$, both ps < 0.003, one-tailed), indicating the consistency of the estimated conflict similarity effects across the two approaches. In the following texts, we will use the terms "*conflict similarity effect*" and "*conflict type effect*" interchangeably.

Moreover, to test the continuity and generalizability of the similarity modulation, we conducted a leave-one-out prediction analysis. We used the behavioral data from Experiment 1 for this test, due to its larger amount of data than Experiment 2. Specifically, we removed data from one of the five similarity levels (as illustrated by the θs in **Fig. 1C**) and used the remaining data to perform the same mixed-effect model (i.e., the two-stage analysis). This yielded one pair of beta coefficients including the similarity regressor and the intercept for each subject, with which we predicted the CSE for the removed similarity level for each subject. We repeated this process for each similarity level once. The predicted results were highly correlated with the original data, with $r = .87$ for the RT and $r = .84$ for the ER, ps < .001.

Experiment 2

Behavioral results. We next conducted an fMRI experiment using a shorter version of the same task with a different sample (n = 35, 17 females) to seek neural evidence of how different conflict types are organized. Using behavioral data, we first validated the experimental design by testing congruency effects in each of the five conflict types (Note S1; Fig. S1 C/D). We then tested the modulation of conflict similarity on the behavioral CSE using the linear mixed effect model as in Experiment 1 (except the two-stage method). Results showed a significant effect of conflict similarity modulation (RT: $\beta = 0.24 \pm 0.04$, $t(1148) = 6.36$, $p < .001$, $\eta_p^2 = .096$; ER: $\beta = 0.33 \pm 0.06$, $t(1206) = 5.81$, $p < .001$, $\eta_p^2 = .124$, Fig. S2C/F), thus replicating the results of Experimental 1 and setting the stage for fMRI analysis. As in Experiment 1, the conflict similarity modulation effect remained significant after regressing out the influence of physical proximity between the stimuli of consecutive trials (Note S2).

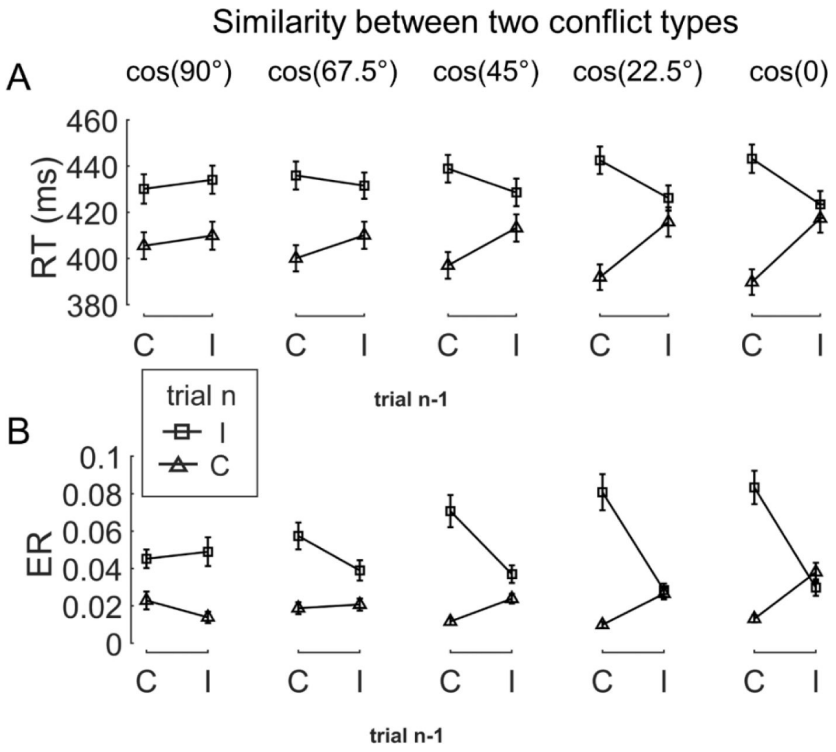


Fig. 2.

The conflict similarity modulation on the behavioral CSE in Experiment 1.

(A) RT and (B) ER are plotted as a function of congruency types on trial n−1 and trial n. Each column shows one similarity level, as indicated by the defined angular difference between two conflict types. Error bars are standard errors. C = congruent; I = incongruent; RT = reaction time; ER = error rate.

Brain activations modulated by conflict type with univariate analyses

In the fMRI analysis, we first replicated the classic congruency effect by searching for brain regions showing higher univariate activation in incongruent than congruent conditions (GLM1, see Methods). Consistent with the literature (Botvinick et al., 2004; Fu et al., 2022), this effect was observed in the pre-supplementary motor area (pre-SMA) (**Fig. 3**, Table S1). We then tested the encoding of conflict type as a cognitive space by identifying brain regions with activation levels parametrically covarying with the coordinates (i.e., axial angle relative to the horizontal axis) in the hypothesized cognitive space. As shown in **Fig. 1B**, change in the angle corresponds to change in spatial Stroop and Simon conflicts in opposite directions. Accordingly, we found the right inferior parietal sulcus (IPS) and the right dorsomedial prefrontal cortex (dmPFC) displayed positive correlation between fMRI activation and the Simon conflict (**Fig. 3**, Fig. S3, Table S1).

To further test if the univariate results explain the conflict similarity modulation of the behavioral CSE (slope in Fig. S2C), we conducted brain-behavioral correlation analyses for regions identified above. Regions with higher spatial Stroop/Simon modulation effects were expected to trigger higher behavioral conflict similarity modulation effect on the CSE. However, none of the three regions (i.e., left MFG, right IPS and right dmPFC, **Fig. 3**) were positively correlated with the behavioral performance, all uncorrected *ps* >.28, one-tailed. In addition, since the conflict type difference covaries with the orientation of the arrow location at the individual level (e.g., the stimulus in a higher level of Simon conflict is always closer to the horizontal axis, see Fig. S4), the univariate modulation effects may not reflect purely conflict type difference. To further tease these factors apart, we used multivariate analyses.

Multivariate patterns of the right dlPFC encodes the conflict similarity

The hypothesis that the brain encodes conflict types in a cognitive space predicts that similar conflict types will have similar neural representations. To test this prediction, we computed the representational similarity matrix (RSM) that encoded correlations of blood-oxygen-level dependent (BOLD) signal patterns between each pair of conflict type (Stroop, St_HSm_L, St_MSm_M, St_LSm_H, and Simon, with H, M and L indicating high, medium and low, respectively, see also **Fig. 1B**) × congruency (congruent, incongruent) × arrow direction (up, down) × run × subject combinations for each of the 360 cortical regions from the Multi-Modal Parcellation (MMP) cortical atlas (Glasser et al., 2016; Jiang et al., 2020). The RSM was then submitted to a linear mixed-effect model as the dependent variable to test whether the representational similarity in each region was modulated by various experimental variables (e.g., conflict type, spatial orientation, stimulus, response, etc., see Methods). The linear mixed-effect model was used to de-correlate conflict type and spatial orientation leveraging the between-subject manipulation of stimulus locations (**Fig. 7** and Fig. S4).

To validate this method, we applied this analysis to test the effects of response/stimulus features and found that representational similarity of the BOLD signal significantly covaried with whether two response/spatial location/arrow directions were the same most strongly in bilateral motor/visual/somatosensory areas, respectively (Fig. S6). We then identified the cortical regions encoding conflict type as a cognitive space by testing whether their RSMs can be explained by the similarity between conflict types. Specifically, we applied three independent criteria: (1) the cortical regions should exhibit a statistically significant positive conflict similarity effect on the RSM; (2) the conflict similarity effect should be stronger in incongruent than congruent trials to reflect flexible adjustment of cognitive control demand when the conflict is present; and (3) the conflict similarity effect should be positively correlated with the behavioral conflict similarity modulation effect on the CSE (see *Behavioral Results* of Experiment 2). The first criterion revealed several cortical regions encoding the conflict similarity, including the Brodmann 8C area (a

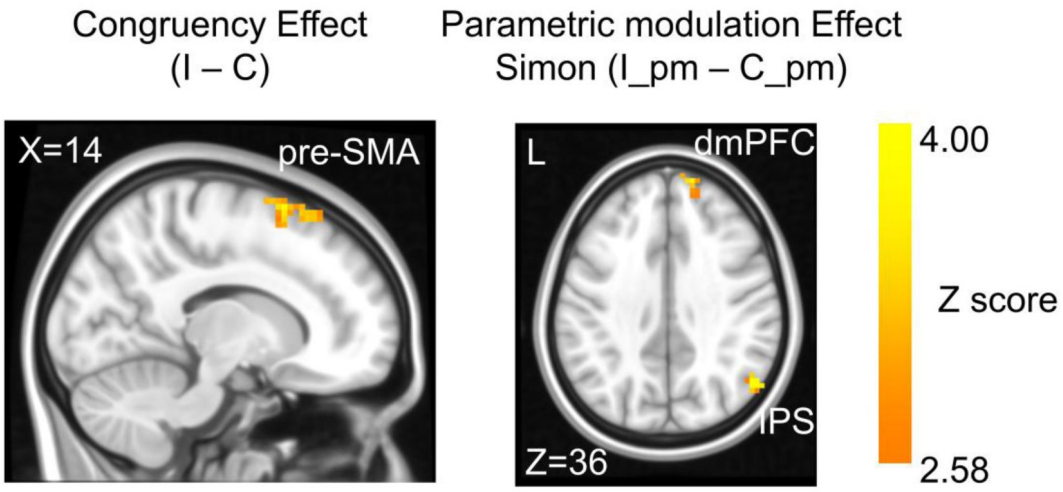


Fig. 3.

The congruency effect and parametric modulation effect detected by uni-voxel analyses.

Results displayed are probabilistic TFCE enhanced and thresholded with voxel-wise $p < .001$ and cluster-wise $p < .05$, both one-tailed. The congruency effect denotes the higher activation in incongruent than congruent condition (left panel). The positive parametric modulation effect (I_pm – C_pm) denotes the higher activation when the conflict type contained a higher ratio of Simon conflict component (right panel). I = incongruent; C = congruent; pm = parametric modulator.

subregion of dlPFC) (Glasser et al., 2016) and a47r in the right hemisphere, and the superior frontal language (SFL) area, 6r, 7Am, 24dd, and ventromedial visual area 1 (VMV1) areas in the left hemisphere (Bonferroni corrected *ps* < 0.0001, one-tailed, **Fig. 4A**). We next tested whether these regions were related to cognitive control by comparing the strength of conflict similarity effect between incongruent and congruent conditions (criterion 2). Results revealed that the left SFL, left VMV1, and right 8C met this criterion, Bonferroni corrected *p*s < .05, one-tailed, suggesting that the representation of conflict type was strengthened when the conflict was present (e.g., **Fig. 4D** and Fig. S5). The inter-subject brain-behavioral correlation analysis (criterion 3) showed that the strength of conflict similarity effect on RSM scaled with the modulation of conflict similarity on the CSE (slope in Fig. S2C) in right 8C (*r* = .52, Bonferroni corrected *p* = .002, one-tailed, **Fig. 4C**) but not in the left SFL and VMV1 (all Bonferroni corrected *ps* > .05, one-tailed). These results are listed in **Table 1**. In addition, we did not find evidence supporting the encoding of congruency in the right 8C area (see Note S6), suggesting that the right 8C area specifically represents conflict similarity.

To examine if the right 8C specifically encodes the cognitive space rather than the domain-general or domain-specific organizations, we tested several additional models (see Methods). Model comparison showed a lower BIC in the Cognitive-Space model (BIC = 5377094) than the Domain-General (BIC = 537127) or Domain-Specific (BIC = 537127) models. Further analysis showed the dimensionality of the representation in the right 8C was 1.19, suggesting the cognitive space was close to 1D. We also tested if the observed conflict similarity effect was driven solely by spatial Stroop or Simon conflicts, and found larger BICs for the models only including the Stroop similarity (i.e., the Stroop-Only model, BIC = 5377122) or Simon similarity (i.e., the Simon-Only model, BIC = 5377096). An additional Stroop+Simon model, including both Stroop-Only and Simon-Only regressors, also showed a worse model fitting (BIC = 5377118). Moreover, we replicated the results with only incongruent trials, considering that the pattern of conflict representations is more manifested when the conflict is present (i.e., on incongruent trials) than not (i.e., on congruent trials). We found a poorer fitting in Domain-general (BIC = 1344129), Domain-Specific (BIC = 1344129), Stroop-Only (BIC = 1344128), Simon-Only (BIC = 1344120), and Stroop+Simon (BIC = 1344157) models than the Cognitive-Space model (BIC = 1344104). These results indicate that the right 8C encodes an integrated cognitive space for resolving Stroop and Simon conflicts. The more detailed model comparison results are listed in **Table 2**.

In sum, we found converging evidence supporting that the right dlPFC (8C area) encoded conflict similarity, which further supports the hypothesis that conflict types are represented in a cognitive space.

Multivariate patterns of visual and oculomotor areas encode stimulus orientation

To tease apart the representation of conflict type from that of perceptual information, we tested the modulation of the spatial orientations of stimulus locations on RSM using the aforementioned RSA. We also applied three independent criteria: (1) the cortical regions should exhibit a statistically significant orientation effect on the RSM; (2) the conflict similarity effect should be stronger in incongruent than congruent trials; and (3) the orientation effect should not interact with the CSE, since the orientation effect was dissociated from the conflict similarity effect and was not expected to influence cognitive control. We observed increasing fMRI representational similarity between trials with more similar orientations of stimulus location in the occipital cortex, such as right V1, right V2, right V4, and right lateral occipital 2 (LO2) areas (Bonferroni corrected *p*s < 0.0001). We also found the same effect in the oculomotor related region, i.e., the left frontal eye field (FEF), and other regions including the right 5m, left 31pv and right parietal area F (PF) (**Fig. 5A**). Then we tested if any of these brain regions were related to the conflict representation by comparing their encoding strength between incongruent and congruent conditions. Results showed that the right V1, right V2, left FEF, and right PF encoded stronger

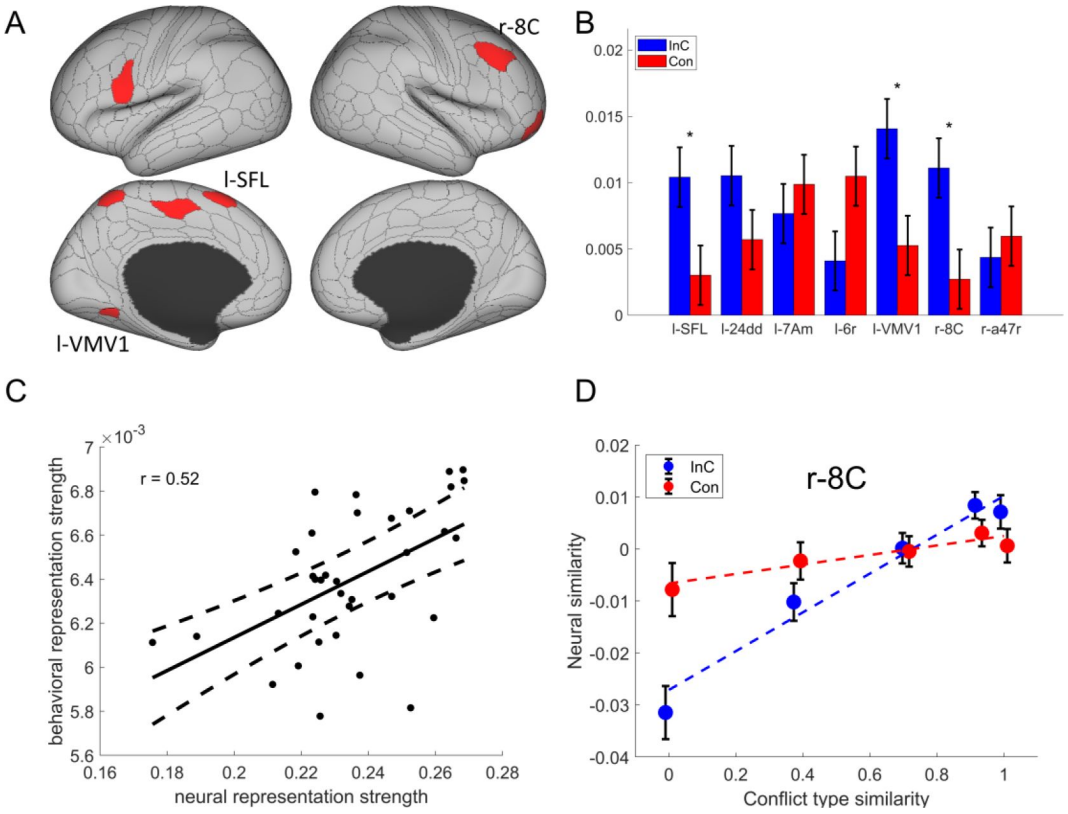


Fig. 4.

The conflict type effect.

(A) Brain regions surviving the Bonferroni correction ($p < 0.0001$) across the 360 regions (criterion 1). Labeled regions are those meeting the criterion 2. (B) Different encoding of conflict type in the incongruent with congruent conditions. * Bonferroni corrected $p < .05$. (C) The brain-behavior correlation of the right 8C (criterion 3). The x-axis shows the beta coefficient of the conflict type effect from the RSA, and the y-axis shows the beta coefficient obtained from the behavioral linear model using the conflict similarity to predict the CSE in Experiment 2. (D) Illustration of the different encoding strength of conflict type similarity in incongruent versus congruent conditions of right 8C. The y-axis is derived from the z-scored Pearson correlation coefficient, consistent with the RSA methodology. See Fig. S5B for a plot with the raw Pearson correlation measurement. l = left; r = right.

Region name	Criterion 1			Criterion 2		Criterion 3	
	t	$\beta \pm SD$	p	t	p	r	p
Conflict type effect							
left SFL	5.12	0.0008 ± 0.0002	1.5×10^{-7}	2.45	.007	.35	.021
left 24dd	6.61	0.0012 ± 0.0002	1.9×10^{-11}	1.61	.053	−.04	.580
left 7Am	7.40	0.0015 ± 0.0002	6.6×10^{-14}	−0.75	.772	.04	.418
left 6r	5.90	0.0012 ± 0.0002	1.8×10^{-9}	−1.94	.974	−.11	.736
left VMV1	7.84	0.0017 ± 0.0002	2.2×10^{-15}	2.83	.002	−.27	.940
right 8C*	5.60	0.0011 ± 0.0002	1.1×10^{-8}	2.46	.007	.52	.001
right a47r	5.04	0.0007 ± 0.0001	2.4×10^{-7}	−0.68	.753	−.04	.589
Orientation effect							
left FEF*	5.17	0.0011 ± 0.0002	1.2×10^{-7}	2.73	.003	.02	.456
left 31pv	6.37	0.0019 ± 0.0003	9.6×10^{-11}	−0.48	.686	.12	.247
right V1*	5.59	0.0007 ± 0.0001	1.1×10^{-8}	2.91	.002	−.08	.671
right V2*	5.09	0.0015 ± 0.0003	1.8×10^{-7}	3.19	.001	.02	.466
right V4	5.02	0.0015 ± 0.0003	2.6×10^{-7}	0.56	.289	−.13	.777
right LO2	5.95	0.0020 ± 0.0003	1.3×10^{-9}	−1.81	.965	.27	.062
right 5m	5.24	0.0011 ± 0.0002	8.1×10^{-8}	−2.12	.983	−.10	.707
right PF*	5.57	0.0009 ± 0.0002	1.2×10^{-8}	3.08	.001	-.01	.523

Note. all p values listed are 1-tailed and uncorrected. The * denotes the regions meeting all three criteria for each effect.

Table 1.

Summary statistics of the cross-subject RSA for regions showing conflict type and orientation effects identified by the three criteria.

Model name	Regressor	Full RSM			RSM_I		
		t	*p*	BIC	*t*	*p*	BIC
Cognitive-Space	similarity	5.60	1.07×10^{-8}	5377094	4.97	3.37×10^{-7}	1344105
Domain-General	similarity	−0.19	.575	5377127	0.40	.346	1344129
Domain-Specific	similarity	0.84	.201	5377127	0.28	.390	1344129
Stroop-Only	similarity	0.20	.423	5377122	1.23	.109	1344128
Simon-Only	similarity	4.19	1.37×10^{-5}	5377096	3.11	.0009	1344120
Stroop+Simon	Stroop	0.06	.478	5377118	1.01	.149	1344157
	Simon	3.30	.0005		3.05	.001	

Table 2.

Model comparison results of the right 8C. RSM_I shows results using incongruent trials only.

orientation effect in the incongruent than the congruent condition, Bonferroni corrected *p*s < .05, one-tailed (Table1, **Fig. 5B**). We then tested if any of these regions was related to the behavioral performance, and results showed that none of them positively correlated with the behavioral conflict similarity modulation effect, all uncorrected *ps* > .45, one-tailed. Thus all regions are consistent with the criterion 3. Like the right 8C area, none of the reported areas directly encoded congruency (see Note S6). Taken together, we found that the visual and oculomotor regions encoded orientations of stimulus location in a continuous manner and that the encoding strength was stronger when the conflict was present.

We hypothesize that the overlapping spatial information of orientation may have facilitated the encoding of conflict types. To explore the relation between conflict type and orientation representations, we conducted representational connectivity (i.e., the similarity between two within-subject RSMs of two regions) (Kriegeskorte et al., 2008) analyses and found that among the orientation effect regions, the right V1 and right V2 showed significant representational connectivity with the right 8C compared to the controlled regions (including those encoding orientation effect but not showing larger encoding strength in incongruent than congruent conditions, as well as eight other regions encoding none of our defined effects in the main RSA, see Methods). Compared with the largest connectivity strength in the controlled regions (i.e., the right V4, $\beta = 0.7963 \pm 0.0314$), we found higher connectivity in the right V1, $\beta = 0.8633 \pm 0.0325$, and right V2, $\beta = 0.8773 \pm 0.0334$, (Fig. S7).

Discussion

Understanding how different types of conflicts are resolved is essential to answer how cognitive control achieves adaptive behavior. However, the dichotomy between domain-general and/or domain-specific processes presents a dilemma (Braem et al., 2014; Egner, 2008). Reconciliation of the two views also suffers from the inability to fully address how infinite conflicts can be resolved by a limited set of cognitive control processes. In this study, we hypothesized that this issue can be addressed if conflicts are organized as a cognitive space. Leveraging the well-known dissociation between the spatial Stroop and Simon conflicts (Li et al., 2014; Liu et al., 2010; Wang et al., 2014), we designed five conflict types that are systematically different from each other. The cognitive space hypothesis predicted that the representational proximity/distance between two conflict types scales with their similarities/dissimilarities, which was tested at both behavioral and neural levels. Behaviorally, we found that the CSEs were linearly modulated by conflict similarity between consecutive trials, replicating and extending our previous study (Yang et al., 2021). BOLD activity patterns in the right dlPFC further showed that the representational similarity between conflict types was modulated by their conflict similarity, and that strength of the modulation was positively associated with the modulation of conflict similarity on the behavioral CSE. We also observed that activity in two brain regions (right IPS and right dlPFC) was parametrically modulated by the conflict type difference, though they did not directly explain the behavioral results. Additionally, we found that the visual regions encoded the spatial orientation of the stimulus location, which might provide the essential concrete information to determine the conflict type. Together, these results support the hypothesis that conflicts are organized in a cognitive space that enables a limited set of cognitive control processes to resolve infinite possible types of conflicts.

Conventionally, the domain-general view of control suggests a common representation for different types of conflicts (**Fig. 6**, left), while the domain-specific view suggests dissociated representations for different types (**Fig. 6**, right). Previous research on this topic often adopts a binary manipulation of conflicts (Braem et al., 2014) (i.e., each domain only has one conflict type) and gathered evidence for the domain-general/specific view with presence/absence of CSE, respectively. Here, we parametrically manipulated the similarity of conflict types and found the CSE systematically vary with conflict similarity level, demonstrating that cognitive control is

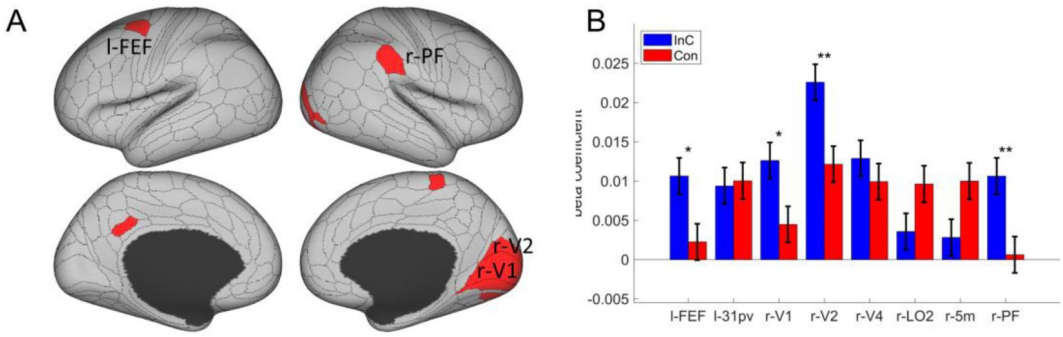


Fig. 5.

The orientation effect.

(A) Brain regions surviving the Bonferroni correction ($p < 0.0001$) across the 360 regions (criterion 1). Labeled regions are those meeting the criterion 2. (B) Different encoding of orientation in the incongruent with congruent conditions. * Bonferroni corrected $p < .05$., ** Bonferroni corrected $p < .01$.

neither purely domain-general nor purely domain-specific, but can be reconciled as a cognitive space (Bellmund et al., 2018)(**Fig. 6**, middle). The model comparison analysis also showed that the Cognitive-Space model best explained the representation in right DLPFC. Specifically, the cognitive space provides a solution to use a single cognitive space organization to encode different types of conflicts that are close (domain-general) or distant (domain-specific) to each other. It also shows the potential for how unlimited conflict types can be coded using limited resources (i.e., as different points in a low-dimensional cognitive space), as suggested by its out-of-sample predictability. Moreover, geometry can also emerge in the cognitive space (Fu et al., 2022), which will allow for decomposition of a conflict type (e.g., how much conflict in each of the dimensions in the cognitive space) so that it can be mapped into the limited set of cognitive control processes. Such geometry enables fast learning of cognitive control settings from similar conflict types by providing a measure of similarity (e.g., as distance in space).

If the dimensionality of the cognitive space of conflict is extremely high, the cognitive space solution would suffer the same criticism as the domain-specificity theory. We argue that the dimensionality is manageable for the human brain, as task information unrelated to differentiating conflicts can be removed. For example, the Simon conflict can be represented in a space consisting of spatial location, stimulus information and responses. Thus, the dimensionality of the cognitive space of conflicts should not exceed the number of represented features. The dimensionality can be further reduced, as humans selectively represent a small number of features when learning task representations (e.g., spatial information is reduced to the horizontal dimension from the 3D space we live in)(Niv, 2019). Consistently, we observed a low dimensional (1.19D) space representing the five conflict types. This is expected since the only manipulated variable is the angular distance between conflict types. The reduced dimensionality does not only require less effort to represent the conflict, but also facilitates generalization of cognitive control settings among different conflict types (Badre et al., 2021).

Although our finding of cognitive space in the right dlPFC differs from other cognitive space studies (Constantinescu et al., 2016; Park et al., 2020; Schuck et al., 2016) that highlighted the orbitofrontal and hippocampus regions, it is consistent with the cognitive control literature. The prefrontal cortex has long been believed to be a key region of cognitive control representation (Mansouri et al., 2007; Miller & Cohen, 2001; Milner, 1963) and is widely engaged in multiple task demands (Cole et al., 2013; Duncan, 2010). However, it is not until recently that the multivariate representation in this region has been examined. For instance, Vaidya et al. (2021) reported that frontal regions presented latent states that are organized hierarchically. Freund et al.(2021) showed that dlPFC encoded the target and congruency in a typical color-word Stroop task. Taken together, we suggest that the right dlPFC might flexibly encode a variety of cognitive spaces to meet the dynamic task demands. In addition, we found no such representation in the left dlPFC (Note S8), indicating a possible lateralization. Previous studies showed that the left dlPFC was related to the expectancy-related attentional set up-regulation, while the right dlPFC was related to the online adjustment of control (Friehs et al., 2020; Vanderhasselt et al., 2009), which is consistent with our findings. Moreover, the right PFC also represents a composition of single rules (Reverberi et al., 2012), which may explain how the spatial Stroop and Simon types can be jointly encoded in a single space.

Previous researchers have proposed an “expected value of control (EVC)” theory, which posits that the brain can evaluate the cost and benefit associated with executing control for a demanding task, such as the conflict task, and specify the optimal control strength (Shenhav et al., 2013). For instance, Grahek et al. (2022) found that more frequently switching goals when doing a Stroop task were achieved by adjusting smaller control intensity. Our work complements the EVC theory by further investigating the neural representation of different conflict conditions and how these representations can be evaluated to facilitate conflict resolution. We found that different conflict conditions can be efficiently represented in a cognitive space encoded by the right dlPFC, and participants with stronger cognitive space representation have also adjusted their conflict control

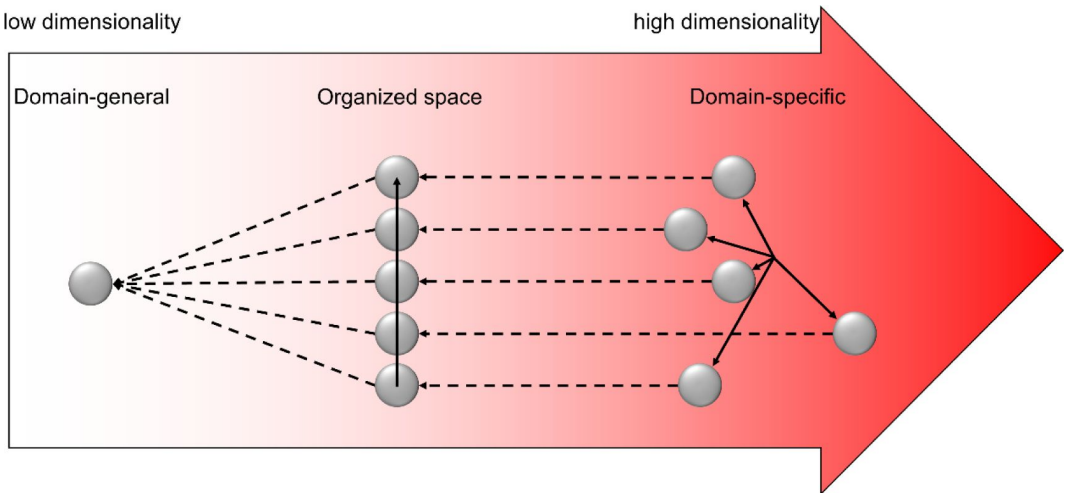


Fig. 6.

Illustration of the hypothesized dimensionalities of different representations.

The shade of the red color indicates the degree of dimensionality (i.e., how many dimensions are needed to represent different states). The dimensionality of domain-general representation is extremely low, with all representations compressed to one dot. The dimensionality of domain-specific representation is extremely high, with each control state encoded in a unique and orthogonal dimension. The dimensionality of the organized representation is modest, enabling distant states to be separated but also allowing close states to share representations. The solid arrows show the axes of different dimensions. The dashed arrows indicate how the representational dimensionality can be reduced by projecting the independent dimensions to a common dimension.

to a greater extent based on the conflict similarity (**Fig 4C**). The finding suggests that the cognitive space organization of conflicts guides cognitive control to adjust behavior. Previous studies have shown that participants may adopt different strategies to represent a task, with the model-based strategies benefitting goal-related behaviors more than the model-free strategies (Rmus et al., 2022). Similarly, we propose that cognitive space could serve as a mental model to assist fast learning and efficient organization of cognitive control settings. Specifically, the cognitive space representation may provide a principle for how our brain evaluates the expected cost of switching and the benefit of generalization between states and selects the path with the best cost-benefit tradeoff (Abrahamse et al., 2016; Shenhav et al., 2013). The proximity between two states in cognitive space could reflect both the expected cognitive demand required to transition and the useful mechanisms to adapt from. The closer the two conditions are in cognitive space, the lower the expected switching cost and the higher the generalizability when transitioning between them. With the organization of a cognitive space, a new conflict can be quickly assigned a location in the cognitive space, which will facilitate the development of cognitive control settings for this conflict by interpolating nearby conflicts and/or projecting the location to axes representing different cognitive control processes, thus leading to a stronger CSE when following a more similar conflict condition. On the other hand, without a cognitive space, there would be no measure of similarity between conflicts on different trials, hence limiting the ability of fast learning of cognitive control setting from similar trials.

The cognitive space in the right dlPFC appears to be an abstraction of concrete information from the visual regions. We found that the right V1 and V2 encoded the spatial orientation of the target location (**Fig. 5**) and showed strong representational connectivity with the right dlPFC (Fig. S7), suggesting that there might be information exchange between these regions. We speculate that the representation of spatial orientation may have provided the essential perceptual information to determine the conflict type (**Fig. 1**) and thus served as the critical input for the cognitive space. The conflict type representation further incorporates the stimulus-response mapping rules to the spatial orientation representation, so that vertically symmetric orientations can be recognized as the same conflict type (Fig. S4). In other words, the representation of conflict type involves the compression of perceptual information (Flesch et al., 2022), which is consistent with the idea of a low-dimensional representation of cognitive control (Badre et al., 2021; MacDowell et al., 2022). The compression and abstraction processes might be why the frontoparietal regions are the top of hierarchy of information processing (Gilbert & Li, 2013) and why the frontoparietal regions are widely engaged in multiple task demands(Duncan, 2013).

Although the spatial orientation information in our design could be helpful to the construction of cognitive space, the cognitive space itself was independent of the stimulus-level representation of the task. We found the conflict similarity modulation on CSE did not change with more training (see Note S3), indicating that the cognitive space did not depend on strategies that could be learned through training. Instead, the cognitive space should be determined by the intrinsic similarity structure of the task design. For example, a previous study (Freitas & Clark, 2015) has found that the CSE across different versions of spatial Stroop and flanker tasks was stronger than that across either of the two conflicts and Simon. In their designs, the stimulus similarity was controlled at the same level, so the difference in CSE was only attributable to the similar dimensional overlap between Stroop and flanker tasks, in contrast to the Simon task. Furthermore, recent studies showed that the cognitive space generally exists to represent structured latent states (e.g., Vaidya & Badre, 2022), mental strategy cost (Grahek et al., 2022), and social hierarchies (Park et al., 2020). Therefore, cognitive space is likely a universal strategy that can be applied to different scenarios.

With conventional univariate analyses, we observed that the overall congruency effect was located at the medial frontal region (i.e., pre-SMA), which is consistent with previous studies (Botvinick et al., 2004; Fu et al., 2022). Beyond that, we also found regions that can be parametrically modulated by conflict type difference, including right IPS, right dlPFC (modulated

by Simon difference). The right lateralization of these regions is consistent with a previous finding (Li et al., 2017). The scaling of brain activities based on conflict difference is potentially important to the representational organization of different types of conflicts. However, we didn't observe their brain-behavioral relevance. One possible reason is that the conflict (dis)similarity is a combination of (dis)similarity of spatial Stroop and Simon conflicts, but each univariate region only reflects difference along a single conflict domain. Also likely, the representational geometry is more of a multivariate problem than what univariate activities can capture (Freund, Etzel, et al., 2021). Future studies may adopt approaches such as repetition suppression induced fMRI adaptation (Badre et al., 2021) to test the role of univariate activities in task representations.

Methodological implications. The use of an experimental paradigm that permits parametric manipulation of conflict similarity provides a way to systematically investigate the organization of cognitive control, as well as its influence on adaptive behaviors. This approach extends traditional paradigms, such as the multi-source interference task (Fu et al., 2022), color Stroop-Simon task (Liu et al., 2010) and similar paradigms that do not afford a quantifiable metric of conflict source similarity. Moreover, the cross-subject RSA provides high sensitivity to the variables of interest and the ability to separate confounding factors. For instance, in addition to dissociating conflict type from orientation, we dissociated target from response, and spatial Stroop distractor from Simon distractor. We further showed cognitive control can both enhance the target representation and suppress the distractor representation (Note S10, Fig. S8), which is in line with previous studies (Polk et al., 2008; Ritz & Shenhav, 2022).

A few limitations of this study need to be noted. To parametrically manipulate the conflict similarity levels, we adopted the spatial Stroop-Simon paradigm that enables parametrical combinations of spatial Stroop and Simon conflicts. However, since this paradigm is a two-alternative forced choice design, the behavioral CSE is not a pure measure of adjusted control but could be partly confounded by bottom-up factors such as feature integration (Hommel et al., 2004). Future studies may replicate our findings with a multiple-choice design (including more varied stimulus sets, locations and responses) with confound-free trial sequences (Braem et al., 2019). Another limitation is that in our design, the spatial Stroop and Simon effects are highly anticorrelated. This constraint may make the five conflict types represented in a unidimensional space (e.g., a circle) embedded in a 2D space. Future studies may test the 2D cognitive space with fully independent conditions. A possible improvement to our current design would be to include left, right, up, and down arrows represented in a grid formation across four spatially separate quadrants, with each arrow mapped to its own response button. However, one potential confounding factor would be that these conditions have different levels of difficulty (i.e., different magnitude of conflict), which may affect the CSE results and their representational similarity. Additionally, our study is not a comprehensive test of the cognitive space hypothesis but aimed primarily to provide original evidence for the geometry of cognitive space in representing conflict information in cognitive control. Future research should examine other aspects of the cognitive space such as its dimensionality, its applicability to other conflict tasks such as Eriksen Flanker task, and its relevance to other cognitive abilities, such as cognitive flexibility and learning.

In sum, we showed that the cognitive control can be organized in an abstract cognitive space that is represented in the right dlPFC and guides cognitive control to adjust goal-directed behavior. The cognitive space hypothesis reconciles the long-standing debate between the domain-general and domain-specific views of cognitive control and provides a parsimonious and more broadly applicable framework for understanding how our brains efficiently and flexibly represents multiple task settings.

Materials and Methods

Subjects

In Experiment 1, we enrolled thirty-three college students (19-28 years old, average of 21.5 ± 2.3 years old; 19 males). In Experiment 2, thirty-six college students were recruited, and one subject was excluded due to not following task instructions. The final sample of Experiment 2 consisted of thirty-five participants (19-29 years old, average of 22.3 ± 2. 5 years old; 17 males). The sample sizes were determined based on our previous study (Yang et al., 2021). All participants reported no history of psychiatric or neurological disorders and were right-handed, with normal or corrected-to-normal vision. The experiments were approved by the Institutional Review Board of the Institute of Psychology, Chinese Academy of Science. Informed consent was obtained from all subjects.

Method Details

Experiment 1

Experimental Design. We adopted a modified spatial Stroop-Simon task (Yang et al., 2021) (**Fig. 1**). The task was programmed with the E-prime 2.0 (Psychological Software Tools, Inc.). The stimulus was an upward or downward black arrow (visual angle of ~ 1°) displayed on a 17-inch LCD monitor with a viewing distance of ~60 cm. The arrow appeared inside a grey square at one of ten locations with the same distance from the center of the screen, including two horizontal (left and right), two vertical (top and bottom), and six corner (orientations of 22.5°, 45°and 67.5°) locations. The distance from the arrow to the screen center was approximately 3°. To dissociate orientation of stimulus locations and conflict types (see below), participants were randomly assigned to two sets of stimulus locations (one included top-right and bottom-left quadrants, and the other included top-left and bottom-right quadrants).

Each trial started with a fixation cross displayed in the center for 100–300 ms, followed by the arrow for 600 ms and another fixation cross for 1100–1300 ms (the total trial length was fixed at 2000 ms). Participants were instructed to respond to the pointing direction of the arrow by pressing a left or right button and to ignore its location. The mapping between the arrow orientations and the response buttons was counterbalanced across participants. The task design introduced two possible sources of conflicts: on one hand, the direction of the arrow is either congruent or incongruent with the vertical location of the arrow, thus introducing a spatial Stroop conflict (Lu & Proctor, 1995; MacLeod, 1991), which contains the dimensional overlap between task-relevant stimulus and task-irrelevant stimulus (Kornblum et al., 1990); on the other hand, the response (left or right button) is either congruent or incongruent with the horizontal location of the arrow, thus introducing a Simon conflict (Lu & Proctor, 1995; Simon & Small, 1969), which contains the dimensional overlap between task-irrelevant stimulus and response (Kornblum et al., 1990). Therefore, the five polar orientations of the stimulus location (from 0 to 90°) defined five unique combinations of spatial Stroop and Simon conflicts, with more similar orientations having more similar composition of conflicts. More generally, the spatial orientation of the arrow location relative to the center of the screen forms a cognitive space of different blending of spatial Stroop and Simon conflicts.

The formal task consisted of 30 runs of 101 trials each, divided into three sessions of ten runs each. The participants completed one session each time and all three sessions within one week. Before each session, the participants performed training blocks of 20 trials repeatedly until the accuracy reached 90% in the most recent block. The trial sequences of the formal task were pseudo-randomly generated to ensure that each of the task conditions and their transitions occurred with equal number of trials.

Experiment 2

Experimental Design. The apparatus, stimuli and procedure were identical to Experiment 1 except for the changes below. The stimuli were back projected onto a screen (with viewing angle being ~3.9° between the arrow and the center of the screen) behind the subject and viewed via a surface mirror mounted onto the head coil. Due to the time constraints of fMRI scanning, the trial numbers decreased to a total of 340, divided into two runs with 170 trials each. To obtain a better hemodynamic model fitting, we generated two pseudo-random sequences optimized with a genetic algorithm (Wager & Nichols, 2003) conducted by the NeuroDesign package (Durnez et al., 2018) (see Note S3 for more detail). In addition, we added 6 seconds of fixation before each run to allow the stabilization of the hemodynamic signal, and 20 seconds after each run to allow the signal to drop to the baseline.

Before scanning, participants performed two practice sessions. The first one contained 10 trials of center-displayed arrow and the second one contained 32 trials using the same design as the main task. They repeated both sessions until their performance accuracy for each session reached 90%, after which the scanning began.

fMRI Image acquisition and preprocessing

Functional imaging was performed on a 3T GE scanner (Discovery MR750) using echo-planar imaging (EPI) sensitive to BOLD contrast [in-plane resolution of 3.5 × 3.5 mm^2, 64 × 64 matrix, 37 slices with a thickness of 3.5 mm and no interslice skip, repetition time (TR) of 2000 ms, echo-time (TE) of 30 ms, and a flip angle of 90°]. In addition, a sagittal T1-weighted anatomical image was acquired as a structural reference scan, with a total of 256 slices at a thickness of 1.0 mm with no gap and an in-plane resolution of 1.0 × 1. 0 mm^2.

Before preprocessing, the first three volumes of the functional images were removed due to the instability of the signal at the beginning of the scan. The anatomical and functional data were preprocessed with the fMRIprep 20.2.0 (Esteban et al., 2019) (RRID:SCR_016216), which is based on Nipype 1.5.1 (Gorgolewski et al., 2011) (RRID:SCR_002502). Specifically, BOLD runs were slice-time corrected using 3dTshift from AFNI 20160207 (Jenkinson et al., 2002) (RRID:SCR_005927). The BOLD time-series were resampled to the MNI152NLin2009cAsym space without smoothing. For a more detailed description of preprocessing, see Note S4. After preprocessing, we resampled the functional data to a spatial resolution of 3 × 3 × 3 mm^3. All analyses were conducted in volumetric space, and surface maps are produced with Connectome Workbench (*https://www.humanconnectome.org/software/connectome-workbench*) for display purpose only.

Quantification and Statistical Analysis

Behavioral analysis

Experiment 1. RT and ER were the two dependent variables analyzed. As for RTs, we excluded the first trial of each block (0.9%, for CSE analysis only), error trials (3.8%), trials with RTs beyond three *SD*s or shorter than 200 ms (1.3%) and post-error trials (3.4%). For the ER analysis, the first trial of each block and trials after an error were excluded. To exclude the possible influence of response repetition, we centered the RT and ER data within the response repetition and response alternation conditions separately by replacing condition-specific mean with the global mean for each subject.

To examine the modulation of conflict similarity on the CSE, we organized trials based on a 5 (previous trial conflict type) × 5 (current trial conflict type) × 2 (previous trial congruency) × 2 (current trial congruency) factorial design. As conflict similarity is commutive between conflict

types, we expected the previous by current trial conflict type factorial design to be a symmetrical (e.g., a conflict 1-conflict 2 sequence in theory has the same conflict similarity modulation effect as a conflict 2-conflict 1 sequence), resulting a total of 15 conditions left for the first two factors of the design (i.e., previous × current trial conflict type). For each previous × current trial conflict type condition, the conflict similarity between the two trials can be quantified as the cosine of their angular difference. In the current design, there were five possible angular difference levels (0, 22.5°, 42.5°, 67.5° and 90°, see **Fig. 1C**). We further coded the previous by current trial congruency conditions (hereafter abbreviated as CSE conditions) as CC, CI, IC and II, with the first and second letter encoding the congruency (C) or incongruency (I) on the previous and current trial, respectively. As the CSE is operationalized as the interaction between previous and current trial congruency, it can be rewritten as a contrast of (CI – CC) – (II – IC). In other words, the load of CSE on CI, CC, II and IC conditions is 1, –1, –1 and 1, respectively. To estimate the modulation of conflict similarity on the CSE, we built a regressor by calculating the Kronecker product of the conflict similarity scores of the 15 previous × current trial conflict similarity conditions and the CSE loadings of previous × current trial congruency conditions. This regressor was regressed against RT and ER data separately, which were normalized across participants and CSE conditions. The regression was performed using a linear mixed-effect model, with the intercept and the slope of the regressor for the modulation of conflict similarity on the CSE as random effects (across both participants and the four CSE conditions). As a control analysis, we built a similar two-stage model (Yang et al., 2021). In the first stage, the CSE [i.e., (CI – CC) – (II – IC)] for each of the previous × current trial conflict similarity condition was computed. In the second stage, CSE was used as the dependent variable and was predicted using conflict similarity across the 15 previous × current trial conflict type conditions. The regression was also performed using a linear mixed effect model with the intercept and the slope of the regressor for the modulation of conflict similarity on the CSE as random effects (across participants).

Experiment 2. Behavioral data was analyzed using the same linear mixed effect model as Experiment 1, with all the CC, CI, IC and II trials as the dependent variable. In addition, to test if fMRI activity patterns may explain the behavioral representations differently in congruent and incongruent conditions, we conducted the same analysis to measure behavioral modulation of conflict similarity on the CSE using congruent (CC and IC) and incongruent (CI and II) trials separately.

Estimation of fMRI activity with univariate general linear model (GLM)

To estimate voxel-wise fMRI activity for each of the experimental conditions, the preprocessed fMRI data of each run were analyzed with the GLM. We conducted three GLMs for different purposes. GLM1 aimed to validate the design of our study by replicating the engagement of frontoparietal activities in conflict processing documented in previous studies (Jiang & Egner, 2014; Li et al., 2017), and to explore the cognitive space related regions that were parametrically modulated by the conflict type. Preprocessed functional images were smoothed using a 6-mm FWHM Gaussian kernel. We included incongruent and congruent conditions as main regressors and appended a parametric modulator for each condition. The modulation parameters for Stroop, St_HSm_L, St_MSm_M, St_LSm_H, and Simon trials were −2, −1, 0, 1 and 2, respectively. In addition, we also added event-related nuisance regressors, including error/missed trials, outlier trials (slower than three SDs of the mean or faster than 200 ms) and trials within two TRs of significant head motion (i.e., outlier TRs, defined as standard DVARS > 1.5 or FD > 0.9 mm from previous TR)(Jiang et al., 2020). On average there were 1.2 outlier TRs for each run. These regressors were convolved with a canonical hemodynamic response function (HRF) in SPM 12 (*http://www.fil.ion.ucl.ac.uk/spm*). We further added volume-level nuisance regressors, including the six head motion parameters, the global signal, the white matter signal, the cerebrospinal fluid signal, and outlier TRs. Low-frequency signal drifts were filtered using a cutoff period of 128 s. The two runs were regarded as different sessions and incorporated into a single GLM to get more

power. This yielded two beta maps (i.e., a main effect map and a parametric modulation map) for the incongruent and congruent conditions, respectively and for each subject. At the group level, paired t-tests were conducted between incongruent and congruent conditions, one for the main effect and the other for the parametric modulation effect. Since the spatial Stroop and Simon conflicts change in the opposite direction to each other, a positive modulation effect would reflect a higher brain activation when there is more Simon conflict, and a negative modulation effect would reflect a higher brain activation for more spatial Stroop conflict. To avoid confusion, we converted the modulation effect of spatial Stroop to positive by using a contrast of [– (I_pm – C_pm)] throughout the results presentation. Results were corrected with the probabilistic threshold-free cluster enhancement (pTFCE) and then thresholded by 3dClustSim function in AFNI (Cox & Hyde, 1997) with voxel-wise $p < .001$ and cluster-wize $p < .05$, both 1-tailed. To visualize the parametric modulation effects, we conducted a similar GLM (GLM2), except we used incongruent and congruent conditions from each conflict type as separate regressors with no parametric modulation. Then we extracted beta coefficients for each regressor and each participant with regions observed in GLM1 as regions of interest, and finally got the incongruent−congruent contrasts for each conflict type at the individual level. We reported the results in **Fig. 3**, Table S1, and Fig. S3. Visualization of the uni-voxel results was made by the MRIcron (*https://www.mccauslandcenter.sc.edu/mricro/mricron/*).

The GLM3 aimed to prepare for the representational similarity analysis (see below). There were several differences compared to GLM1. The unsmoothed functional images after preprocessing were used. This model included 20 event-related regressors, one for each of the 5 (conflict type) × 2 (congruency condition) × 2 (arrow direction) conditions. The event-related nuisance regressors were similar to GLM1, but with additional regressors of response repetition and post-error trials to account for the nuisance inter-trial effects. To fully expand the variance, we conducted one GLM analysis for each run. After this procedure, a voxel-wise fMRI activation map was obtained per condition, run and subject.

Representational similarity analysis (RSA)

To measure the neural representation of conflict similarity, we adopted the RSA. RSAs were conducted on each of the 360 cortical regions of a volumetric version of the MMP cortical atlas (Glasser et al., 2016). To de-correlate the factors of conflict type and orientation of stimulus location, we leveraged the between-subject manipulation of stimulus locations and conducted RSA in a cross-subject fashion (**Fig. 7** and Fig. S4). Previous studies (e.g., J. Chen et al., 2017) have demonstrated that consistent multi-voxel activation patterns exist across individuals, and successful applications of cross-subject RSA (see review by Freund, Etzel, et al., 2021) and cross-subject decoding approaches (Jiang et al., 2016; Tusche et al., 2016) have also been reported. The beta estimates from GLM3 were noise-normalized by dividing the original beta coefficients by the square root of the covariance matrix of the error terms (Nili et al., 2014). For each cortical region, we calculated the Pearson's correlations between fMRI activity patterns for each run and each subject, yielding a 1400 (20 conditions × 2 runs × 35 participants) × 1400 RSM. The correlations were calculated in a cross-voxel manner using the fMRI activation maps obtained from GLM3 described in the previous section. We excluded within-subject cells from the RSM (thus also excluding the within-run similarity as suggested by Walther et al., (2016)), and the remaining cells were converted into a vector, which was then z-transformed and submitted to a linear mixed effect model as the dependent variable. The linear mixed effect model also included regressors of conflict similarity and orientation similarity. Importantly, conflict similarity was based on how Simon and spatial Stroop conflicts are combined and hence was calculated by first rotating all subject's stimulus location to the top-right and bottom-left quadrants, whereas orientation was calculated using original stimulus locations. As a result, the regressors representing conflict similarity and orientation similarity were de-correlated. Similarity between two conditions was measured as the cosine value of the angular difference. Other regressors included a target similarity regressor (i.e., whether the arrow directions were identical), a response similarity regressor (i.e., whether the correct responses were identical); a spatial Stroop

distractor regressor (i.e., vertical distance between two stimulus locations); a Simon distractor regressor (i.e., horizontal distance between two stimulus locations). Additionally, we also included a regressor denoting the similarity of Group (i.e., whether two conditions are within the same subject group, according to the stimulus-response mapping). We also added two regressors including ROI-mean fMRI activations for each condition of the pair to remove the possible uni-voxel influence on the RSM. A last term was the intercept. To control the artefact due to dependence of the correlation pairs sharing the same subject, we included crossed random effects (i.e., row-wise and column-wise random effects) for the intercept, conflict similarity, orientation and the group factors (G. Chen et al., 2017). Individual effects for each regressor were also extracted from the model for brain-behavioral correlation analyses. In brain-behavioral analyses, only the RT was used as behavioral measure to be consistent with the fMRI results, where the error trials were regressed out.

The statistical significance of these beta estimates was based on the outputs of the mixed-effect model estimated with the "fitlme" function in Matlab 2022a. Since symmetric cells from the RSM matrix were included in the mixed-effect model, we adjusted the t and p values with the true degree of freedom, which is half of the cells included minus the number of fixed regressors. Multiple comparison correction was applied with the Bonferroni approach across all cortical regions at the $p < 0.0001$ level. To test if the representation strengths are different between congruent and incongruent conditions, we also conducted the RSA using only congruent (RDM_C) and incongruent (RDM_I) trials separately. The contrast analysis was achieved by an additional model with both RDM_C and RDM_I included, adding the congruency and the interaction between conflict type (and orientation) and congruency as both fixed and random factors. The difference between incongruent and congruent representations was indicated by a significant interaction effect. To visualize the difference, we plotted the effect-related patterns (the predictor multiplied by the slope, plus the residual) as a function of the similarity levels (**Fig. 4D**), and a summary RSM for incongruent and congruent conditions based on the raw Pearson correlation coefficients, respectively (Fig. S5).

Model comparison and representational dimensionality

To estimate if the right 8C specifically encodes the cognitive space, rather than the domain-general or domain-specific structures, we conducted two more RSAs. We replaced the cognitive space-based conflict similarity matrix in the RSA we reported above (hereafter referred to as the Cognitive-Space model) with one of the alternative model matrices, with all other regressors equal. The domain-general model treats each conflict type as equivalent, so each two conflict types only differ in the magnitude of their conflict. Therefore, we defined the domain-general matrix as the difference in their congruency effects indexed by the group-averaged RT in Experiment 2. Then the z-scored model vector was sign-flipped to reflect similarity instead of distance. The domain-specific model treats each conflict type differently, so we used a diagonal matrix, with within-conflict type similarities being 1 and all cross-conflict type similarities being 0.

Moreover, to examine if the cognitive space is driven solely by the Stroop or Simon conflicts, we tested a spatial Stroop-Only (hereafter referred to as "Stroop-Only") and a Simon-Only model, with each conflict type projected onto the spatial Stroop (vertical) axis or Simon (horizontal) axis, respectively. The similarity between any two conflict types was defined using the Jaccard similarity index (Jaccard, 1901), that is, their intersection divided by their union. We also included a model assuming the Stroop and Simon dimensions are independently represented in the brain, adding up the Stroop-Only and Simon-Only regressors (hereafter referred to as the Stroop+Simon model). We conducted similar RSAs as reported above, replacing the original conflict similarity regressor with the Strrop-Only, Simon-Only, or both regressors (for the Stroop+Simon model), and then calculated their Bayesian information criterions (BICs).

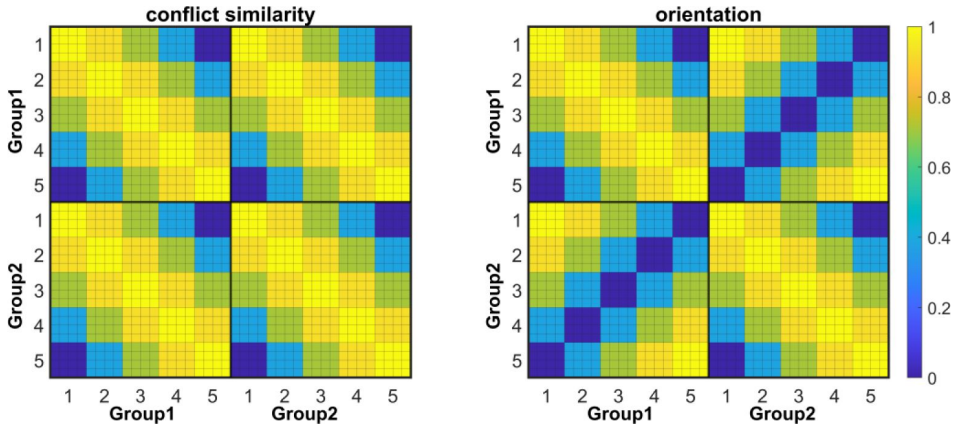


Figure 7.

Schematic of the orthogonality between conflict similarity and orientation.

The within-subject RSMs (e.g., Group1-Group1) for conflict similarity and orientation are all the same, but the cross-group correlations (e.g., Group2-Group1) are different. Therefore, we can separate the contribution of these two effects when including them as different regressors in the same linear regression model. The plotted matrices here include only one subject each from Group 1 and Group 2. Numbers 1-5 indicate the conflict type conditions, for spatial Stroop, St_HSm_L, St_MSm_M, St_LSm_H, and Simon, respectively. The thin lines separate four different sub-conditions, i.e., target arrow (up, down) × congruency (incongruent, congruent), within each conflict type.

To better capture the dimensionality of the representational space, we estimated its dimensionality using the participation ratio (Ito & Murray, 2023). Since we excluded the within-subject cells from the whole RSM, the whole RSM is an incomplete matrix and could not be used. To resolve this issue, we averaged the cells corresponding to each pair of conflict types to obtain an averaged 5×5 RSM matrix, similar to the matrix shown in **Fig. 1C**. We then estimated the participation ratio using the formula:

$$\dim = \frac{(\sum_i^m \lambda_i)^2}{\sum_i^m \lambda_i^2},$$

where λ_i is the eigenvalue of the RSM and m is the number of eigenvalues.

Representational connectivity analysis

To explore the possible relevance between the conflict type and the orientation effects, we conducted representational connectivity (Kriegeskorte et al., 2008) between regions showing evidence encoding conflict similarity and orientation similarity. We hypothesized that this relationship should exist at the within-subject level, so we conducted this analysis using within-subject RSMs excluding the diagonal. Specifically, the z-transformed RSM vector of each region were extracted and submitted to a mixed linear model, with the RSM of the conflict type region (i.e., the right 8C) as the dependent variable, and the RSM of one of the orientation regions (e.g., right V2) as the predictor. Intercept and the slope of the regressor were set as random effects at the subject level. The mixed effect model was conducted for each pair of regions, respectively. Considering there might be strong intrinsic correlations across the RSMs induced by the nuisance factors, such as the within-subject similarity, we added two sets of regions as control. First, we selected regions without showing any effects of interest (i.e., uncorrected *ps* > 0.3 for all the conflict type, orientation, congruency, target, response, spatial Stroop distractor and Simon distractor effects). Second, we selected regions of orientation effect meeting the first but not the second criterion, to account for the potential correlation between regions of the two partly orthogonal regressors (Fig. S6). Existence of representational connectivity was defined by a connectivity slope higher than 95% of the standard error above the mean of any control region.

Acknowledgements

We thank Eliot Hazeltine for valuable comments on a previous version of this manuscript. The work was supported by the National Natural Science Foundation of China and the German Research Foundation (NSFC 62061136001/DFG TRR-169) to X.L. and China Postdoctoral Science Foundation (2019M650884) to G.Y.

References

Abrahamse E., Braem S., Notebaert W., Verguts T (2016) **Grounding cognitive control in associative learning** *Psychological Bulletin* **142**:693–728 https://doi.org/10.1037/bul0000047

Akcay C., Hazeltine E (2011) **Domain-specific conflict adaptation without feature repetitions** *Psychonomic Bulletin & Review* **18**:505–511 https://doi.org/10.3758/s13423-011-0084-y

Assem M., Glasser M. F., Van Essen D. C., Duncan J. (2020) **A Domain-General Cognitive Core Defined in Multimodally Parcellated Human Cortex** *Cerebral Cortex* **30**:4361–4380 https://doi.org/10.1093/cercor/bhaa023

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

Behrens T. E. J., Muller T. H., Whittington J. C. R., Mark S., Baram A. B., Stachenfeld K. L., Kurth-Nelson Z (2018) **What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior** *Neuron* **100**:490–509 https://doi.org/10.1016/j.neuron.2018.10.002

Bellmund J. L. S., Gardenfors P., Moser E. I., Doeller C. F (2018) **Navigating cognition: Spatial codes for human thinking** *Science (New York, N.Y.)* **362** https://doi.org/10.1126/science.aat6766

Botvinick M. M., Cohen J. D., Carter C. S (2004) **Conflict monitoring and anterior cingulate cortex: an update** *Trends in Cognitive Sciences* **8**:539–546 https://doi.org/10.1016/j.tics.2004.10.003

Braem S., Abrahamse E. L., Duthoo W., Notebaert W (2014) **What determines the specificity of conflict adaptation? A review, critical analysis, and proposed synthesis** *Frontiers in Psychology* **5** https://doi.org/10.3389/fpsyg.2014.01134

Braem S., Bugg J. M., Schmidt J. R., Crump M. J. C., Weissman D. H., Notebaert W., Egner T (2019) **Measuring Adaptive Control in Conflict Tasks** *Trends in Cognitive Sciences* **23**:769–783 https://doi.org/10.1016/j.tics.2019.07.002

Chen G., Taylor P. A., Shin Y.-W., Reynolds R. C., Cox R. W. J. N (2017) **Chen, G., Taylor, P. A., Shin, Y.-W., Reynolds, R. C., & Cox, R. W. J. N. (2017). Untangling the relatedness among correlations, Part II: Inter-subject correlation group analysis through linear mixed-effects modeling. 147, 825–840.** *Untangling the relatedness among correlations, Part II: Inter-subject correlation group analysis through linear mixed-effects modeling* **147**:825–840

Chen J., Leong Y. C., Honey C. J., Yong C. H., Norman K. A., Hasson U (2017) **Shared memories reveal shared structure in neural activity across individuals** *Nature Neuroscience* **20**:115–125 https://doi.org/10.1038/nn.4450

Cole M. W., Reynolds J. R., Power J. D., Repovs G., Anticevic A., Braver T. S (2013) **Multi-task connectivity reveals flexible hubs for adaptive task control** *Nature Neuroscience* **16**:1348–1355 https://doi.org/10.1038/nn.3470

Constantinescu A. O., O'Reilly J. X., Behrens T. E. J. (2016) **Organizing conceptual knowledge in humans with a gridlike code** *Science (New York, N.Y.)* **352**:1464–1468 https://doi.org/10.1126/science.aaf0941

Cosmides L., Tooby J, Hirschfeld L. A., Gelman S. A. (1994) **Origins of domain specificity: The evolution of functional organization** *Mapping the mind: Domain specificity in cognition and culture*

Cox R. W., Hyde J. S (1997) **Software tools for analysis and visualization of fMRI data** *NMR in Biomedicine* **10**:171–178 https://doi.org/10.1002/(sici)1099-1492(199706/08)10:4/5<171::Aid-nbm453>3.0.Co;2-l

Duncan J (2010) **The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour** *Trends in Cognitive Sciences* **14**:172–179 https://doi.org/10.1016/j.tics.2010.01.004

Duncan J (2013) **The structure of cognition: attentional episodes in mind and brain** *Neuron* **80**:35–50 https://doi.org/10.1016/j.neuron.2013.09.015

Durnez J., Blair R., Poldrack R. A (2018) **Neurodesign: Optimal Experimental Designs for Task fMRI** *bioRxiv* **119594** https://doi.org/10.1101/119594

Egner T (2007) **Congruency sequence effects and cognitive control** *Cognitive, Affective & Behavioral Neuroscience* **7**:380–390 https://doi.org/10.3758/cabn.7.4.380

Egner T (2008) **Multiple conflict-driven control mechanisms in the human brain** *Trends in Cognitive Sciences* **12**:374–380 https://doi.org/10.1016/j.tics.2008.07.001

Egner T., Delano M., Hirsch J (2007) **Separate conflict-specific cognitive control mechanisms in the human brain** *NeuroImage* **35**:940–948 https://doi.org/10.1016/j.neuroimage.2006.11.061

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** *Nature Methods* **16**:111–116 https://doi.org/10.1038/s41592-018-0235-4

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* https://doi.org/10.1016/j.neuron.2022.01.005

Freitas A. L., Bahar M., Yang S., Banai R (2007) **Contextual adjustments in cognitive control across tasks** *Psychological Science* **18**:1040–1043 https://doi.org/10.1111/j.1467-9280.2007.02022.x

Freitas A. L., Clark S. L (2015) **Generality and specificity in cognitive control: conflict adaptation within and across selective-attention tasks but not across selective-attention and Simon tasks** *Psychological Research* **79**:143–162 https://doi.org/10.1007/s00426-014-0540-1

Freund M. C., Bugg J. M., Braver T. S (2021) **A Representational Similarity Analysis of Cognitive Control during Color-Word Stroop** *Journal of Neuroscience* **41**:7388–7402 https://doi.org/10.1523/JNEUROSCI.2956-20.2021

Freund M. C., Etzel J. A., Braver T. S (2021) **Neural Coding of Cognitive Control: The Representational Similarity Analysis Approach** *Trends in Cognitive Sciences* **25**:622–638 https://doi.org/10.1016/j.tics.2021.03.011

Friehs M. A., Klaus J., Singh T., Frings C., Hartwigsen G (2020) **Perturbation of the right prefrontal cortex disrupts interference control** *NeuroImage* **222** https://doi.org/10.1016/j.neuroimage.2020.117279

Fu Z., Beam D., Chung J. M., Reed C. M., Mamelak A. N., Adolphs R., Rutishauser U (2022) **The geometry of domain-general performance monitoring in the human medial frontal cortex** *Science (New York, N.Y.)* **376** https://doi.org/10.1126/science.abm9922

Gilbert C. D., Li W (2013) **Top-down influences on visual processing** *Nat Rev Neurosci* **14**:350–363 https://doi.org/10.1038/nrn3476

Glasser M. F., Coalson T. S., Robinson E. C., Hacker C. D., Harwell J., Yacoub E., Van Essen D. C. (2016) **A multi-modal parcellation of human cerebral cortex** *Nature* **536**:171–178 https://doi.org/10.1038/nature18933

Gorgolewski K., Burns C. D., Madison C., Clark D., Halchenko Y. O., Waskom M. L., Ghosh S. S (2011) **Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python** *Frontiers in Neuroinformatics* **5** https://doi.org/10.3389/fninf.2011.00013

Grahek I., Leng X., Fahey M. P., Yee D., Shenhav A (2022) **). Empirical and Computational Evidence for Reconfiguration Costs During Within-Task Adjustments in Cognitive Control. Proceedings of the Annual Meeting of the Cognitive Science Society**

Hazeltine E., Lightman E., Schwarb H., Schumacher E. H. (2011) **The boundaries of sequential modulations: evidence for set-level control** *Journal of Experimental Psychology: Human Perception and Performance* **37**:1898–1914 https://doi.org/10.1037/a0024662

Hommel B., Proctor R. W., Vu K. P (2004) **A feature-integration account of sequential effects in the Simon task** *Psychological Research* **68**:1–17 https://doi.org/10.1007/s00426-003-0132-y

Ito T., Murray J. D (2023) **Multitask representations in the human cortex transform along a sensory-to-motor hierarchy** *Nat Neurosci* **26**:306–315 https://doi.org/10.1038/s41593-022-01224-0

Jaccard P (1901) **Étude comparative de la distribution florale dans une portion des Alpes et des Jura** *Bull Soc Vaudoise Sci Nat* **37**:547–579 https://doi.org/10.5169/seals-266450

Jenkinson M., Bannister P., Brady M., Smith S (2002) **Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images** *NeuroImage* **17**:825–841 https://doi.org/10.1006/nimg.2002.1132

Jiang J., Egner T (2014) **Using neural pattern classifiers to quantify the modularity of conflict-control mechanisms in the human brain** *Cerebral Cortex* **24**:1793–1805 https://doi.org/10.1093/cercor/bht029

Jiang J., Summerfield C., Egner T (2016) **Visual Prediction Error Spreads Across Object Features in Human Visual Cortex** *Journal of Neuroscience* **36**:12746–12763 https://doi.org/10.1523/JNEUROSCI.1546-16.2016

Jiang J., Wang S. F., Guo W., Fernandez C., Wagner A. D (2020) **Prefrontal reinstatement of contextual task demand is predicted by separable hippocampal patterns** *Nat Commun* **11** https://doi.org/10.1038/s41467-020-15928-z

Kan I. P., Teubner-Rhodes S., Drummey A. B., Nutile L., Krupa L., Novick J. M (2013) **To adapt or not to adapt: the question of domain-general cognitive control** *Cognition* **129**:637–651 https://doi.org/10.1016/j.cognition.2013.09.001

Kim C., Chung C., Kim J (2012) **Conflict adjustment through domain-specific multiple cognitive control mechanisms** *Brain Research* **1444**:55–64 https://doi.org/10.1016/j.brainres.2012.01.023

Kornblum S., Hasbroucq T., Osman A (1990) **Dimensional overlap: cognitive basis for stimulus-response compatibility--a model and taxonomy** *Psychological Review* **97**:253–270 https://doi.org/10.1037/0033-295x.97.2.253

Kriegeskorte N., Mur M., Bandettini P (2008) **Representational similarity analysis - connecting the branches of systems neuroscience** *Frontiers in Systems Neuroscience* **2** https://doi.org/10.3389/neuro.06.004.2008

Li Q., Nan W., Wang K., Liu X (2014) **Independent processing of stimulus-stimulus and stimulus-response conflicts** *PloS One* **9** https://doi.org/10.1371/journal.pone.0089249

Li Q., Yang G., Li Z., Qi Y., Cole M. W., Liu X (2017) **Conflict detection and resolution rely on a combination of common and distinct cognitive control networks** *Neuroscience and Biobehavioral Reviews* **83**:123–131 https://doi.org/10.1016/j.neubiorev.2017.09.032

Liu X., Banich M. T., Jacobson B. L., Tanabe J. L (2004) **Common and distinct neural substrates of attentional control in an integrated Simon and spatial Stroop task as assessed by event-related fMRI** *NeuroImage* **22**:1097–1106 https://doi.org/10.1016/j.neuroimage.2004.02.033

Liu X., Park Y., Gu X., Fan J (2010) **Dimensional overlap accounts for independence and integration of stimulus-response compatibility effects** *Attention, Perception, & Psychophysics* **72**:1710–1720 https://doi.org/10.3758/APP.72.6.1710

Lu C. H., Proctor R. W (1995) **The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects** *Psychonomic Bulletin & Review* **2**:174–207 https://doi.org/10.3758/BF03210959

MacDowell C. J., Tafazoli S., Buschman T. J (2022) **A Goldilocks theory of cognitive control: Balancing precision and efficiency with low-dimensional control states** *Curr Opin Neurobiol* **76** https://doi.org/10.1016/j.conb.2022.102606

MacLeod C. M (1991) **Half a century of research on the Stroop effect: an integrative review** *Psychological Bulletin* **109**:163–203

Magen H., Cohen A (2007) **Modularity beyond perception: evidence from single task interference paradigms** *Cognitive Psychology* **55**:1–36 https://doi.org/10.1016/j.cogpsych.2006.09.003

Mansouri F. A., Buckley M. J., Tanaka K (2007) **Mnemonic function of the dorsolateral prefrontal cortex in conflict-induced behavioral adjustment** *Science (New York, N.Y.)* **318**:987–990 https://doi.org/10.1126/science.1146384

Miller E. K., Cohen J. D (2001) **An integrative theory of prefrontal cortex function** *Annual Review of Neuroscience* **24**:167–202 https://doi.org/10.1146/annurev.neuro.24.1.167

Milner B (1963) **Effects of Different Brain Lesions on Card Sorting - Role of Frontal Lobes** *Archives of Neurology* **9** https://doi.org/DOI10.1001/archneur.1963.00460070100010

Musslick S., Cohen J. D (2021) **Rationalizing constraints on the capacity for cognitive control** *Trends Cogn Sci* **25**:757–775 https://doi.org/10.1016/j.tics.2021.06.001

Nili H., Wingfield C., Walther A., Su L., Marslen-Wilson W., Kriegeskorte N (2014) **A toolbox for representational similarity analysis** *PLoS Computational Biology* **10** https://doi.org/10.1371/journal.pcbi.1003553

Niv Y (2019) **Learning task-state representations** *Nat Neurosci* **22**:1544–1553 https://doi.org/10.1038/s41593-019-0470-8

Park S. A., Miller D. S., Nili H., Ranganath C., Boorman E. D (2020) **Map Making: Constructing, Combining, and Inferring on Abstract Cognitive Maps** *Neuron* **107**:1226–1238 https://doi.org/10.1016/j.neuron.2020.06.030

Peterson B. S., Kane M. J., Alexander G. M., Lacadie C., Skudlarski P., Leung H. C., Gore J. C. (2002) **An event-related functional MRI study comparing interference effects in the Simon and Stroop tasks** *Brain Research: Cognitive Brain Research* **13**:427–440 https://doi.org/10.1016/s0926-6410(02)00054-x

Polk T. A., Drake R. M., Jonides J. J., Smith M. R., Smith E. E (2008) **Attention enhances the neural processing of relevant features and suppresses the processing of irrelevant features in humans: a functional magnetic resonance imaging study of the Stroop task** *Journal of Neuroscience* **28**:13786–13792 https://doi.org/10.1523/JNEUROSCI.1026-08.2008

Reverberi C., Gorgen K., Haynes J. D (2012) **Compositionality of rule representations in human prefrontal cortex** *Cereb Cortex* **22**:1237–1246 https://doi.org/10.1093/cercor/bhr200

Ritz H., Shenhav A (2022) **Humans reconfigure target and distractor processing to address distinct task demands** *bioRxiv* https://doi.org/10.1101/2021.09.08.459546

Rmus M., Ritz H., Hunter L. E., Bornstein A. M., Shenhav A (2022) **Humans can navigate complex graph structures acquired during latent learning** *Cognition* **225** https://doi.org/10.1016/j.cognition.2022.105103

Schmidt J. R., Weissman D. H (2014) **Congruency sequence effects without feature integration or contingency learning confounds** *PloS One* **9** https://doi.org/10.1371/journal.pone.0102337

Schuck N. W., Cai M. B., Wilson R. C., Niv Y (2016) **Human Orbitofrontal Cortex Represents a Cognitive Map of State Space** *Neuron* **91**:1402–1412 https://doi.org/10.1016/j.neuron.2016.08.019

Shenhav A., Botvinick M. M., Cohen J. D (2013) **The expected value of control: an integrative theory of anterior cingulate cortex function** *Neuron* **79**:217–240 https://doi.org/10.1016/j.neuron.2013.07.007

Simon J. R., Small A. M. (1969) **Processing auditory information: interference from an irrelevant cue** *Journal of Applied Psychology* **53**:433–435 https://doi.org/10.1037/h0028034

Torres-Quesada M., Funes M. J., Lupianez J (2013) **Dissociating proportion congruent and conflict adaptation effects in a Simon-Stroop procedure** *Acta Psychologica* **142**:203–210 https://doi.org/10.1016/j.actpsy.2012.11.015

Tusche A., Bockler A., Kanske P., Trautwein F. M., Singer T (2016) **Decoding the Charitable Brain: Empathy, Perspective Taking, and Attention Shifts Differentially Predict Altruistic Giving** *Journal of Neuroscience* **36**:4719–4732 https://doi.org/10.1523/JNEUROSCI.3392-15.2016

Vaidya A. R., Badre D (2022) **Abstract task representations for inference and control** *Trends Cogn Sci* **26**:484–498 https://doi.org/10.1016/j.tics.2022.03.009

Vaidya A. R., Jones H. M., Castillo J., Badre D (2021) **Neural representation of abstract task structure during generalization** *Elife* **10**:1–26 https://doi.org/10.7554/eLife.63226

Vanderhasselt M. A., De Raedt R., Baeken C. (2009) **Dorsolateral prefrontal cortex and Stroop performance: tackling the lateralization** *Psychonomic Bulletin & Review* **16**:609–612 https://doi.org/10.3758/PBR.16.3.609

Wager T. D., Nichols T. E (2003) **Optimization of experimental design in fMRI: a general framework using a genetic algorithm** *NeuroImage* **18**:293–309 https://doi.org/10.1016/S1053-8119(02)00046-0

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 https://doi.org/10.1016/j.neuroimage.2015.12.012

Wang K., Li Q., Zheng Y., Wang H., Liu X (2014) **Temporal and spectral profiles of stimulus-stimulus and stimulus-response conflict processing** *NeuroImage* **89**:280–288 https://doi.org/10.1016/j.neuroimage.2013.11.045

Wu T., Spagna A., Chen C., Schulz K. P., Hof P. R., Fan J (2020) **Supramodal Mechanisms of the Cognitive Control Network in Uncertainty Processing** *Cerebral Cortex* **30**:6336–6349

Yang G., Nan W., Zheng Y., Wu H., Li Q., Liu X (2017) **Distinct cognitive control mechanisms as revealed by modality-specific conflict adaptation effects** *Journal of Experimental Psychology: Human Perception and Performance* **43**:807–818 https://doi.org/10.1037/xhp0000351

Yang G., Xu H., Li Z., Nan W., Wu H., Li Q., Liu X (2021) **The congruency sequence effect is modulated by the similarity of conflicts** *Journal of Experimental Psychology: Learning, Memory, and Cognition* **47**:1705–1719 https://doi.org/10.1037/xlm0001054

Author information

Guochun Yang

CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing 100101, China, Department of Psychology, University of Chinese Academy of Sciences, Beijing 100101, China, Department of Psychological and Brain Sciences, University of Iowa, Iowa City, IA 52242, USA, Cognitive Control Collaborative, University of Iowa, Iowa City, IA 52242, USA
ORCID iD: 0000-0002-0516-8772

Haiyan Wu
Centre for Cognitive and Brain Sciences and Department of Psychology, University of Macau, Taipa, Macau 999078, China
ORCID iD: 0000-0001-8869-6636

Qi Li
Beijing Key Laboratory of Learning and Cognition, School of Psychology, Capital Normal University, Beijing 100048, China

Xun Liu
CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing 100101, China, Department of Psychology, University of Chinese Academy of Sciences, Beijing 100101, China
For correspondence: liux@psych.ac.cn
ORCID iD: 0000-0003-1366-8926

Zhongzheng Fu
Department of Neurosurgery, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA, Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA

Jiefeng Jiang
Department of Psychological and Brain Sciences, University of Iowa, Iowa City, IA 52242, USA, Cognitive Control Collaborative, University of Iowa, Iowa City, IA 52242, USA

Editors

Reviewing Editor
David Badre
Brown University, United States of America

Senior Editor
Michael Frank
Brown University, United States of America

Reviewer #1 (Public Review):

People can perform a wide variety of different tasks, and a long-standing question in cognitive neuroscience is how the properties of different tasks are represented in the brain. The authors develop an interesting task that mixes two different sources of difficulty, and find that the brain appears to represent this mixture on a continuum, in the prefrontal areas involved in resolving task difficulty. While these results are interesting and in several ways compelling, they overlap with previous findings and rely on novel statistical analyses that may require further validation.

Strengths
1. The authors present an interesting and novel task for combining the contributions of stimulus-stimulus and stimulus-response conflict. While this mixture has been measured in the multi-source interference task (MSIT), this task provides a more graded mixture between these two sources of difficulty.

2. The authors do a good job triangulating regions that encoding conflict similarity, looking for the conjunction across several different measures of conflict encoding. These conflict

measures use several best-practice approaches towards estimating representational similarity.

3. The authors quantify several salient alternative hypothesis and systematically distinguish their core results from these alternatives.

4. The question that the authors tackle is important to cognitive control, and they make a solid contribution.

Concerns
1. The evidence from this previous work for mixtures between different conflict sources makes the framing of 'infinite possible types of conflict' feel like a strawman. The authors cite classic work (e.g., Kornblum et al., 1990) that develops a typology for conflict which is far from infinite. I think few people would argue that every possible source and level of difficulty will have to be learned separately. This work provides confirmatory evidence that task difficulty is represented parametrically (e.g., consistent with the n-back, MOT, and random dot motion literature).

2. The degree of Stroop vs Simon conflict is perfectly negatively correlated across conditions. This limits their interpretation of an integrated cognitive space, as they cannot separately measure Stroop and Simon effects. The author's control analyses have limited ability to overcome this task limitation. While these results are consistent with parametric encoding, they cannot adjudicate between combined vs separated representations.

Reviewer #2 (Public Review):

This study examines the construct of "cognitive spaces" as they relate to neural coding schemes present in response conflict tasks. The authors use a novel experimental design in which different types of response conflict (spatial Stroop, Simon) are parametrically manipulated. These conflict types are hypothesized to be encoded jointly, within an abstract "cognitive space", in which distances between task conditions depend only on the similarity of conflict types (i.e., where conditions with similar relative proportions of spatial-Stroop versus Simon conflicts are represented with similar activity patterns). Authors contrast such a representational scheme for conflict with several other conceptually distinct schemes, including a domain-general, domain-specific, and two task-specific schemes. The authors conduct a behavioral and fMRI study to test which of these coding schemes is used by prefrontal cortex. Replicating the authors' prior work, this study demonstrates that sequential behavioral adjustments (the congruency sequence effect) are modulated as a function of the similarity between conflict types. In fMRI data, univariate analyses identified activation in left prefrontal and dorsomedial frontal cortex that was modulated by the amount of Stroop or Simon conflict present, and representational similarity analyses (RSA) that identified coding of conflict similarity, as predicted under the cognitive space model, in right lateral prefrontal cortex.

This study tackles an important question regarding how distinct types of conflict might be encoded in the brain within a computationally efficient representational format. The ideas postulated by the authors are interesting ones and the statistical methods are generally rigorous. The evidence supporting the authors claims, however, is limited by confounds in the experimental design and by lack of clarity in reporting the testing of alternative hypotheses within the method and results.

(1) Model comparison

The authors commendably performed a model comparison within their study, in which they formalized alternative hypotheses to their cognitive space hypothesis. We greatly appreciate the motivation for this idea and think that it strengthened the manuscript. Nevertheless,

some details of this model comparison were difficult for us to understand, which in turn has limited our understanding of the strength of the findings.

The text indicates the domain-general model was computed by taking the difference in congruency effects per conflict condition. Does this refer to the "absolute difference" between congruency effects? In the rest of this review, we assume that the absolute difference was indeed used, as using a signed difference would not make sense in this setting. Nevertheless, it may help readers to add this information to the text.

Regarding the Stroop-Only and Simon-Only models, the motivation for using the Jaccard metric was unclear. From our reading, it seems that all of the other models --- the cognitive space model, the domain-general model, and the domain-specific model --- effectively use a Euclidean distance metric. (Although the cognitive space model is parameterized with cosine similarities, these similarity values are proportional to Euclidean distances because the points all lie on a circle. And, although the domain-general model is parameterized with absolute differences, the absolute difference is equivalent to Euclidean distance in 1D.) Given these considerations, the use of Jaccard seems to differ from the other models, in terms of parameterization, and thus potentially also in terms of underlying assumptions. Could authors help us understand why this distance metric was used instead of Euclidean distance? Additionally, if Jaccard must be used because this metric seems to be non-standard in the use of RSA, it would likely be helpful for many readers to give a little more explanation about how it was calculated.

When considering parameterizing the Stroop-Only and Simon-Only models with Euclidean distances, one concern we had is that the joint inclusion of these models might render the cognitive space model unidentifiable due to collinearity (i.e., the sum of the Stroop-Only and Simon-Only models could be collinear with the cognitive space model). Could the authors determine whether this is the case? This issue seems to be important, as the presence of such collinearity would suggest to us that the design is incapable of discriminating those hypotheses as parameterized.

(2) Issue of uniquely identifying conflict coding

We certainly appreciate the efforts that authors have taken to address potential confounders for encoding of conflict in their original submission. We broach this question not because we wish authors to conduct additional control analyses, but because this issue seems to be central to the thesis of the manuscript and we would value reading the authors' thoughts on this issue in the discussion.

To summarize our concerns, conflict seems to be a difficult variable to isolate within aggregate neural activity, at least relative to other variables typically studied in cognitive control, such as task-set or rule coding. This is because it seems reasonable to expect that many more nuisance factors covary with conflict --- such as univariate activation, level of cortical recruitment, performance measures, arousal --- than in comparison with, for example, a well-designed rule manipulation. Controlling for some of these factors post-hoc through regression is commendable (as authors have done here), but such a method will likely be incomplete and can provide no guarantees on the false positive rate.

Relatedly, the neural correlates of conflict coding in fMRI and other aggregate measures of neural activity are likely of heterogeneous provenance, potentially including rate coding (Fu et al., 2022), temporal coding (Smith et al., 2019), modulation of coding of other more concrete variables (Ebitz et al., 2020, 10.1101/2020.03.14.991745; see also discussion and reviews of Tang et al., 2016, 10.7554/eLife.12352), or neuromodulatory effects (e.g., Aston-Jones & Cohen, 2005). Some of these origins would seem to be consistent with "explicit" coding of conflict (conflict as a representation), but others would seem to be more consistent with epiphenomenal coding of conflict (i.e., conflict as an emergent process). Again, these concerns

could apply to many variables as measured via fMRI, but at the same time, they seem to be more pernicious in the case of conflict. So, if authors consider these issues to be germane, perhaps they could explicitly state in the discussion whether adopting their cognitive space perspective implies a particular stance on these issues, how they interpret their results with respect to these issues, and if relevant, qualify their conclusions with uncertainty on these issues.

(3) Interpretation of measured geometry in 8C

We appreciate the inclusion of the measured similarity matrices of area 8C, the key area the results focus on, to the supplemental, as this allows for a relatively model-agnostic look at a portion of the data. Interestingly, the measured similarity matrix seems to mismatch the cognitive space model in a potentially substantive way. Although the model predicts that the "pure" Stroop and Simon conditions will have maximal self-similarity (i.e., the Stroop-Stroop and Simon-Simon cells on the diagonal), these correlations actually seem to be the lowest, by what appears to be a substantial margin (particularly the Stroop-Stroop similarities). What should readers make of this apparent mismatch? Perhaps authors could offer their interpretation on how this mismatch could fit with their conclusions.

Reviewer #3 (Public Review):

Yang and colleagues investigated whether information on two task-irrelevant features that induce response conflict is represented in a common cognitive space. To test this, the authors used a task that combines the spatial Stroop conflict and the Simon effect. This task reliably produces a beautiful graded congruency sequence effect (CSE), where the cost of congruency is reduced after incongruent trials. The authors measured fMRI to identify brain regions that represent the graded similarity of conflict types, the congruency of responses, and the visual features that induce conflicts. They applied univariate, multivariate, and connectivity analyses to fMRI data to identify brain regions that represent the graded similarity of conflict types, the congruency of responses, and the visual features that induce conflicts. They further directly assessed the dimensionality of represented conflict space.

The authors identified the right dlPFC (right 8C), which shows 1) stronger encoding of graded similarity of conflicts in incongruent trials and 2) a positive correlation between the strength of conflict similarity type and the CSE on behavior. The dlPFC has been shown to be important for cognitive control tasks. As the dlPFC did not show a univariate parametric modulation based on the higher or lower component of one type of conflict (e.g., having more spatial Stroop conflict or less Simon conflict), it implies that dissimilarity of conflicts is represented by a linear increase or decrease of neural responses. Therefore, the similarity of conflict is represented in multivariate neural responses that combine two sources of conflict.

The strength of the current approach lies in the clear effect of parametric modulation of conflict similarity across different conflict types. The authors employed a clever cross-subject RSA that counterbalanced and isolated the targeted effect of conflict similarity, decorrelating orientation similarity of stimulus positions that would otherwise be correlated with conflict similarity. A pattern of neural response seems to exist that maps different types of conflict, where each type is defined by the parametric gradation of the yoked spatial Stroop conflict and the Simon conflict on a similarity scale. The similarity of patterns increases in incongruent trials and is correlated with CSE modulation of behavior.

The main significance of the paper lies in the evidence supporting the use of an organized "cognitive space" to represent conflict information as a general control strategy. The authors thoroughly test this idea using multiple approaches and provide convincing support for their findings. However, the universality of this cognitive strategy remains an open question.

The task presented in the study involved two sources of conflict information through a single salient visual input, which might have encouraged the utilization of a common space. The similarity space was analyzed at the level of between-individuals (i.e., cross-subject RSA) to mitigate potential confounds in the design, such as congruency and the orientation of stimulus positions. This approach makes it challenging to establish a direct link between the quality of conflict space representation and the patterns of behavioral adaptations within individuals.

Furthermore, it remains unclear at which cognitive stages during response selection such a unified space is recruited. Can we effectively map any sources of conflict into a single scale? Is this unified space adaptively adjusted within the same brain region? Additionally, does the amount of conflict solely define the dimensions of this unified space across many conflict-inducing tasks? These questions remain open for future studies to address.

Taken together, this study presents an exciting possibility that information requiring high levels of cognitive control could be flexibly mapped into cognitive map-like representations that both benefit and bias our behavior. Further characterization of the representational geometry and generalization of the current results look promising ways to understand representations for cognitive control.

Author Response

The following is the authors' response to the original reviews.

Reviewer #1:

People can perform a wide variety of different tasks, and a long-standing question in cognitive neuroscience is how the properties of different tasks are represented in the brain. The authors develop an interesting task that mixes two different sources of difficulty, and find that the brain appears to represent this mixture on a continuum, in the prefrontal areas involved in resolving task difficulty. While these results are interesting and in several ways compelling, they overlap with previous findings and rely on novel statistical analyses that may require further validation.

Strengths

1. *The authors present an interesting and novel task for combining the contributions of stimulus-stimulus and stimulus-response conflict. While this mixture has been measured in the multi-source interference task (MSIT), this task provides a more graded mixture between these two sources of difficulty*

1. *The authors do a good job triangulating regions that encoding conflict similarity, looking for the conjunction across several different measures of conflict encoding*

1. *The authors quantify several salient alternative hypothesis and systematically distinguish their core results from these alternatives*

1. *The question that the authors tackle is of central theoretical importance to cognitive control, and they make an interesting an interesting contribution to this question*

We would like to thank the reviewer for the positive evaluation of our manuscript and the constructive comments and suggestions. Your feedback has been invaluable in our efforts to enhance the accessibility of our manuscript and strengthen our findings. In response to your suggestion, we reanalyzed our data using the approach proposed by Chen et al.'s (2017, NeuroImage) and applied stricter multiple comparison correction thresholds in our reporting. This reanalysis largely replicated our previous results, thereby reinforcing the robustness of our findings. We also have examined several alternative models and results supported the integration of the spatial Stroop and Simon conflicts within the cognitive space. In addition, we enriched the theoretical framework of our manuscript by connecting the cognitive space with other important theories such as the "Expected Value of Control" theory. We have incorporated your feedback, revisions and additional analyses into the manuscript. As a result, we firmly believe that these changes have significantly improved the quality of our work. We have provided detailed responses to your comments below.

> *1. It's not entirely clear what the current task can measure that is not known from the MSIT, such as the additive influence of conflict sources in Fu et al. (2022), Science. More could be done to distinguish the benefits of this task from MSIT.*

We agree that the MSIT task incorporates Simon and Eriksen Flanker conflict tasks and can efficiently detect the additivity of conflict effects across orthogonal tasks. Like the MSIT, our task incorporates Simon with spatial Stroop conflicts and can test the same idea. For example, a previous study from our lab (Li et al., 2014) used the combined spatial Stroop-Simon condition with the arrows displayed on diagonal corners and found evidence for the additive hypothesis. However, the MSIT cannot be used to test whether/how different conflicts are parametrically represented in a low-dimensional space, a question that is important to address the debate of domain-general and domain-specific cognitive control.

To this end, our current study adopted the spatial Stroop-Simon task for the unique purpose of parametrically modulating conflict similarity. As far as we know, there is no way to define the similarity between the combined Simon_Flanker conflict condition and the Simon/Flanker conditions in the MSIT. In contrast, with the spatial Stroop-Simon paradigm, we can define the similarity with the cosine of the angle difference across the two conditions in question.

We have added the following texts in the discussion part to emphasize the 51 difference between our paradigm and other studies.

"The use of an experimental paradigm that permits parametric manipulation of conflict similarity provides a way to systematically investigate the organization of cognitive control, as well as its influence on adaptive behaviors. This approach extends traditional paradigms, such as the multi-source interference task (Fu et al., 2022), color Stroop-Simon task (Liu et al., 2010) and similar paradigms that do not afford a quantifiable metric of conflict source similarity."

References:

Li, Q., Nan, W., Wang, K., & Liu, X. (2014). Independent processing of stimulus-stimulus and stimulus-response conflicts. PloS One, 9(2), e89249.

> *1. The evidence from this previous work for mixtures between different conflict sources make the framing of 'infinite possible types of conflict' feel like a strawman. The authors cite classic work (e.g., Kornblum et al., 1990) that develops a typology for conflict which is far from infinite, and I think few people would argue that every possible source of difficulty will have to be learned separately. Such an issue is addressed in theories like 'Expected Value of Control', where optimization of control policies can address unique combinations of task demands.*

The notion that there might be infinite conflicts arises when we consider the quantitative feature of cognitive control. If each combination of the Stroop-Simon combination is regarded as a conflict condition, there would be infinite combinations, and it is our major goal to investigate how these infinite conflict conditions are represented effectively in a space with finite dimensions. We agree that it is unnecessary to dissociate each of these conflict conditions into a unique conflict type, since they may not differ substantially. However, we argue that understanding variant conflicts within a purely categorical framework (e.g., Simon and Flanker conflict in MSIT) is insufficient, especially because it leads to dichotomic conclusions that do not capture how combinations of conflicts are organized in the brain, as our study addresses.

There could be different perspectives on how our cognitive control system flexibly encodes and resolves multiple conflicts. The cognitive space assumption we held provides a principle by which we can represent multiple conflicts in a lower dimensional space efficiently. While the “Expected Value of Control” theory addresses when and how much cognitive control to apply based on control demand, the “cognitive space” view seeks to explain how the conflict, which defines cognitive control demand, is encoded in the brain. Thus, we argue that these two lines of work are different yet complementary. The geometry of cognitive space of conflict can benefit the adjustment of cognitive control for upcoming conflicts. For example, our brain may evaluate the similarity/distance (and thus cost) between the consecutive conflict conditions, and selects the path with best cost-benefit tradeoff to switch from one state to another. This idea is conceptually similar to a recent study by Grahek et al. (2022) demonstrating that more frequently switching states were encoded as closer together than less frequently switching states in a “drift-threshold” space.

Nevertheless, Grahek et al (2022) investigated how cognitive control changes based on the expected value of control theory within the same conflict, whereas our study aims to examine organization of different conflict.

We have added the implications of cognitive space view in the discussion to indicate the potential values of our finding to understand the EVC account and the difference between the two theories.

“Previous researchers have proposed an “expected value of control (EVC)” theory, which posits that the brain can evaluate the cost and benefit associated with executing control for a demanding task, such as the conflict task, and specify the optimal control strength (Shenhav et al., 2013). For instance, Grahek et al. (2022) found that more frequently switching goals when doing a Stroop task were achieved by adjusting smaller control intensity. Our work complements the EVC theory by further investigating the neural representation of different conflict conditions and how these representations can be evaluated to facilitate conflict resolution. We found that different conflict conditions can be efficiently represented in a cognitive space encoded by the right dlPFC, and participants with stronger cognitive space representation have also adjusted their conflict control to a greater extent based on the conflict similarity (Fig 4C). The finding suggests that the cognitive space organization of conflicts guides cognitive control to adjust behavior. Previous studies have shown that participants may adopt different strategies to represent a task, with the model-based

strategies benefitting goal-related behaviors more than the model-free strategies (Rmus et al., 2022). Similarly, we propose that cognitive space could serve as a mental model to assist fast learning and efficient organization of cognitive control settings. Specifically, the cognitive space representation may provide a principle for how our brain evaluates the expected cost of switching and the benefit of generalization between states and selects the path with the best cost-benefit tradeoff (Abrahamse et al., 2016; Shenhav et al., 2013). The proximity between two states in cognitive space could reflect both the expected cognitive demand required to transition and the useful mechanisms to adapt from. The closer the two conditions are in cognitive space, the lower the expected switching cost and the higher the generalizability when transitioning between them. With the organization of a cognitive space, a new conflict can be quickly assigned a location in the cognitive space, which will facilitate the development of cognitive control settings for this conflict by interpolating nearby conflicts and/or projecting the location to axes representing different cognitive control processes, thus leading to a stronger CSE when following a more similar conflict condition. On the other hand, without a cognitive space, there would be no measure of similarity between conflicts on different trials, hence limiting the ability of fast learning of cognitive control setting from similar trials."

Reference:

Grahek, I., Leng, X., Fahey, M. P., Yee, D., & Shenhav, A. Empirical and Computational Evidence for Reconfiguration Costs During Within-Task Adjustments in Cognitive Control. CogSci.

> 1. *Wouldn't a region that represented each conflict source separately still show the same pattern of results? The degree of Stroop vs Simon conflict is perfectly negatively correlated across conditions, so wouldn't a region that* just *tracks Stoop conflict show these RSA patterns? The authors show that overall congruency is not represented in DLPFC (which is surprising), but they don't break it down by whether this is due to Stroop or Simon congruency (I'm not sure their task allows for this).*

To estimate the unique contributions of the spatial Stroop and Simon conflicts, we performed a model-comparison analysis. We constructed a Stroop-Only model and a Simon-Only model, with each conflict type projected onto the Stroop (vertical) axis or Simon (horizontal) axis, respectively. The similarity between any two conflict types was defined using the Jaccard similarity index (Jaccard, P., 1901), that is, their intersection divided by their union. By replacing the cognitive spacebased conflict similarity regressor with the Stroop-Only and Simon-Only regressors, we calculated their BICs. Results showed that the BIC was larger for Stroop-Only (5377122) and Simon-Only (5377096) than for the Cognitive-Space model (5377094). An additional Stroop+Simon model, including both Stroop-Only and Simon-Only regressors, also showed a poorer model fitting (BIC = 5377118) than the Cognitive-Space model. Considering that the pattern of conflict representations is more manifested when the conflict is present (i.e., on incongruent trials) than not (i.e., on congruent trials), we also conducted the model comparison using the incongruent trials only. Results showed that Stroop-Only (1344128), Simon-Only (1344120), and Stroop+Simon (1344157) models all showed higher BIC values than the CognitiveSpace model (1344104). These results indicate that the right 8C encodes an integrated cognitive space for resolving Stroop and Simon conflicts. Therefore, we believe the cognitive space has incorporated both dimensions. We added these additional analyses and results to the revised manuscript.

"To examine if the right 8C specifically encodes the cognitive space rather than the domain-general or domain-specific organizations, we tested several additional models (see Methods). Model comparison showed a lower BIC in the Cognitive-Space model (BIC = 5377094) than the Domain-General (BIC = 537127) or Domain-Specific (BIC = 537127) models. Further analysis showed the dimensionality of the representation in the right 8C was 1.19, suggesting the cognitive space was close to 1D. We also tested if the observed conflict similarity effect was

driven solely by spatial Stroop or Simon conflicts, and found larger BICs for the models only including the Stroop similarity (i.e., the Stroop-Only model, BIC = 5377122) or Simon similarity (i.e., the Simon-Only model, BIC = 5377096). An additional Stroop+Simon model, including both StroopOnly and Simon-Only regressors, also showed a worse model fitting (BIC = 5377118). Moreover, we replicated the results with only incongruent trials, considering that the pattern of conflict representations is more manifested when the conflict is present (i.e., on incongruent trials) than not (i.e., on congruent trials). We found a poorer fitting in Domain-general (BIC = 1344129), Domain-Specific (BIC = 1344129), Stroop-Only (BIC = 1344128), Simon-Only (BIC = 1344120), and Stroop+Simon (BIC = 1344157) models than the Cognitive-Space model (BIC = 1344104). These results indicate that the right 8C encodes an integrated cognitive space for resolving Stroop and Simon conflicts. The more detailed model comparison results are listed in Table 2."

We reason that we did not observe an overall congruency effect in the RSA results is because our definition of congruency here differed from traditional definitions (i.e., contrast between incongruent and congruent conditions). In the congruency regressor of our RSA model, we defined representational similarity as 1 if calculated between two incongruent, or two congruent trials, and 0 if between incongruent and congruent trials. Thus, our definition of the congruency regressor reflects whether multivariate patterns differ between incongruent and congruent trials, rather than whether activity strengths differ. Indeed, we did observe the latter form of congruency effects, with stronger univariate activities in pre-SMA for incongruent versus congruent conditions. We have added this in the Note S6 ("The multivariate representations of conflict type and orientation are different from the congruency effect"):

"Neither did we observe a multivariate congruency effect (i.e., the pattern difference between incongruent and congruent conditions compared to that within each condition) in the right 8C or any other regions. Note the definition of congruency here differed from traditional definitions (i.e., contrast between activity strength of incongruent and congruent conditions), with which we found stronger univariate activities in pre-SMA for incongruent versus congruent conditions."

We could not determine whether the null effect of the congruency regressor was due to Stroop or Simon congruency alone, because congruency levels of the two types always covary. On all trials of the compound conditions (Conf 2-4), whenever the Stroop dimension was incongruent, the Simon dimension was also incongruent, and vice versa for the congruent condition. Thus, the contribution of spatial Stroop or Simon alone to the congruency effect could not be tested using compound conditions. Although we have pure spatial Stroop or Simon conditions, within-Stroop and withinSimon trial pairs constituted only 8% of cells in the representational similarity matrix. This was insufficient to determine whether the null congruency effect was due to solely Stroop or Simon.

Overall, with the added analysis we found that the data in the right 8C area supports conflict representations that are organized based on both Simon and spatial Stroop conflict. Although the current experimental design does not allow us to identify whether the null effect of the congruency regressor was driven by either conflict or both, we clarified that the congruency regressor did not test the 205 conventional congruency effect and the null finding does not contradict previous 206 research.

Reference:

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat(37), 547-579.

> *1. The authors use a novel form of RSA that concatenates patterns across conditions, runs and subjects into a giant RSA matrix, which is then used for linear mixed effects analysis. This appears to be necessary because conflict type and visual orientation are perfectly confounded within the subject (although, if I understand, the conflict type x congruence interaction wouldn't have the same concern about visual confounds, which shouldn't depend on congruence). This is an interesting approach but should be better justified, preferably with simulations validating the sensitivity and specificity of this method and comparing it to more standard methods.*

The confound exists for both the conflict type and the conflict type × congruence interaction in our design, since both incongruent and congruent conditions include stimuli from the full orientation space. For example, for the spatial Stroop type, the congruent condition could be either an up arrow at the top or a down arrow at the bottom. Similarly, the incongruent condition could be either an up arrow at the bottom or a down arrow at the top. Therefore, both the congruent and incongruent conditions are perfectly confounded with the orientation.

We reanalyzed the data using the well-documented approach by Chen et al. (2017, Neuroimage), as suggested by the reviewer. The new analysis replicated our previously reported results (Fig. 4-5, S4-S7). As Chen et al (2017) has provided abundant simulations to validate this approach, we did not run any further simulations.

> *1. A chief concern is that the same pattern contributes to many entries in the DV, which has been addressed in previous work using row-wise and column-wise random effects (Chen et al., 2017, Neuroimage). It would also be informative to know whether the results hold up to removing within-run similarity, which can bias similarity measures (Walther et al., 2016, Neuroimage).*

Thank you for the comment. In our revised manuscript, we followed your suggestion and adopted the approach proposed by Chen et al. (2017). Specifically, we included both the upper and lower triangle of the representational similarity matrix (excluding the diagonal). Moreover, we also removed all the within-subject similarity (thus also excluding the within-run similarity as suggested by Walther et al. (2016)) to minimize the bias of the potentially strong within-subject similarity. In addition, we added both the row-wise and column-wise random effects to capture the dependence of cells within each column and each row, respectively (Chen et al., 2017).

Results from this approach largely replicated our previous results. The right 8C again showed significant conflict similarity representation, with greater representational strength in incongruent than congruent condition, and positively correlated to behavioral performance. The orientation effect was also identified in the visual (e.g., right V1) and oculomotor (e.g., left FEF) regions.

We have revised the methodology and the results in the revised manuscript:

"Representational similarity analysis (RSA).

For each cortical region, we calculated the Pearson's correlations between fMRI activity patterns for each run and each subject, yielding a 1400 (20 conditions × 2 runs × 35 participants) × 1400 RSM. The correlations were calculated in a cross297 voxel manner using the fMRI activation maps obtained from GLM3 described in the previous section. We excluded within-subject cells from the RSM (thus also excluding the within-run similarity as suggested by Walther et al., (2016)), and the remaining cells were converted into a vector, which was then z-transformed and submitted to a linear mixed effect model as the dependent

variable. The linear mixed effect model also included regressors of conflict similarity and orientation similarity. Importantly, conflict similarity was based on how Simon and spatial Stroop conflict are combined and hence was calculated by first rotating all subject's stimulus location to the top right and bottom-left quadrants, whereas orientation was calculated using original stimulus locations. As a result, the regressors representing conflict similarity and orientation similarity were de-correlated. Similarity between two conditions was measured as the cosine value of the angular difference. Other regressors included a target similarity regressor (i.e., whether the arrow directions were identical), a response similarity regressor (i.e., whether the correct responses were identical); a spatial Stroop distractor regressor (i.e., vertical distance between two stimulus locations); a Simon distractor regressor (i.e., horizontal distance between two stimulus locations). Additionally, we also included a regressor denoting the similarity of Group (i.e., whether two conditions are within the same subject group, according to the stimulus-response mapping). We also added two regressors including ROI316 mean fMRI activations for each condition of the pair to remove the possible uni-voxel influence on the RSM. A last term was the intercept. To control the artefact due to dependence of the correlation pairs sharing the same subject, we included crossed random effects (i.e., row-wise and column-wise random effects) for the intercept, conflict similarity, orientation and the group factors (G. Chen et al., 2017)."

Reference:

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

> *1. Another concern is the extent to which across-subject similarity will only capture consistent patterns across people, making this analysis very similar to a traditional univariate analysis (and unlike the traditional use of RSA to capture subject-specific patterns).*

With proper normalization, we assume voxels across different subjects should show some consistent localizations, although individual differences can be high. J. Chen et al. (2017) has demonstrated that consistent multi-voxel activation patterns exist across individuals. Previous studies have also successfully applied cross-subject RSA (see review by Freund et al, 2021) and cross-subject decoding approaches (e.g., Jiang et al., 2016; Tusche et al., 2016), so we believe cross-subject RSA should be feasible to capture distributed activation patterns shared at the group level. We added this argument in the revised manuscript:

"Previous studies (e.g., J. Chen et al., 2017) have demonstrated that consistent multivoxel activation patterns exist across individuals, and successful applications of cross-subject RSA (see review by Freund, Etzel, et al., 2021) and cross-subject decoding approaches (Jiang et al., 2016; Tusche et al., 2016) have also been reported."

In the revised manuscript, we also tested whether the representation in right 8C held for within-subject data. We reasoned that the conflict similarity effects identified by cross-subject RSA should be replicable in within-subject data, although the latter is not able to dissociate the conflict similarity effect from the orientation effect. We performed similar RSA for within-subject RSMs, excluding the within-run cells. We replaced the perfectly confounded factors of conflict similarity and orientation with a common factor called similarity_orientation. Other confounding factor pairs were addressed similarly. Results showed a significant effect of similarity_orientation, $t(13993) = 3.270$, $p = .0005$, 1-tailed. Given the specific representation of conflict similarity identified by the cross-subject RSA, we believe that the within-subject data of right 8C probably showed similar conflict similarity modulation effects as the cross-subject data, although future research that orthogonalizes conflict type and orientation is needed to fully answer this question. We added this result in the revised section Note S7.

"Note S7. The cross-subject RSA captures similar effects with the within-subject RSA Considering the variability in voxel-level functional localizations among individuals, one may question whether the cross-subject RSA results were biased by the consistent multi-voxel patterns across subjects, distinct from the more commonly utilized withinsubject RSA. We reasoned that the cross-subject RSA should have captured similar effects as the within-subject RSA if we observe the conflict similarity effect in right 8C with the latter analysis. Therefore, we tested whether the representation in right 8C held for within-subject data. Specifically, we performed similar RSA for withinsubject RSMs, excluding the within-run cells. We replaced the perfectly confounded factors of conflict similarity and orientation with a common factor called similarity_orientation. Other confounding factor pairs (i.e., target versus response, and Stroop distractor versus Simon distractor) were addressed similarly. Results showed a significant effect of similarity_orientation, t(13993) = 3.270, p = .0005, 1tailed. Given the specific representation of conflict similarity identified by the crosssubject RSA, the within-subject data of right 8C may show similar conflict similarity modulation effects as the cross-subject data. Further research is needed to fully dissociate the representation of conflict and the representation of visual features such as orientation."

Reference:

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. Nature Neuroscience, 20(1), 115-125.

Freund, M. C., Etzel, J. A., & Braver, T. S. (2021). Neural Coding of Cognitive Control: The Representational Similarity Analysis Approach. Trends in Cognitive Sciences, 25(7), 622-638.

Jiang, J., Summerfield, C., & Egner, T. (2016). Visual Prediction Error Spreads Across Object Features in Human Visual Cortex. J Neurosci, 36(50), 12746-12763.

Tusche, A., Bockler, A., Kanske, P., Trautwein, F. M., & Singer, T. (2016). Decoding the Charitable Brain: Empathy, Perspective Taking, and Attention Shifts Differentially Predict Altruistic Giving. Journal of Neuroscience, 36(17), 4719-4732.

> *1. Finally, the authors should confirm all their results are robust to less liberal methods of multiplicity correction. For univariate analysis, they should report the effects from the standard p < .001 cluster forming threshold for univariate analysis (or TFCE). For multivariate analyses, FDR can be quite liberal. The authors should consider whether their mixed-effects analyses allow for group-level randomization, and consider (relatively powerful) Max-Stat randomization tests (Nichols & Holmes, 2002, Hum Brain Mapp).*

In our revised manuscript, we have corrected the univariate results using the probabilistic TFCE (pTFCE) approach by Spisak et al. (2019). This approach estimates the conditional probability of cluster extent based on Bayes' rule. Specifically, we applied pTFCE on our univariate results (i.e., the z-maps of our contrasts). This returned enhanced Z-score maps, which were then thresholded based on simulated cluster size thresholds using 3dClustSim. A cluster-forming threshold of p < .001 was employed. Results showed only the pre-SMA was activated in the incongruent > congruent contrast, and right IPS and right dmPFC were activated in the linear Simon modulation effect. Further tests also showed these regions were not correlated with the behavioral performance, uncorrected ps >.28. These results largely replicated our previous results. We have revised the method and results accordingly.

Methods:

"Results were corrected with the probabilistic threshold-free cluster enhancement(pTFCE) and then thresholded by 3dClustSim function in AFNI (Cox & Hyde, 1997) with voxel-wise p <

.001 and cluster-wize p < .05, both 1-tailed."

Results:

"In the fMRI analysis, we first replicated the classic congruency effect by searching for brain regions showing higher univariate activation in incongruent than congruent conditions (GLM1, see Methods). Consistent with the literature (Botvinick et al., 2004; Fu et al., 2022), this effect was observed in the pre-supplementary motor area (preSMA) (Fig. 3, Table S1). We then tested the encoding of conflict type as a cognitive space by identifying brain regions with activation levels parametrically covarying with the coordinates (i.e., axial angle relative to the horizontal axis) in the hypothesized cognitive space. As shown in Fig. 1B, change in the angle corresponds to change in spatial Stroop and Simon conflicts in opposite directions. Accordingly, we found the right inferior parietal sulcus (IPS) and the right dorsomedial prefrontal cortex (dmPFC) displayed positive correlation between fMRI activation and the Simon conflict (Fig. 3, Fig. S3, Table S1)."

We appreciate the reviewer's suggestion to apply the Max-Stat randomization tests (Nichols & Holmes, 2002) for the multivariate analyses. However, the representational similarity matrix was too large (1400×1400) to be tested with a balanced randomization approach (i.e., the Max-Stat), due to (1) running even 1000 times for all ROIs cost very long time; (2) the distribution generated from normal times of randomization (e.g., 5000 iterations) would probably be unbalanced, since the full range of possible samples that could be generated by a complete randomization is not adequately represented. Instead, we adopted a very strict Bonferroni correction p < 0.0001/360 when reporting the regression results from RSA. Notebally, Chen et al (2017) has shown that their approach could control the FDR at an acceptable level.

Reference:

Spisák, T., Spisák, Z., Zunhammer, M., Bingel, U., Smith, S., Nichols, T., & Kincses,T. (2019). Probabilistic TFCE: A generalized combination of cluster size and voxel intensity to increase statistical power. NeuroImage, 185, 12-26.

Chen, G., Taylor, P. A., Shin, Y.-W., Reynolds, R. C., & Cox, R. W. J. N. (2017). Untangling the relatedness among correlations, Part II: Inter-subject correlation group analysis through linear mixed-effects modeling. 147, 825-840.

> *Minor concerns:*
>
> 1. *I appreciate the authors wanting to present the conditions in a theory-agnostic way, but the framing of 5 conflict types was confusing. I think framing the conditions as a mixture of 2 conflict types (Stroop and Simon) makes more sense, especially given the previous work on MSIT.*

We have renamed the Type1-5 as spatial Stroop, StHSmL, StMSmM, StLSmH, and Simon conditions, respectively. H, L, and M indicate high, low andmedium similarity with the corresponding conflict, respectively. This is alsoconsistent with the naming of our previous work (Yang et al., 2021).

Reference:

Yang, G., Xu, H., Li, Z., Nan, W., Wu, H., Li, Q., & Liu, X. (2021). The congruency sequence effect is modulated by the similarity of conflicts. Journal of Experimental Psychology: Learning, Memory, and Cognition, 47(10), 1705-1719.

> *1. It would be helpful to have more scaffolding for the key conflict & orientation analyses. A schematic in the main text that outlines these contrasts would be very helpful (e.g. similar to S4).*

We have inserted Figure 7 in the revised manuscript. In this figure, we plotted the schematic of the difference between the conflict similarity 467 and orientation regressors according to their cross-group representational similarity 468 matrices.

> *1. Figure 4D could be clearer, both in labeling and figure caption. 'Modeled similarity' could be relabelled to something more informative, like 'conflict type (or mixture) similarity'. Alternatively, it would be helpful to show a summary RDM for region r-8C. For example, breaking it down by just conflict type and congruence.*

We have relabeled the x-axis to “Conflict type similarity” and y-axis to “Neural similarity” for Figure 4D in the revised manuscript.

We have also added a summary RSM figure in Fig. S5 to show the different similarity patterns between incongruent and congruent conditions.

> *1. It may be helpful to connect your work to how people have discussed multiple forms of conflict monitoring and control with respect to target and distractor features e.g., Lindsay & Jacoby, 1994, JEP:HPP; Mante, Sussillo et al., 2013, Nature; Soutschek et al., 2015, JoCN; Jackson et al., 2021, Comm Bio; Ritz & Shenhav, 2022, bioRxiv*

We have added an analysis to examine how cognitive control modulates target and distractor representation. To this end, we selected the left V4, a visual region showing joint representation of target, Stroop distractor and Simon distractor, as the region of interest. We tested whether these representation strengths differed between incongruent and congruent conditions, finding the representation of target was stronger and representations of both distractors were weaker in the incongruent condition. This suggests that cognitive control modulates the stimuli in both directions. We added the results in Note S10 and Fig. S8, and also added discussion of it in “Methodological implications”.

“Note S10. Cognitive control enhances target representation and suppresses distractor representation Using the separability of confounding factors afforded by the cross-subject RSA, we examined how representations of targets and distractors are modulated by cognitive control. The key assumption is that exerting cognitive control may enhance target representation and suppress distractor representation. We hypothesized that stimuli are represented in visual areas, so we chose a visual ROI from the main RSA results showing joint representation of target, spatial Stroop distractor and Simon distractor ($p < .005$, 1-tail, uncorrected). Only the left V4 met this criterion. We then tested representations with models similar to the main text for incongruent only trials, congruent only trials, and the incongruent – congruent contrast. The contrast model additionally used interaction between the congruency and target, Stroop distractor and Simon distractor terms. Results showed that in the incongruent condition, when we employ more cognitive control, the target representation was enhanced ($t(237990) = 2.59$, $p = .029$, Bonferroni corrected) and both spatial Stroop ($t(237990) = –4.18$, $p < .001$, Bonferroni corrected) and Simon ($t(237990) = –3.14$, $p = .005$, Bonferroni corrected) distractor representations were suppressed (Fig. S8). These are consistent with the idea that the top-down control modulates the stimuli in both directions (Polk et al., 2008; Ritz & Shenhav, 2022).”

Discussion:

"Moreover, the cross-subject RSA provides high sensitivity to the variables of interest and the ability to separate confounding factors. For instance, in addition to dissociating conflict type from orientation, we dissociated target from response, and spatial Stroop distractor from Simon distractor. We further showed cognitive control can both enhance the target representation and suppress the distractor representation (Note S10, Fig. S8), which is in line with previous studies (Polk et al., 2008; Ritz & Shenhav, 2022)."

> *1. For future work, I would recommend placing stimuli along the whole circumference, to orthogonalize Stroop and Simon conflict within-subject.*

We thank the reviewer for this highly helpful suggestion. Expanding the 547 conflict conditions to a full conflict space and replicating our current results could 548 provide stronger evidence for the cognitive space view.

In the revised manuscript, we added this as a possible future design:

"A possible improvement to our current design would be to include left, right, up, and down arrows presented in a grid formation across four spatially separate quadrants, with each arrow mapped to its own response button. However, one potential confounding factor would be that these conditions have different levels of difficulty (i.e., different magnitude of conflict), which may affect the CSE results and their representational similarity."

> ### *Reviewer #2:*
>
> *Summary, general appraisal*
>
> *This study examines the construct of "cognitive spaces" as they relate to neural coding schemes present in response conflict tasks. The authors utilize a novel paradigm, in which subjects must map the direction of a vertically oriented arrow to either a left or right response. Different types of conflict (spatial Stroop, Simon) are parametrically manipulated by varying the spatial location of the arrow (a taskirrelevant feature). The vertical eccentricity of the arrow either agrees or conflicts with the arrow's direction (spatial Stroop), while the horizontal eccentricity of the arrow agrees or conflicts with the side of the response (Simon). A neural coding model is postulated in which the stimuli are embedded in a cognitive space, organized by distances that depend only on the similarity of congruency types (i.e., where conditions with similar relative proportions of spatial-Stroop versus Simon congruency are represented with similar activity patterns). The authors conduct a behavioral and fMRI study to provide evidence for such a representational coding scheme. The behavioral findings replicate the authors' prior work in demonstrating that conflict-related cognitive control adjustments (the congruency sequence effect) shows strong modulation as a function of the similarity between conflict types. With the fMRI neural activity data, the authors report univariate analyses that identified activation in left prefrontal and dorsomedial frontal cortex modulated by the amount of Stroop or Simon conflict present, and multivariate representational similarity analyses (RSA) that identified right lateral prefrontal activity encoding conflict similarity and correlated with the behavioral effects of conflict similarity.*
>
> *This study tackles an important question regarding how distinct types of conflict, which have been previously shown to elicit independent forms of cognitive control adjustments, might be encoded in the brain within a computationally efficient representational format. The ideas postulated by the authors are interesting ones and the utilized methods are rigorous.*

We would like to express our sincere appreciation for the reviewer's positive evaluation of our manuscript and the constructive comments and suggestions. Through careful consideration of your feedback, we have endeavored to make our manuscript more accessible to readers and further strengthened our findings. In response to your suggestion, we reanalyzed our data with the approach proposed by Chen et al.'s (2017, NeuroImage). This reanalysis largely replicated our previous results, reinforcing the validity of our findings. Additionally, we conducted tests with several alternative models and found that the cognitive space hypothesis best aligns with our observed data. We have incorporated these revisions and additional analyses into the manuscript based on your valuable feedback. As a result, we believe that these changes and additional analyses have significantly enhanced the quality of our manuscript. We have provided detailed responses to your comments below.

> *However, the study has critical limitations that are due to a lack of clarity regarding theoretical hypotheses, serious confounds in the experimental design, and a highly non-standard (and problematic) approach to RSA. Without addressing these issues it is hard to evaluate the contribution of the authors findings to the computational cognitive neuroscience literature.*

1. *The primary theoretical question and its implications are unclear. The paper would greatly benefit from more clearly specifying potential alternative hypotheses and discussing their implications. Consider, for example, the case of parallel conflict monitors. Say that these conflict monitors are separately tuned for Stroop and Simon conflict, and are located within adjacent patches of cortex that are both contained within a single cortical parcel (e.g., as defined by the Glasser atlas used by the authors for analyses). If RSA was conducted on the responses of such a parcel to this task, it seems highly likely that an activation similarity matrix would be observed that is quite similar (if not identical) to the hypothesized one displayed in Figure 1. Yet it would seem like the authors are arguing that the "cognitive space" representation is qualitatively and conceptually distinct from the "parallel monitor" coding scheme. Thus, it seems that the task and analytic approach is not sufficient to disambiguate these different types of coding schemes or neural architectures.*

   The authors also discuss a fully domain-general conflict monitor, in which different forms of conflict are encoded within a single dimension. Yet this alternative hypothesis is also not explicitly tested nor discussed in detail. It seems that the experiment was designed to orthogonalize the "domain-general" model from the "cognitive space" model, by attempting to keep the overall conflict uniform across the different stimuli (i.e., in the design, the level of Stroop congruency parametrically trades off with the level of Simon congruency). But in the behavioral results (Fig. S1), the interference effects were found to peak when both Stroop and Simon congruency are present (i.e., Conf 3 and 4), suggesting that the "domain-general" model may not be orthogonal to the "cognitive space" model. One of the key advantages of RSA is that it provides the ability to explicitly formulate, test and compare different coding models to determine which best accounts for the pattern of data. Thus, it would seem critical for the authors to set up the design and analyses so that an explicit model comparison analysis could be conducted, contrasting the domain-general, domain-specific, and cognitive space accounts.

We appreciate the reviewer pointing out the need to formally test alternative models. In the revised manuscript, we have added and compared a few alternative models, finding the Cognitive-Space model (the one with graded conflict similarity levels as we reported) provided the best fit to our data. Specifically, we tested the following five models against the Cognitive-Space model:

(1) Domain-General model. This model treats each conflict type as equivalent, so each two conflict types only differ in the magnitude of their conflict. Therefore, we defined the domain-general matrix as the difference in their effects indexed by the group-averaged RT in Experiment 2. Then the z-scored model vector was sign-flipped to reflect similarity instead of distance. This model showed non-significant conflict type effects ($t(951989) = 0.92$, $p = .179$) and poorer fit (BIC = 5377126) than the Cognitive-Space model (BIC = 5377094).

(2) Domain-Specific model. This model treats each conflict type differently, so we used a diagonal matrix, with within-conflict type similarities being 1 and all crossconflict type similarities being 0. This model also showed non-significant effects ($t(951989) = 0.84$, $p = .201$) and poorer fit (BIC = 5377127) than the Cognitive-Space model.

(3) Stroop-Only model. This model assumes that the right 8C only encodes the spatial Stroop conflict. We projected each conflict type to the Stroop (vertical) axis and calculated the similarity between any two conflict types as the Jaccard similarity index (Jaccard, 1901), that is, their intersection divided by their union. This model also showed non-significant effects ($t(951989) = 0.20$, $p = .423$) and poorer fit (BIC = 5377122) than the Cognitive-Space model.

(4) Simon-Only model. This model assumes that the right 8C only encodes the Simon conflict. We projected each conflict type to the Simon (horizontal) axis and calculated the similarity like the Stroop-Only model. This model showed significant effects ($t(951989) = 4.19$, $p < .001$) but still quantitatively poorer fit (BIC = 5377096) than the Cognitive-Space model.

(5) Stroop+Simon model. This model assumes the spatial Stroop and Simon conflicts are parallelly encoded in the brain, similar to the "parallel monitor" hypothesis suggested by the reviewer. It includes both Stroop-Only and Simon-Only regressors. This model showed nonsignificant effect for the Stroop regressor ($t(951988) = 0.06$, $p = .478$) and significant effect for the Simon regressor ($t(951988) = 3.30$, $p < .001$), but poorer fit (BIC = 5377118) than the Cognitive-Space model.

“Moreover, we replicated these results with only incongruent trials (i.e., when conflict is present), considering that the pattern of conflict representations is more manifested when the conflict is present (i.e., on incongruent trials) than not (i.e., on congruent trials). We found a poorer fitting in Domain-general (BIC = 1344129), Domain-Specific (BIC = 1344129), Stroop-Only (BIC = 1344128), Simon-Only (BIC = 1344120), and Stroop+Simon (BIC = 1344157) models than the Cognitive-Space model (BIC = 1344104).”

In summary, these results indicate that the right 8C encodes an integrated cognitive space for resolving Stroop and Simon conflicts. We added the above results to the revised manuscript.

The above analysis approach was added to the method “Model comparison and representational dimensionality”, and the results were added to the “Multivariate patterns of the right dlPFC encodes the conflict similarity” in the revised manuscript.

Methods:

“Model comparison and representational dimensionality To estimate if the right 8C specifically encodes the cognitive space, rather than the domain-general or domain-specific structures, we conducted two more RSAs. We replaced the cognitive space-based conflict similarity matrix in the RSA we reported above (hereafter referred to as the Cognitive-Space model) with one of the alternative model matrices, with all other regressors equal. The domain-general model treats each conflict type as equivalent, so each two conflict types only differ in the magnitude of their conflict. Therefore, we defined the domain-general matrix as the difference in their congruency effects indexed by the group-averaged RT in Experiment 2. Then the zscored model vector was sign-flipped to reflect similarity instead of distance. The

domain-specific model treats each conflict type differently, so we used a diagonal matrix, with within-conflict type similarities being 1 and all cross-conflict type similarities being 0.

Moreover, to examine if the cognitive space is driven solely by the Stroop or Simon conflicts, we tested a spatial Stroop-Only (hereafter referred to as “Stroop-Only”) and a Simon-Only model, with each conflict type projected onto the spatial Stroop (vertical) axis or Simon (horizontal) axis, respectively. The similarity between any two conflict types was defined using the Jaccard similarity index (Jaccard, 1901), that is, their intersection divided by their union. We also included a model assuming the Stroop and Simon dimensions are independently represented in the brain, adding up the StroopOnly and Simon-Only regressors (hereafter referred to as the Stroop+Simon model). We conducted similar RSAs as reported above, replacing the original conflict similarity regressor with the Strrop-Only, Simon-Only, or both regressors (for the Stroop+Simon model), and then calculated their Bayesian information criterions (BICs).”

Results:

“To examine if the right 8C specifically encodes the cognitive space rather than the domain-general or domain-specific organizations, we tested several additional models (see Methods). Model comparison showed a lower BIC in the Cognitive-Space model (BIC = 5377094) than the Domain-General (BIC = 537127) or Domain-Specific (BIC = 537127) models. Further analysis showed the dimensionality of the representation in the right 8C was 1.19, suggesting the cognitive space was close to 1D. We also tested if the observed conflict similarity effect was driven solely by spatial Stroop or Simon conflicts, and found larger BICs for the models only including the Stroop similarity (i.e., the Stroop-Only model, BIC = 5377122) or Simon similarity (i.e., the Simon-Only model, BIC = 5377096). An additional Stroop+Simon model, including both StroopOnly and Simon-Only regressors, also showed a worse model fitting (BIC = 5377118). Moreover, we replicated the results with only incongruent trials, considering that the pattern of conflict representations is more manifested when the conflict is present (i.e., on incongruent trials) than not (i.e., on congruent trials). We found a poorer fitting in Domain-general (BIC = 1344129), Domain-Specific (BIC = 1344129), Stroop-Only (BIC = 1344128), Simon-Only (BIC = 1344120), and Stroop+Simon (BIC = 1344157) models than the Cognitive-Space model (BIC = 1344104). These results indicate that the right 8C encodes an integrated cognitive space for resolving Stroop and Simon conflicts. The more detailed model comparison results are listed in Table 2.”

Reference:

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat(37), 547-579.

> *2a) Relatedly, the reasoning for the use of the term "cognitive space" is unclear. The mere presence of graded coding for two types of conflict seems to be a low bar for referring to neural activity patterns as encoding a "cognitive space". It is discussed that cognitive spaces/maps allow for flexibility through inference and generalization. But no links were made between these cognitive abilities and the observed representational structure.*

In the revised manuscript, we have clarified that we tested a specific prediction of the cognitive space hypothesis: the geometry of the cognitive space predicts that more similar conflict types will have more similar neural representations,leading to the CSE and RSA patterns tested in this study. These results add to the literature by providing empirical evidence on how different conflict types are encoded in the brain. We agree that this study is not a comprehensive test of the cognitive space hypothesis. Thus, in the revised manuscript we explicitly clarified that this study is a test of the geometry of the cognitive space hypothesis.

Critically, the cognitive space view holds that the representations of different abstract information are organized continuously and the representational geometry in the cognitive space are determined by the similarity among the represented information (Bellmund et al., 2018).

"The present study aimed to test the geometry of cognitive space in conflict representation. Specifically, we hypothesize that different types of conflict are represented as points in a cognitive space. Importantly, the distance between the points, which reflects the geometry of the cognitive space, scales with the difference in the sources of the conflicts being represented by the points."

We have also discussed the limitation of the results and stressed the need for more research to fully test the cognitive space hypothesis.

“Additionally, our study is not a comprehensive test of the cognitive space hypothesis but aimed primarily to provide original evidence for the geometry of cognitive space in representing conflict information in cognitive control. Future research should examine other aspects of the cognitive space such as its dimensionality, its applicability to other conflict tasks such as Eriksen Flanker task, and its relevance to other cognitive abilities, such as cognitive flexibility and learning.

> *2b) Additionally, no explicit tests of generality (e.g., via cross-condition generalization) were provided.*

To examine the generality of cognitive space across conditions, we conducted a leave-one-out prediction analysis. We used the behavioral data from Experiment 1 for this test, due to its larger amount of data than Experiment 2. Specifically, we removed data from one of the five similarity levels (as illustrated by the θs in Fig. 1C) and used the remaining data to perform the same mixed-effect model as reported in the main text (i.e., the two-stage analysis). This yielded one pair of beta coefficients including the similarity regressor and the intercept for each subject, with which we predicted the CSE for the removed similarity level for each subject. We repeated this process for each similarity level once. The predicted results were highly correlated with the original data, with r = .87 for the RT and r = .84 for the ER, ps < .001. We have added this analysis and result to the “Conflict type 706 similarity modulated behavioral congruency sequence effect (CSE)” section.

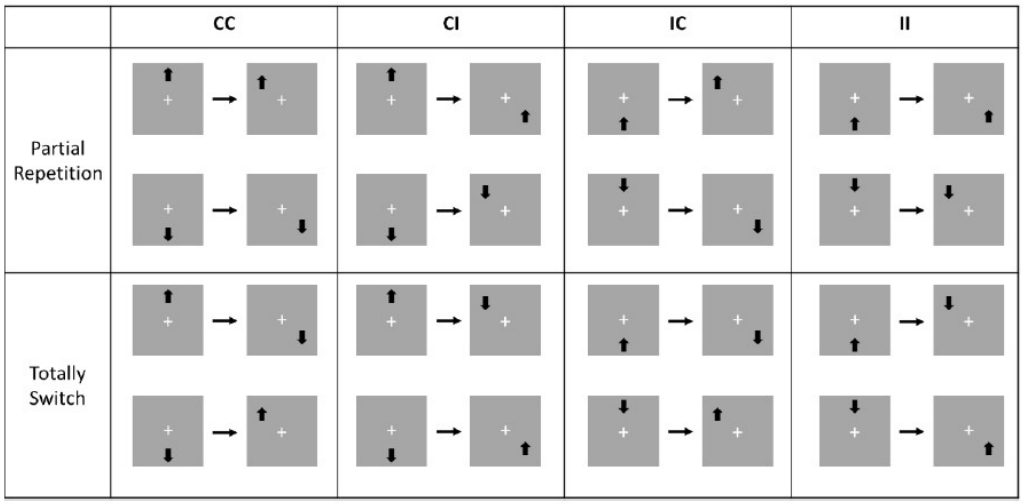
“Moreover, to test the continuity and generalizability of the similarity modulation, we conducted a leave-one-out prediction analysis. Specifically, we removed data from one of the five similarity levels (as illustrated by the θs in Fig. 1C) and used the remaining data to perform the same mixed-effect model (i.e., the two-stage analysis). This yielded one pair of beta coefficients including the similarity regressor and the intercept for each subject, with which we predicted the CSE for the removed similarity level for each subject. We repeated this process for each similarity level once. The predicted results were highly correlated with the original data, with r = .87 for the RT and r = .84 for the ER, ps < .001."

> *2c) Finally, although the design elicits strong CSE effects, it seems somewhat awkward to consider CSE behavioral patterns as a reflection of the kind of abilities supported by a cognitive map (if this is indeed the implication that was intended). In fact, CSE effects are well-modeled by simpler "model-free" associative learning processes, that do not require elaborate representations of abstract structures.*

We argue the conflict similarity modulation of CSEs we observed cannot be explained by the “model-free” stimulus-driven associative learning process. This mainly refers to the feature integration account proposed by Hommel et al. (2004), which explains poorer performance in CI and IC trials (compared with CC and II trials) with the partial repetition cost caused by the

breaking of stimulus-response binding. Although we cannot remove its influence on the within-type trials (similarity level 5, θ = 0), it should not affect the cross-type trials (similarity level 1-4, θ = 90°, 67.5°, 45° and 22.5°, respectively), because the CC, CI, IC, II trials had equal probabilities of partially repeated and fully switched trials (see the Author response image 1 for an example of trials across Conf 1 and Conf 3 conditions). Thus, feature integration cannot explain the gradual CSE decrease from similarity level 1 to 4, which sufficiently reproduce the full effect, as suggested by the leave-one-out prediction analysis mentioned above. We thus conclude that the similarity modulation of CSE cannot be explained by the stimulus-driven associative learning.

Author response image 1.



Notably, however, our findings are aligned with an associative learning account of cognitive control (Abrahamse et al., 2016), which extends association learning from stimulus/response level to cognitive control. In other words, abstract cognitive control state can be learned and generalized like other sensorimotor features. This view explicitly proposes that "transfer occurs to the extent that two tasks overlap", a hypothesis directly supported by our CSE results (see also Yang et al., 2021). Extending this, our fMRI results provide the neural basis of how cognitive control can generalize through a representation of cognitive space. The cognitive space view complements associative learning account by providing a fundamental principle for the learning and generalization of control states. Given the widespread application of CSE as indicator of cognitive control generalization (Braem et al., 2014), we believe that it can be recognized as a kind of ability supported by the cognitive space. This was further supported by the brain-behavioral correlation: stronger encoding of cognitive space was associated with greater bias of trial-wise behavioral adjustment by the consecutive conflict similarity.

We have incorporated these ideas into the discussion:

"Similarly, we propose that cognitive space could serve as a mental model to assist fast learning and efficient organization of cognitive control settings. Specifically, the cognitive space representation may provide a principle for how our brain evaluates the expected cost of switching and the benefit of generalization between states and selects the path with the best cost-benefit tradeoff (Abrahamse et al., 2016; Shenhav et al., 2013). The proximity between two states in cognitive space could reflect both the expected cognitive demand required to transition and the useful mechanisms to adapt from. The closer the two conditions are in cognitive space, the lower the expected switching cost and the higher the

generalizability when transitioning between them. With the organization of a cognitive space, a new conflict can be quickly assigned a location in the cognitive space, which will facilitate the development of cognitive control settings for this conflict by interpolating nearby conflicts and/or projecting the location to axes representing different cognitive control processes, thus leading to a stronger CSE when following a more similar conflict condition."

References:

Hommel, B., Proctor, R. W., & Vu, K. P. (2004). A feature-integration account of sequential effects in the Simon task. Psychological Research, 68(1), 1-17. Abrahamse, E., Braem, S., Notebaert, W., & Verguts, T. (2016). Grounding cognitive control in associative learning. Psychological Bulletin, 142(7), 693-728.

Yang, G., Xu, H., Li, Z., Nan, W., Wu, H., Li, Q., & Liu, X. (2021). The congruency sequence effect is modulated by the similarity of conflicts. Journal of 770 Experimental Psychology: Learning, Memory, and Cognition, 47(10), 1705-1719.

Braem, S., Abrahamse, E. L., Duthoo, W., & Notebaert, W. (2014). What determines the specificity of conflict adaptation? A review, critical analysis, and proposed synthesis. Frontiers in Psychology, 5, 1134.

> 1. *More generally, it seems problematic that Stroop and Simon conflict in the paradigm parametrically trade-off against each other. A more powerful design would have de-confounded Stroop and Simon conflict so that each could be separately estimation via (potentially orthogonal) conflict axes. Additionally, incorporating more varied stimulus sets, locations, or responses might have enabled various tests of generality, as implied by a cognitive space account.*

We thank the reviewer for these valuable suggestions. We argue that the current design is adequate to test the prediction that more similar conflict types have more similar neural representations. That said, we agree that further examination using more powerful experimental designs are needed to fully test the cognitive space account of cognitive control. We also agree that employing more varied stimulus sets,locations and responses would further extend our findings. We have included this as a future research direction in the revised manuscript.

We have revised our discussion about the limitation as:

"A few limitations of this study need to be noted. To parametrically manipulate the conflict similarity levels, we adopted the spatial Stroop-Simon paradigm that enables parametrical combinations of spatial Stroop and Simon conflicts. However, since this paradigm is a two-alternative forced choice design, the behavioral CSE is not a pure measure of adjusted control but could be partly confounded by bottom-up factors such as feature integration (Hommel et al., 2004). Future studies may replicate our findings with a multiple-choice design (including more varied stimulus sets, locations and responses) with confound-free trial sequences (Braem et al., 2019). Another limitation is that in our design, the spatial Stroop and Simon effects are highly anticorrelated. This constraint may make the five conflict types represented in a unidimensional space (e.g., a circle) embedded in a 2D space. Future studies may test the 2D cognitive space with fully independent conditions. A possible improvement to our current design would be to include left, right, up, and down arrows presented in a grid formation across four spatially separate quadrants, with each arrow mapped to its own response button. However, one potential confounding factor would be that these conditions have different levels of difficulty (i.e., different magnitude of conflict), which may affect the CSE results and their representational similarity."

1. *Serious confounds in the design render the results difficult to interpret. As much prior neuroimaging and behavioral work has established, "conflict" per se is perniciously correlated with many conceptually different variables. Consequently, it is very difficult to distinguish these confounding variables within aggregate measures of neural activity like fMRI. For example, conflict is confounded with increased time-on-task with longer RT, as well as conflict-driven increases in coding of other task variables (e.g., task-set related coding; e.g., Ebitz et al. 2020 bioRxiv). Even when using much higher resolution invasive measures than fMRI (i.e., eCoG), researchers have rightly been wary of making strong conclusions about explicit encoding of conflict (Tang et al, 2019; eLife). As such, the researchers would do well to be quite cautious and conservative in their analytic approach and interpretation of results.*

We acknowledge the findings showing that encoding of conflicts may not be easily detected in the brain. However, recent studies have shown that the representational similarity analysis can effectively detect representations of conflict tasks (e.g., the color Stroop) using factorial designs (Freund et al., 2021a; 2021b).

In our analysis, we are aware of the potential impact of time-on-task (e.g., RT) on univariate activation levels and subsequent RSA patterns. To address this issue, we added univariate fMRI activation levels as nuisance regressors to the RSA. To de confound conflict from other factors such as orientation of stimuli related to the center of the screen, we also applied the cross-subject RSA approach. Furthermore, we were cautious about determining regions that encoded conflict control. We set three strict criteria: (1) Regions must show a conflict similarity modulation effect; (2) regions must show higher representational strength in the incongruent condition compared with the congruent condition; and (3) regions must correlate with behavioral performance. With these criteria, we believe that the results we reported are already conservative. We would be happy to implement any additional criteria the reviewer recommends.

Reference:

Freund, M. C., Etzel, J. A., & Braver, T. S. (2021a). Neural Coding of Cognitive Control: The Representational Similarity Analysis Approach. Trends in Cognitive Sciences, 25(7), 622-638.

Freund, M. C., Bugg, J. M., & Braver, T. S. (2021b). A Representational Similarity 823 Analysis of Cognitive Control during Color-Word Stroop. Journal of 824 Neuroscience, 41(35), 7388-7402.

1. *This issue is most critical in the interpretation of the fMRI results as reflecting encoding of conflict types. A key limitation of the design, that is acknowledged by the authors is that conflict is fully confounded within-subject by spatial orientation. Indeed, the limited set of stimulus-response mappings also cast doubt on the underlying factors that give rise to the CSE modulations observed by the authors in their behavioral results. The CSE modulations are so strong - going from a complete absence of current x previous trial-type interaction in the cos(90) case all the way to a complete elimination of any current trial conflict when the prior trial was incongruent in the cos(0) case - that they cause suspicion that they are actually driven by conflict-related control adjustments rather than sequential dependencies in the stimulus-response mappings that can be associatively learned.*

Unlike the fMRI data, we cannot tease apart the effects of conflict similarity and orientation in a similar manner as the cross-subject RSA for behavioral CSEs. However, we have a few reasons that the orientation and other bottom-up factors should not be the factors driving the similarity modulation effect.

First, we did not find any correlation between the regions showing orientation effects and behavioral CSEs. This suggests that orientation does not directly contribute to the CSE modulation.

Second, if the CSE modulation is purely driven by the association learning of the stimulus-response mapping, we should observe a stronger modulation effect after more extensive training. However, our results do not support this prediction. Using data from Experiment 1, we found that the modulation effect remained constant across the three sessions (see Note S3).

“Note S3. Modulation of conflict similarity on behavioral CSEs does not change across time We tested if the conflict similarity modulation on the CSE is susceptible to training. We collected the data of Experiment 1 across three sessions, thus it is possible to examine if the conflict similarity modulation effect changes across time. To this end, we added conflict similarity, session and their interaction into a mixed-effect linear model, in which the session was set as a categorical variable. With a post-hoc analysis of variance (ANOVA), we calculated the statistical significance of the interaction term. This approach was applied to both the RT and ER. Results showed no interaction effect in either RT, $F(2,1479) = 1.025$, $p = .359$, or ER, $F(2,1479) = 0.789$, $p = .455$. This result suggests that the modulation effect does not change across time. “

Third, the observed similarity modulation on the CSE, particularly for similarity levels 1-4, should not be attributed to the stimulus-response associations, such as feature integration, as have been addressed in response to comment 2.c.

Finally, other bottom-up factors, such as the spatial location proximity did not drive the CSE modulation results, which we have addressed in the original manuscript in Note S2.

"Note S2. Modulation of conflict similarity on behavioral CSEs cannot be explained by the physical proximity

In our design, the conflict similarity might be confounded by the physical proximity between stimulus (i.e., the arrow) of two consecutive trials. That is, when arrows of the two trials appear at the same quadrant, a higher conflict similarity also indicates a higher physical proximity (Fig. 1A). Although the opposite is true if arrows of the two trials appear at different quadrants, it is possible the behavioral effects can be biased by the within quadrant trials. To examine if the physical distance has confounded the conflict similarity modulation effect, we conducted an additional analysis.

We defined the physical angular difference across two trials as the difference of their polar angles relative to the origin. Therefore, the physical angular difference could vary from 0 to 180°. For each CSE conditions (i.e., CC, CI, IC and II), we grouped the trials based on their physical angular distances, and then averaged trials with the same previous by current conflict type transition but different orders (e.g., StHSmL−StLSmH and StLSmH−StHSmL) within each subject. The data were submitted to a mixed-effect model with the conflict similarity, physical proximity (i.e., the opposite of the physical angular difference) as fixed-effect predictors, and subject and CSE condition as random effects. Results showed significant conflict similarity modulation effects in both Experiment 1 (RT: $\beta = 0.09 \pm 0.01$, $t(7812) = 13.74$, $p < .001$, $\eta p2 = .025$; 875 ER: $\beta = 0.09 \pm 0.01$, $t(7812) = 7.66$, $p < .001$, $\eta p2 = .018$) and Experiment 2 (RT: $\beta = 876\ 0.21 \pm 0.02$, $t(3956) = 9.88$, $p < .001$, $\eta p2 = .043$; ER: $\beta = 0.20 \pm 0.03$, $t(4201) = 6.11$, 877 $p < .001$, $\eta p2 = .038$). Thus, the observed modulation of conflict similarity on behavioral 878 CSEs cannot be explained by physical proximity."

> *1. To their credit, the authors recognize this confound, and attempt to address it analytically through the use of a between-subject RSA approach. Yet the solution is itself problematic, because it doesn't actually deconfound conflict from orientation. In particular, the RSA model assumes that whatever components of neural activity encode orientation produce this encoding within the same voxellevel patterns of activity in each subject. If they are not (which is of course likely), then orthogonalization of these variables will be incomplete. Similar issues underlie the interpretation target/response and distractor coding. Given these issues, perhaps zooming out to a larger spatial scale for the between-subject RSA might be warranted. Perhaps whole-brain at the voxel level with a high degree of smoothing, or even whole-brain at the parcel level (averaging per parcel). For this purpose, Schaefer atlas parcels might be more useful than Glasser, as they more strongly reflect functional divisions (e.g., motor strip is split into mouth/hand divisions; visual cortex is split into central/peripheral visual field divisions). Similarly, given the lateralization of stimuli, if a within-parcel RSA is going to be used, it seems quite sensible to pool voxels across hemispheres (so effectively using 180 parcels instead of 360).*

Doing RSA at the whole-brain level is an interesting idea. However, it does not allow the identification of specific brain regions representing the cognitive space. Additionally, increasing the spatial scale would include more voxels that are not involved in representing the information of interest and may increase the noise level of data. Given these concerns, we did not conduct the whole-brain level RSA.

We agree that smoothing data can decrease cross-subject variance in voxel distribution and may increase the signal-noise ratio. We reanalyzed the results for the right 8C region using RSA on smoothed beta maps (6-mm FWHM Gaussian kernel). This yielded a significant conflict similarity effect, $t(951989) = 5.55$, $p < .0001$, replicating the results on unsmoothed data ($t(951989) = 5.60$, $p < .0001$). Therefore, we retained the results from unsmoothed data in the main text, and added the results based on smoothed data to the supplementary material (Note S9).

“Note S9. The cross-subject pattern similarity is robust against individual differences Due to individual differences, the multivoxel patterns extracted from the same brain mask may not reflect exactly the same brain region for each subject. To reduce the influence of individual difference, we conducted the same cross-subject RSA using data smoothed with a 6-mm FWHM Gaussian kernel. Results showed a significant conflict similarity effect, $t(951989) = 5.55$, $p < .0001$, replicating the results on unsmoothed data ($t(951989) = 5.60$, $p < .0001$). “

We also used the bilateral 8C area as a single mask and conducted the same RSA. We found a significant conflict type similarity effect, $t(951989) = 4.36$, $p < .0001$. However, the left 8C alone showed no such representation, $t(951989) = 0.38$, $p = .351$, consistent with the right lateralized representation of cognitive space we reported in Note S8. Therefore, we used ROIs from each hemisphere separately.

“Note S8. The lateralization of conflict type representation

We observed the right 8C but not the left 8C represented the conflict type similarity. A further test is to show if there is a lateralization. We tested several regions of the left dlPFC, including the i6-8, 8Av, 8C, p9-46v, 46, 9-46d, a9-46v (Freund, Bugg, et al., 2021). We found that none of these regions show the representation of conflict type, all uncorrected $ps > .35$. These results indicate that the conflict type is specifically represented in the right dlPFC. “

We have also discussed the lateralization in the manuscript:

"In addition, we found no such representation in the left dlPFC (Note S8), indicating a possible lateralization. Previous studies showed that the left dlPFC was related to the expectancy-related attentional set up-regulation, while the right dlPFC was related to the online adjustment of control (Friehs et al., 2020; Vanderhasselt et al., 2009), which is consistent with our findings. Moreover, the right PFC also represents a composition of single rules (Reverberi et al., 2012), which may explain how the spatial Stroop and Simon types can be jointly encoded in a single space."

> 1. *The strength of the results is difficult to interpret due to the non-standard analysis method. The use of a mixed-level modeling approach to summarize the empirical similarity matrix is an interesting idea, but nevertheless is highly non-standard within RSA neuroimaging methods. More importantly, the way in which it was implemented makes it potentially vulnerable to a high degree of inaccuracy or bias. In this case, this bias is likely to be overly optimistic (high false positive rate). No numerical or formal defense was provided for this mixed-level model approach. As a result, the use of this method seems quite problematic, as it renders the strength of the observed results difficult to interpret. Instead, the authors are encouraged using a previously published method of conducting inference with between-subject RSA, such as the bootstrapping methods illustrated in Kragel et al. (2018; Nat Neurosci), or in potentially adopting one of the Chen et al. methods mentioned above, that have been extensively explored in terms of statistical properties.*
>
> *No numerical or formal defense was provided for this mixed-level model approach. As a result, the use of this method seems quite problematic, as it renders the strength of the observed results difficult to interpret. Instead, the authors are encouraged using a previously published method of conducting inference with between-subject RSA, such as the bootstrapping methods illustrated in Kragel et al. (2018; Nat Neurosci), or in potentially adopting one of the Chen et al. methods mentioned above, that have been extensively explored in terms of statistical properties.*

In our revised manuscript, we have adopted the approach proposed by Chen et al. (2017). Specifically, we included both the upper and lower triangle of the representational similarity matrix (excluding the diagonal). Moreover, we also removed all the within-subject similarity (thus also excluding the within-run similarity) to minimize the bias of the potentially strong within-subject similarity (note we also analyzed the within-subject data and found significant effects for the similarity modulation, though this effect cannot be attributed to the conflict similarity or orientation alone. We added this part in Note S7, see below). In addition, we added both the row-wise and column-wise random effects to capture the dependence of cells within each column/row (Chen et al., 2017). We have revised the method part as:

"We excluded within-subject cells from the RSM (thus also excluding the withinrun similarity as suggested by Walther et al., (2016)), and the remaining cells were converted into a vector, which was then z-transformed and submitted to a linear mixed effect model as the dependent variable. The linear mixed effect model also included regressors of conflict similarity and orientation similarity. Importantly, conflict similarity was based on how Simon and spatial Stroop conflicts are combined and hence was calculated by first rotating all subject's stimulus location to the topright and bottom-left quadrants, whereas orientation was calculated using original stimulus locations. As a result, the regressors representing conflict similarity and orientation similarity were de-correlated. Similarity between two conditions was measured as the cosine value of the angular difference. Other regressors included a target similarity regressor (i.e., whether the arrow directions were identical), a response similarity regressor (i.e., whether the correct responses were identical); a spatial Stroop distractor regressor (i.e., vertical distance between two stimulus locations); a Simon distractor regressor (i.e., horizontal distance between two stimulus locations). Additionally, we also included a

regressor denoting the similarity of Group (i.e., whether two conditions are within the same subject group, according to the stimulus-response mapping). We also added two regressors including ROImean fMRI activations for each condition of the pair to remove the possible univoxel influence on the RSM. A last term was the intercept. To control the artefact due to dependence of the correlation pairs sharing the same subject, we included crossed random effects (i.e., row-wise and column-wise random effects) for the intercept, conflict similarity, orientation and the group factors (G. Chen et al., 2017)."

Results from this approach highly replicated our original results. Specifically, we found the right 8C again showed a strong conflict similarity effect, a higher representational strength in the incongruent condition compared to the congruent condition, and a significant correlation with the behavioral CSE. The orientation effect was also identified in the visual (e.g., right V1) and oculomotor (e.g., left FEF) regions.

We revised the results accordingly:

For the conflict type effect:

"The first criterion revealed several cortical regions encoding the conflict similarity, including the Brodmann 8C area (a subregion of dlPFC(Glasser et al., 2016)) and a47r in the right hemisphere, and the superior frontal language (SFL) area, 6r, 7Am, 24dd, and ventromedial visual area 1 (VMV1) areas in the left hemisphere (Bonferroni corrected ps < 0.0001, one-tailed, Fig. 4A). We next tested whether these regions were related to cognitive control by comparing the strength of conflict similarity effect between incongruent and congruent conditions (criterion 2). Results revealed that the left SFL, left VMV1, and right 8C met this criterion, Bonferroni corrected ps < .05, one-tailed, suggesting that the representation of conflict type was strengthened when conflict was present (e.g., Fig. 4D). The intersubject brain-behavioral correlation analysis (criterion 3) showed that the strength of conflict similarity effect on RSM scaled with the modulation of conflict similarity on the CSE (slope in Fig. S2C) in right 8C (r = .52, Bonferroni corrected p = .002, onetailed, Fig. 4C, Table 1) but not in the left SFL and VMV1 (all Bonferroni corrected ps > .05, one-tailed). "

For the orientation effect:

"We observed increasing fMRI representational similarity between trials with more similar orientations of stimulus location in the occipital cortex, such as right V1, right V2, right V4, and right lateral occipital 2 (LO2) areas (Bonferroni corrected ps < 0.0001). We also found the same effect in the oculomotor related region, i.e., the left 997 frontal eye field (FEF), and other regions including the right 5m, left 31pv and right parietal area F (PF) (Fig. 5A). Then we tested if any of these brain regions were related to the conflict representation by comparing their encoding strength between incongruent and congruent conditions. Results showed that the right V1, right V2, left FEF, and right PF encoded stronger orientation effect in the incongruent than the congruent condition, Bonferroni corrected ps < .05, one-tailed (Table1, Fig. 5B). We then tested if any of these regions was related to the behavioral performance, and results showed that none of them positively correlated with the behavioral conflict similarity modulation effect, all uncorrected ps > .45, one-tailed. Thus all regions are consistent with the criterion 3."

"Note S7. The cross-subject RSA captures similar effects with the within-subject RSA Considering the variability in voxel-level functional localizations among individuals, one may question whether the cross-subject RSA results were biased by the consistent multi-voxel patterns across subjects, distinct from the more commonly utilized withinsubject RSA. We reasoned that the cross-subject RSA should have captured similar effects as the within-subject RSA if we observe the conflict similarity effect in right 8C with the latter analysis. Therefore, we tested whether the representation in right 8C held for within-subject data. Specifically, we performed similar RSA for withinsubject RSMs, excluding the within-run cells. We replaced

the perfectly confounded factors of conflict similarity and orientation with a common factor called similarity_orientation. Other confounding factor pairs (i.e., target versus response, and Stroop distractor versus Simon distractor) were addressed similarly. Results showed a significant effect of similarity_orientation, t(13993) = 3.270, p = .0005, 1tailed. Given the specific representation of conflict similarity identified by the crosssubject RSA, the within-subject data of right 8C may show similar conflict similarity modulation effects as the cross-subject data. Further research is needed to fully dissociate the representation of conflict and the representation of visual features such as orientation."

> 1. *Another potential source of bias is in treating the subject-level random effect coefficients (as predicted by the mixed-level model) as independent samples from a random variable (in the t-tests). The more standard method for inference would be to use test statistics derived from the mixed-model fixed effects, as those have degrees of freedom calculations that are calibrated based on statistical theory.*

In our revised manuscript, we reported the statistical p values calculated from the mixed-effect models. Note that because we used the Chen et al. (2017) method, which includes data from the symmetric matrix, we corrected the degrees of freedom and estimated the true p values based on the t statistics of model results. For the I versus C comparison results, we calculated the p values by combining I and C RSMs into a larger model and then adding the condition type, as well as the interaction between the regressors of interest (conflict similarity and orientation) and the condition type. We made the statistical inference based on the interaction effect.

We have revised the corresponding methods as:

"The statistical significance of these beta estimates was based on the outputs of the mixed-effect model estimated with the "fitlme" function in Matlab 2022a. Since symmetric cells from the RSM matrix were included in the mixed-effect model, we adjusted the t and p values with the true degree of freedom, which is half of the cells included minus the number of fixed regressors. Multiple comparison correction was applied with the Bonferroni approach across all cortical regions at the $p < 0.0001$ level. To test if the representation strengths are different between congruent and incongruent conditions, we also conducted the RSA using only congruent (RDM_C) and incongruent (RDM_I) trials separately. The contrast analysis was achieved by an additional model with both RDM_C and RDM_I included, adding the congruency and the interaction between conflict type (and orientation) and congruency as both fixed and random factors. The difference between incongruent and congruent representations was indicated by a significant interaction effect."

> ### *Reviewer #3:*
>
> *Yang and colleagues investigated whether information on two task-irrelevant features that induce response conflict is represented in a common cognitive space. To test this, the authors used a task that combines the spatial Stroop conflict and the Simon effect. This task reliably produces a beautiful graded congruency sequence effect (CSE), where the cost of congruency is reduced after incongruent trials. The authors measured fMRI to identify brain regions that represent the graded similarity of conflict types, the congruency of responses, and the visual features that induce conflicts.*
>
> *Using several theory-driven exclusion criteria, the authors identified the right dlPFC (right 8C), which shows 1) stronger encoding of graded similarity of conflicts in incongruent trials and 2) a positive correlation between the strength of conflict similarity type and the CSE on behavior. The dlPFC has been shown to be important for cognitive control tasks. As the dlPFC did not show a univariate parametric modulation based on the higher or lower component of one type of conflict (e.g., having more spatial Stroop conflict or less*

Simon conflict), it implies that dissimilarity of conflicts is represented by a linear increase or decrease of neural responses. Therefore, the similarity of conflict is represented in multivariate neural responses that combine two sources of conflict.

The strength of the current approach lies in the clear effect of parametric modulation of conflict similarity across different conflict types. The authors employed a clever cross-subject RSA that counterbalanced and isolated the targeted effect of conflict similarity, decorrelating orientation similarity of stimulus positions that would otherwise be correlated with conflict similarity. A pattern of neural response seems to exist that maps different types of conflict, where each type is defined by the parametric gradation of the yoked spatial Stroop conflict and the Simon conflict on a similarity scale. The similarity of patterns increases in incongruent trials and is correlated with CSE modulation of behavior.

We would like to thank the reviewer for the positive evaluation of our manuscript and for providing constructive comments. By addressing these comments, we believe that we have made our manuscript more accessible for the readers while also strengthening our findings. In particular, we have tested a few alternative models and confirmed that the cognitive space hypothesis best fits the data. We have also demonstrated the geometric properties of the cognitive space by examining the continuity and dimensionality of the space, further supporting our main arguments. We have incorporated revisions and additional analyses to the manuscript based on your feedback. Overall, we believe that these changes and additional analyses have significantly improved the manuscript. Please find our detailed responses below.

However, several potential caveats need to be considered.

1. *One caveat to consider is that the main claim of recruitment of an organized "cognitive space" for conflict representation is solely supported by the exclusion criteria mentioned earlier. To further support the involvement of organized space in conflict representation, other pieces of evidence need to be considered. One approach could be to test the accuracy of out-of-sample predictions to examine the continuity of the space, as commonly done in studies on representational spaces of sensory information. Another possible approach could involve rigorously testing the geometric properties of space, rather than fitting RSM to all conflict types. For instance, in Fig 6, both the organized and domain-specific cognitive maps would similarly represent the similarity of conflict types expressed in Fig1c (as evident from the preserved order of conflict types). The RSM suggests a low-dimensional embedding of conflict similarity, but the underlying dimension remains unclear.*

Following the reviewer's first suggestion, we conducted a leave-one-out prediction approach to examine the continuity of the cognitive space. We used the behavioral data from Experiment 1 for this test, due to its larger amount of data than Experiment 2. Specifically, we removed data from one of the five similarity levels (as illustrated by the θs in Fig. 1C) and used the remaining data to perform the same mixed-effect model as reported in the main text (i.e., the two-stage analysis). This yielded one pair of beta coefficients including the similarity regressor and the intercept for each subject, with which we predicted the CSE for the removed similarity level at subject level. We repeated this process for each similarity level once. The predicted results were highly correlated with the original data, with $r = .87$ for the RT and $r = .84$ for the ER, $ps < .001$. We have added this analysis and result to the "Conflict type similarity modulated behavioral congruency sequence effect (CSE)" 1079 section:

"Moreover, to test the continuity and generalizability of the similarity modulation, we conducted a leave-one-out prediction analysis. We used the behavioral data from Experiment

1 for this test, due to its larger amount of data than Experiment 2. Specifically, we removed data from one of the five similarity levels (as illustrated by the θs in Fig. 1C) and used the remaining data to perform the same mixed-effect model (i.e., the two-stage analysis). This yielded one pair of beta coefficients including the similarity regressor and the intercept for each subject, with which we predicted the CSE for the removed similarity level for each subject. We repeated this process for each similarity level once. The predicted results were highly correlated with the original data, with r = .87 for the RT and r = .84 for the ER, ps < .001."

To estimate if the domain-specific model could explain the results we observed in right 8C, we conducted a model-comparison analysis. The domain-specific model treats each conflict type differently, so we used a diagonal matrix, with within-conflict type similarities being 1 and all cross-conflict type similarities being 0. This model showed non-significant effects (t(951989) = 0.84, p = .201) and poorer fit (BIC = 5377127) than the cognitive space model (t(951989) = 5.60, p = 1.1×10−8, BIC = 5377094). We also compared other alternative models and found the cognitive space model best fitted the data. We have included these results in the revised manuscript:

"To examine if the right 8C specifically encodes the cognitive space rather than the domain-general or domain-specific organizations, we tested several additional models (see Methods). Model comparison showed a lower BIC in the Cognitive-Space model (BIC = 5377094) than the Domain-General (BIC = 537127) or Domain-Specific (BIC = 537127) models. Further analysis showed the dimensionality of the representation in the right 8C was 1.19, suggesting the cognitive space was close to 1D. We also tested if the observed conflict similarity effect was driven solely by spatial Stroop or Simon conflicts, and found larger BICs for the models only including the Stroop similarity (i.e., the Stroop-Only model, BIC = 5377122) or Simon similarity (i.e., the Simon-Only model, BIC = 5377096). An additional Stroop+Simon model, including both StroopOnly and Simon-Only regressors, also showed a worse model fitting (BIC = 5377118). Moreover, we replicated the results with only incongruent trials, considering that the pattern of conflict representations is more manifested when the conflict is present (i.e., on incongruent trials) than not (i.e., on congruent trials). We found a poorer fitting in Domain-general (BIC = 1344129), Domain-Specific (BIC = 1344129), Stroop-Only (BIC = 1344128), Simon-Only (BIC = 1344120), and Stroop+Simon (BIC = 1344157) models than the Cognitive-Space model (BIC = 1344104). These results indicate that the right 8C encodes an integrated cognitive space for resolving Stroop and Simon conflicts. The more detailed model comparison results are listed in Table 2."

We also estimated the dimensionality of the right 8C with the averaged RSM and found the dimensionality of the cognitive space was ~ 1.19, very close to a 1D space. This result is consistent with our experimental design, as the only manipulated variable is the angular distance between conflict types. We have added these results and the methods to the revised manuscript.

Results:

"Further analysis showed the dimensionality of the representation in the right 8C was 1.19, suggesting the cognitive space was close to 1D."

Methods:

"To better capture the dimensionality of the representational space, we estimated its dimensionality using the participation ratio (Ito & Murray, 2023). Since we excluded the within-subject cells from the whole RSM, the whole RSM is an incomplete matrix and could not be used. To resolve this issue, we averaged the cells corresponding to each pair of conflict types to obtain an averaged 5×5 RSM matrix, similar to the matrix shown in Fig. 1C. We then estimated the participation ratio using the formula:

$$dim = \frac{(\sum_i^m \lambda_i)^2}{\sum_i^m \lambda_i^2},$$

where λi is the eigenvalue of the RSM and m is the number of eigenvalues.

> 1. *Another important factor to consider is how learning within the confined task space, which always negatively correlates the two types of conflicts within each subject, may have influenced the current results. Is statistical dependence of conflict information necessary to use the organized cognitive space to represent conflicts from multiple sources? Answering this question would require a paradigm that can adjust multiple sources of conflicts parametrically and independently. Investigating such dependencies is crucial in order to better understand the adaptive utility of the observed cognitive space of conflict similarity.*

As the central goal of our design was to test the geometry of neural representations of conflict, we manipulated the conflict similarity. The anticorrelated Simon and spatial Stroop conflict aimed to make the overall magnitude of conflict similar among different conflict types. We agree that with the current design the likely cognitive space is not a full 2D space with Simon and spatial Stroop being two dimensions. Instead, the likely cognitive space is a subspace (e.g., a circle) embedded in the 2D space, due to the constraint of anticorrelated Simon and spatial Stroop conflict across conflict types. Nevertheless, the subspace can also be used to test the geometry that similar conflict types share similar neural representations.

To test the full 2D cognitive space, a possible revision of our current design is to have multiple hybrid conditions (like Type 2-4) that cover the whole space. For instance, imagine arrow locations in the first quadrant space. We could have a 3×3 design with 9 conflict conditions, where their horizontal/vertical coordinates could be one of the combinations of 0, 0.5 and 1. This way, the spatial Stroop and Simon conditions would be independent of each other. Notably, however, one potential confounding factor would be that these conditions have different levels of difficulty (i.e., different magnitude of conflict), which may affect the CSE results and their representational similarity.
We have added the above limitations and future designs to the revised 1156 manuscript.

“Another limitation is that in our design, the spatial Stroop and Simon effects are highly anticorrelated. This constraint may make the five conflict types represented in a unidimensional space (e.g., a circle) embedded in a 2D space. Future studies may test the 2D cognitive space with fully independent conditions. A possible improvement to our current design would be to include left, right, up, and down arrows presented in a grid formation across four spatially separate quadrants, with each arrow mapped to its own response button. However, one potential confounding factor would be that these conditions have different levels of difficulty (i.e., different magnitude of conflict), which may affect the CSE results and their representational similarity.”

> *Major comments:*
>
> 1. *The RSM result (and the absence of univariate effect) seem to be a good first step to claim the use of cognitive space of conflict. Yet, the presence of an organized (unidimensional; Fig. 6) and continuous cognitive space should be further tested and backed up.*

We thank the reviewer for recognizing the methods and results of our current work. Indeed, the utilization of a parametric design and RSA to examine organization of neural representations is a widely embraced methodology in the field of cognitive neuroscience (e.g., Freund et al., 2021; Ritz et al., 2022). Our current study aimed primarily to provide original

evidence for whether similar conflicts are represented similarly in the brain, which reflects the geometry of conflict representations (i.e., the structure of differences between conflict representations). We have used multiple criteria to back up the findings by showing the representation is sensitive to the presence of conflict and has behavioral relevance.

We agree that the cognitive space account of cognitive control requires further validation. Therefore, in the revised manuscript, we have added several additional tests to strengthen the evidence supporting the organized cognitive space representation. Firstly, we tested five alternative models (Domain-General, Domain Specific, Stroop-Only, Simon-Only and Stroop+Simon models), and found that the Cognitive-Space model best fitted our data. Secondly, we explicitly calculated the dimensionality of the representation and observed a low dimensionality (1.19D). We have added these results to the "Multivariate patterns of the right dlPFC encodes the conflict similarity" section in the revised manuscript (see also the response to Comment 1).

Furthermore, we utilized data from Experiment 1 to demonstrate the continuity of the cognitive space by showing its ability to predict out-of-sample data. We have included this result to the "Conflict type similarity modulated behavioral congruency sequence effect (CSE)" section in the revised manuscript:

"Moreover, to test the continuity and generalizability of the similarity modulation, we conducted a leave-one-out prediction analysis. We used the behavioral data from Experiment 1 for this test, due to its larger amount of data than Experiment 2. Specifically, we removed data from one of the five similarity levels (as illustrated by the θs in Fig. 1C) and used the remaining data to perform the same mixed-effect model (i.e., the two-stage analysis). This yielded one pair of beta coefficients including the similarity regressor and the intercept for each subject, with which we predicted the CSE for the removed similarity level for each subject. We repeated this process for each similarity level once. The predicted results were highly correlated with the original data, with r = .87 for the RT and r = .84 for the ER, ps < .001."

References:

Freund, M. C., Bugg, J. M., & Braver, T. S. (2021). A Representational Similarity Analysis of Cognitive Control during Color-Word Stroop. Journal of Neuroscience, 41(35), 7388-7402.

Ritz, H., & Shenhav, A. (2022). Humans reconfigure target and distractor processing to address distinct task demands. bioRxiv. doi:10.1101/2021.09.08.459546

> *1. Is the conflict similarity effect not driven by either coding of the weak to strong gradient of the spatial Stroop conflict or the Simon conflict? For example, would simply identifying brain regions that selectively tuned to the Simon conflict continuously enough to create a graded similarity in Fig. C.*

We recognize that our current design and analyzing approach cannot fully exclude the possibility that the current results are driven solely by either Stroop or Simon conflicts, since their gradients are correlated to the conflict similarity gradient we defined. To estimate their unique contributions, we performed a model-comparison analysis. We constructed a Stroop-Only model and a Simon-Only model, with each conflict type projected onto the Stroop (vertical) axis or Simon (horizontal) axis, respectively. The similarity between any two conflict types was defined using the Jaccard similarity index (Jaccard, P., 1901), that is, their intersection divided by their union. By replacing the cognitive space-based conflict similarity regressor with the Stroop-Only and Simon-Only regressors, we calculated their BICs. Results showed that the BIC was larger for Stroop-Only (5377122) and Simon-Only (5377096) than for the cognitive space model (5377094). An additional Stroop+Simon model, including both

Stroop-Only and Simon-Only regressors, also 1220 showed a poorer model fitting (BIC = 5377118) than the cognitive space model.

Moreover, we replicated the results with only incongruent trials. We found a poorer fitting in Stroop-Only (BIC = 1344128), Simon-Only (BIC = 1344120), and Stroop+Simon (BIC = 1344157) models than the Cognitive-Space model (BIC = 1344104). These results indicate that the right 8C encodes an integrated cognitive space for resolving Stroop and Simon conflicts. Therefore, we believe the cognitive space has incorporated both dimensions. We added these additional analyses and results to the revised manuscript (see also the response to the above Comment 1).

> *1. Is encoding of conflict similarity in the unidimensional organized space driven by specific requirements of the task or is this a general control strategy? Specifically, is the recruitment of organized space something specific to the task that people are trained to work with stimuli that negatively correlate the spatial Stroop conflict and the Simon conflict?*

We argue that this encoding is a general control strategy. In our task design, we asked the participants to respond to the target arrow and ignore the location that appeared randomly for them. So, they were not trained to deal with the stimuli in any certain way. We also found the conflict similarity modulation on CSE did not change with more training (We added this result in Note S3), indicating that the cognitive space did not depend on strategies that could be learned through training.

“Note S3. Modulation of conflict similarity on behavioral CSEs does not change across time We tested if the conflict similarity modulation on the CSE is susceptible to training. We collected the data of Experiment 1 across three sessions, thus it is possible to examine if the conflict similarity modulation effect changes across time. To this end, we added conflict similarity, session and their interaction into a mixed-effect linear model, in which the session was set as a categorical variable. With a post-hoc analysis of variance (ANOVA), we calculated the statistical significance of the interaction term.

This approach was applied to both the RT and ER. Results showed no interaction effect in either RT, $F(2,1479) = 1.025$, $p = .359$, or ER, $F(2,1479) = 0.789$, $p = .455$. This result suggests that the modulation effect does not change across time."

Instead, the cognitive space should be determined by the intrinsic similarity structure of the task design. A previous study (Freitas et al., 2015) has found that the CSE across different versions of spatial Stroop and flanker tasks was stronger than that across either of the two conflicts and Simon. In their designs, the stimulus similarity was controlled at the same level, so the difference in CSE was only attributable to the similar dimensional overlap between Stroop and flanker tasks, in contrast to the Simon task. Furthermore, recent studies showed that the cognitive space generally exists to represent structured latent states (e.g., Vaidya et al., 2022), mental strategy cost (Grahek et al., 2022), and social hierarchies (Park et al., 2020). Therefore, we argue that cognitive space is likely a universal strategy that can be applied to different scenarios.

We added this argument in the discussion:

“Although the spatial orientation information in our design could be helpful to the construction of cognitive space, the cognitive space itself was independent of the stimulus-level representation of the task. We found the conflict similarity modulation on CSE did not change with more training (see Note S3), indicating that the cognitive space did not depend on strategies that could be learned through training. Instead, the cognitive space should be determined by the intrinsic similarity structure of the task design. For example, a previous study (Freitas et al, 2015) has found that the CSE across different versions of spatial Stroop

and flanker tasks was stronger than that across either of the two conflicts and Simon. In their designs, the stimulus similarity was controlled at the same level, so the difference in CSE was only attributable to the similar dimensional overlap between Stroop and flanker tasks, in contrast to the Simon task. Furthermore, recent studies showed that the cognitive space generally exists to represent structured latent states (e.g., Vaidya et al., 2022), mental strategy cost (Grahek et al., 2022), and social hierarchies (Park et al., 2020). Therefore, cognitive space is likely a universal strategy that can be applied to different scenarios."

Reference:

Freitas, A. L., & Clark, S. L. (2015). Generality and specificity in cognitive control: conflict adaptation within and across selective-attention tasks but not across selective-attention and Simon tasks. Psychological Research, 79(1), 143-162.

Vaidya, A. R., Jones, H. M., Castillo, J., & Badre, D. (2021). Neural representation of 1280 abstract task structure during generalization. Elife, 10, 1-26.

Grahek, I., Leng, X., Fahey, M. P., Yee, D., & Shenhav, A. Empirical and 1282 Computational Evidence for Reconfiguration Costs During Within-Task 1283 Adjustments in Cognitive Control. CogSci.

Park, S. A., Miller, D. S., Nili, H., Ranganath, C., & Boorman, E. D. (2020). Map 1285 Making: Constructing, Combining, and Inferring on Abstract Cognitive Maps. 1286 Neuron, 107(6), 1226-1238 e1228. doi:10.1016/j.neuron.2020.06.030

> *1. The observed pattern seems to suggest that there is conflict similarity space that is defined by the combination of the conflict similarity (i.e., the strength of conflicts) and the sources of conflict (i.e., the Simon vs the spatial Stroop). What are the rational reasons to separate conflicts of different sources (beyond detecting incongruence)? And how are they used for better conflict resolutions?*

The necessity of separating conflicts of different sources lies in that the spatial Stroop and the Simon effects are resolved with different mechanisms. The behavioral congruency effects of a combined conflict from two different sources were shown to be the summation of the two conflict sources (Liu et al., 2010), suggesting that the conflicts are resolved independently. Moreover, previous studies have shown that different sources of conflict are resolved with different brain regions (Egner, 2008; Li et al., 2017), and at different processing stages (Wang et al., 2013). Therefore, when multiple sources of conflict occur simultaneously or sequentially, it should be more efficient to resolve the conflict by identifying the sources.

We have added this argument to the revised manuscript:

“The rationale behind defining conflict similarity based on combinations of different conflict sources, such as spatial-Stroop and Simon, stems from the evidence that these sources undergo independent processing (Egner, 2008; Li et al., 2014; Liu et al., 2010; Wang et al., 2014). Identifying these distinct sources is critical in efficiently resolving potentially infinite conflicts."

Reference:

Egner, T. (2008). Multiple conflict-driven control mechanisms in the human brain. Trends in Cognitive Sciences, 12(10), 374-380.

Li, Q., Yang, G., Li, Z., Qi, Y., Cole, M. W., & Liu, X. (2017). Conflict detection and 1307 resolution rely on a combination of common and distinct cognitive control networks. Neuroscience and Biobehavioral Reviews, 83, 123-131.

Wang, K., Li, Q., Zheng, Y., Wang, H., & Liu, X. (2014). Temporal and spectral 1310 profiles of stimulus-stimulus and stimulus-response conflict processing. NeuroImage, 89, 280-288.

Liu, X., Park, Y., Gu, X., & Fan, J. (2010). Dimensional overlap accounts for independence and integration of stimulus-response compatibility effects. Attention, Perception, & Psychophysics, 72(6), 1710-1720.

> *1. The congruency effect is larger in conflict type 2, 3, 4 consistently compared to conflict 1 and 5. Are these expected under the hypothesis of unified cognitive space of conflict similarity? Is the pattern of similarity modeled in RSA?*

Yes, this is expected. The spatial Stroop and Simon effects have been shown to be additive and independent (Li et al., 2014). Therefore, the congruency effects of conflict type 2, 3 and 4 would be the weighted sum of the spatial Stroop and Simon effects. The weights can be defined by the sine and cosine of the polar angle.

For instance, in Type 2, wy = sin(67.5°) and wx = cos(67.5°). The sum of the two 1321 weight values (i.e., 1.31) is larger than 1, leading to a larger congruency effect than 1322 the pure spatial Stroop (Conf 1) and Simon (Conf 5) conditions.

Note that this hypothesis underlies the Stroop+Simon model, which assumes the Stroop and Simon dimensions are independently represented in the brain and drive the behavior in an additive fashion. Moreover, the observed difference of behavioral congruency effects may have reflected the variance in the Domain-General model, which treats all conflict types as equivalent, with the only difference between each two conflict types in the magnitude of their conflict. Therefore, we did not model the behavioral congruency effects as a covariance regressor in the major RSA. Instead, we conducted a model comparison analysis by comparing these models and the Cognitive-Space model. Results showed worse model fitting of both the Domain-general and Stroop+Simon models. Specially, the regressor of congruency effect difference in the Domain-General model was not significant (p = .575), which also suggests that the higher congruency effect in conflict type 2, 3 and 4 should not influence the Cognitive-Space model results. We have added these methods and results to the revised manuscript (see also our response to Comment 1):

Methods:

“Model comparison and representational dimensionality

To estimate if the right 8C specifically encodes the cognitive space, rather than the domain-general or domain-specific structures, we conducted two more RSAs. We replaced the cognitive space-based conflict similarity matrix in the RSA we reported above (hereafter referred to as the Cognitive-Space model) with one of the alternative model matrices, with all other regressors equal. The domain-general model treats each conflict type as equivalent, so each two conflict types only differ in the magnitude of their conflict. Therefore, we defined the domain-general matrix as the difference in their congruency effects indexed by the group-averaged RT in Experiment 2. Then the z scored model vector was sign-flipped to reflect similarity instead of distance. The domain-specific model treats each conflict type differently, so we used a diagonal matrix, with within-conflict type similarities being 1 and all cross-conflict type similarities being 0.

Moreover, to examine if the cognitive space is driven solely by the Stroop or Simon conflicts, we tested a spatial Stroop-Only (hereafter referred to as “Stroop-Only”) and a Simon-Only model, with each conflict type projected onto the spatial Stroop (vertical) axis or Simon (horizontal) axis, respectively. The similarity between any two conflict types was defined using the Jaccard similarity index (Jaccard, 1901), that is, their intersection divided by their

union. We also included a model assuming the Stroop and Simon dimensions are independently represented in the brain, adding up the Stroop Only and Simon-Only regressors. We conducted similar RSAs as reported above, replacing the original conflict similarity regressor with the Strrop-Only, Simon-Only, or both regressors, and then calculated their Bayesian information criterions (BICs)."

Reference:

Li, Q., Nan, W., Wang, K., & Liu, X. (2014). Independent processing of stimulus stimulus and stimulus-response conflicts. PloS One, 9(2), e89249.

> *1. Please clarify the observed patterns of CSE effects in relation to the hypothesis of common cognitive space of conflict. In particular, right 8C shows that the patterns become dissimilar in incongruent trials compared to congruent trials. How does this direction of the effect fit to the common unidimensional cognitive space account? And how does such a representation contribute to the CES effects?*

The behavioral CSE patterns provide initial evidence for the cognitive space hypothesis. Previous studies have debated whether cognitive control relies on domain-general or domain-specific representations, with much evidence gathered from behavioral CSE patterns. A significant CSE across two conflict conditions typically suggests domain-general representations of cognitive control, while an absence of CSE suggests domain-specific representations. The cognitive space view proposes that conflict representations are neither purely domain-general nor purely domain-specific, but rather exist on a continuum. This view predicts that the CSE across two conflict conditions should depend on the representational distance between them within this cognitive space. Our finding that CSE values systematically vary with conflict similarity level support this hypothesis. We have added this point in the discussion of the revised manuscript:

“Previous research on this topic often adopts a binary manipulation of conflict(Braem et al., 2014) (i.e., each domain only has one conflict type) and gathered evidence for the domain-general/specific view with presence/absence of CSE, respectively. Here, we parametrically manipulated the similarity of conflict types and found the CSE systematically vary with conflict similarity level, demonstrating that cognitive control is neither purely domain-general nor purely domain-specific, but can be reconciled as a cognitive space(Bellmund et al., 2018) (Fig. 6, middle).

Fig. 4D was plotted to show the steeper slope of the conflict similarity effect for incongruent versus congruent conditions. Note the y-aixs displays z-scored Pearson correlation values, so the grand mean of each condition was 0. The values for the first two similarity levels (level 1 and 2) were lower for incongruent than congruent conditions, seemingly indicating lower average similarity. However, this was not the case. The five similarity levels contained different numbers of data points (see Fig. 1C), so levels 4 and 5 should be weighted more heavily than levels 1 and 2. When comparing the grand mean of raw Pearson correlation values, the incongruent condition (0.0053) showed a tendency toward higher similarity than the congruent condition (0.0040), $t(475998) = 1.41$, $p = .079$. We have also plotted another version of Fig. 4D in Fig. S5, in which the raw Pearson correlation values were used.

The greater representation of conflict type in incongruent condition compared to congruent condition (as evidenced by a steeper slope) suggests that the conflict representation was driven by the incongruent condition. This is probably due to the stronger involvement of cognitive control in incongruent condition (than congruent condition), which in turn leads to more distinct patterns across different conflict types. This is consistent with the fact that the congruent condition is typically a baseline, where any conflict related effects should be weaker.

The representation of cognitive space may contribute to the CSE as a mental model. This model allows our brain to evaluate the cost and benefit associated with transitioning between different conflict conditions. When two consecutive trials are characterized by more similar conflict types, their representations in the cognitive space will be closer, resulting in a less costly transition. As a consequence, stronger CSEs are observed. We revised the corresponding discussion part as:

"Similarly, we propose that cognitive space could serve as a mental model to assist fast learning and efficient organization of cognitive control settings. Specifically, the cognitive space representation may provide a principle for how our brain evaluates the expected cost of switching and the benefit of generalization between states and selects the path with the best cost-benefit tradeoff (Abrahamse et al., 2016; Shenhav et al., 2013). The proximity between two states in cognitive space could reflect both the expected cognitive demand required to transition and the useful mechanisms to adapt from. The closer the two conditions are in cognitive space, the lower the expected switching cost and the higher the generalizability when transitioning between them. With the organization of a cognitive space, a new conflict can be quickly assigned a location in the cognitive space, which will facilitate the development of cognitive control settings for this conflict by interpolating nearby conflicts and/or projecting the location to axes representing different cognitive control processes, thus leading to a stronger CSE when following a more similar conflict condition."

> *Minor comments:*
>
> *1. Some of the labels of figure axes are unclear (e.g., Fig4C) about what they represent.*

In Fig. 4C, the x-axis label is "neural representational strength", which refers to the beta coefficient of the conflict type effect computed from the main RSA, denoting the strength of the conflict type representation in neural patterns. The y-axis label is "behavioral representational strength", which refers to the beta coefficient obtained from the behavioral linear model using conflict similarity to predict the CSE in Experiment 2; it reflects how strong the conflict similarity modulates the behavioral 1440 CSE. We apologize for any confusion from the brief axis labels. We have added expanded descriptions to the figure caption of Fig. 4C.