Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint posted
9 May 2023 (this version)

Sent for peer review
21 February 2023

Posted to bioRxiv
14 February 2023

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

Abstract

Cognitive control resolves conflict between task-relevant and -irrelevant information to enable goal-directed behavior. As conflict can arise from different sources (e.g., sensory input, internal representations), how a finite set of cognitive control processes can effectively address huge array of conflict remains a major challenge. We hypothesize that different conflict 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. To test this hypothesis, we designed a task with five types of conflict that could be conceptually parameterized along one dimension. Over two experiments, both human performance and fMRI activity patterns in the right dorsolateral prefrontal (dlPFC) support that different types of conflict are organized in a cognitive space. The findings suggest that cognitive space can be a dimension reduction tool to effectively organize neural representations of conflict for cognitive control.

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 two different sources of difficulty and reports that the brain appears to represent these sources as a mixture on a continuum, in the prefrontal areas involved in resolving task difficulty. While these results are **useful**, they overlap with previous findings, leave open several design and logical concerns, and rely on novel statistical analyses that may require further validation, so they are currently **incomplete**.

Introduction

Cognitive control enables humans to behave purposefully by modulating neural processing to resolve conflict 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 conflict and how they should be resolved can vary greatly(1). For example, 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 conflicts. A key unsolved question in cognitive control is how our brain efficiently resolves a nearly infinite number of different types of conflict.

A first step to addressing this question is to examine the commonalities and/or dissociations across different types of conflict that can be categorized into different *domains*. Examples of the domains of conflict include experimental paradigm(2),(3), sensory modality(4),(5), or conflict type regarding the dimensional overlap of conflict processes(6),(7).

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 conflict. This is supported by the generalizable control adjustment (i.e., encountering a conflict trial from one type can facilitate conflict resolution of another type) (2),(8) and similar neural patterns(9),(10) 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(11),(12), 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(13). Therefore, the domain-general processes is evolutionarily less advantageous for humans to deal with the diverse situations requiring high efficiency(14). On the other hand, the *domain-specific* theories argue that different types of conflict are handled by distinct cognitive control processes (e.g., where and how information processing should be modulated)(15),(16). However, according to the domain-specific view, the potentially unlimited conflict situations require a large variety of preexisting control processes, which is biologically implausible(17).

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 et al.(18) 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 conflict. Similarly, different brain networks may show domain-generality (i.e., representing multiple conflicts) or domain-specificity (i.e., representing individual conflicts separately)(7), (19). Even within the same brain area (e.g., medial frontal cortex), Fu et al.(20) 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[21], 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 conflict using a system with limited, dissociable dimensions. A construct with a similar function is the *cognitive space*[22], which extends the idea of cognitive map[23] to the representation of abstract information. Critically, the cognitive space view holds that the representations of different abstract information are organized continuously and the locations of representations in the cognitive space are determined by the similarity among the represented information[22].

In the human brain, it has been shown that abstract[23],[24] and social[25] 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[25]. In the field of cognitive control, recent studies have begun to conceptualize different control states within a cognitive space[26]. For example, Fu et al.[20] mapped different conflict conditions to locations in a low/high dimensional cognitive space to demonstrate the domain-general/domain-specific problems; Grahek et al.[27] 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[22] and non-spatial knowledge[23]. A challenge to answering this question lies in how to construct control states with continuous levels of similarity. Our recent work[28] 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[22],[23]. Recent research has suggested that the abstract task structure could be encoded and implemented by the frontoparietal network[29],[30], but whether a similar neural system encodes the cognitive space of cognitive control remains untested.

We hypothesize that different types of conflict are represented as points in a cognitive space. The dimensions in the cognitive space of conflict 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 conflict with low dimensionality[26],[31]. Different domains can be represented conjunctively in a single cognitive space to achieve domain-general cognitive control, as conflict 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 conflict may be located in the frontoparietal network due to its essential roles in conflict resolution[20],[32] and abstract task representation[30].

In this study, we adjusted the paradigm from our previous study[28] 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.

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[6],[33]. 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[33],[34]. 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 conflict[28]. 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.

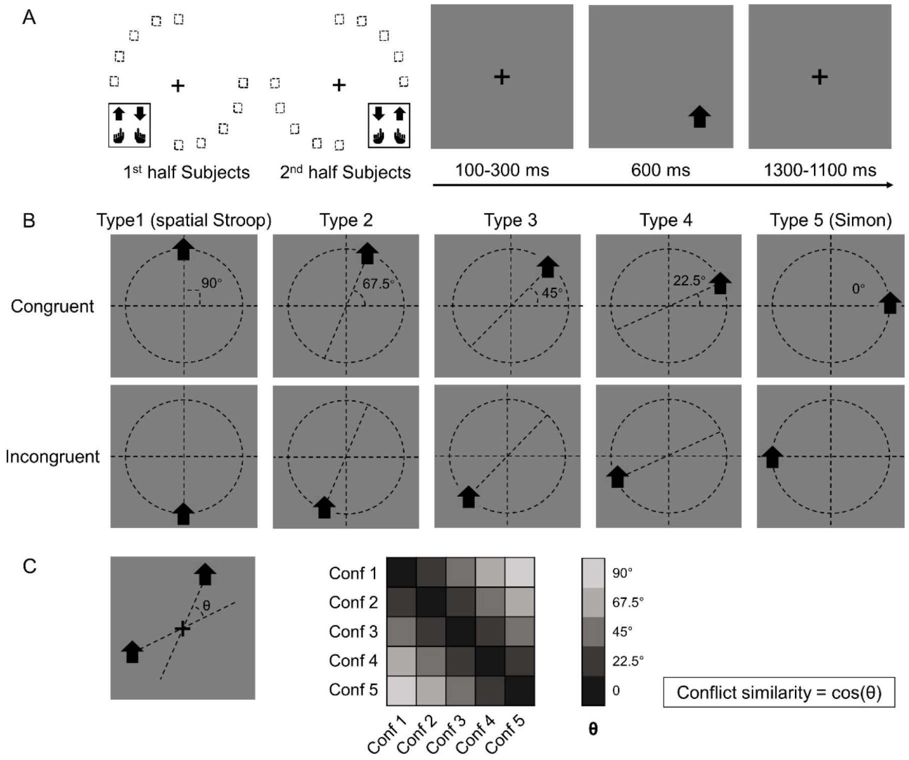


Fig. 1.

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 conflict, 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 Conflict 1 and Conflict 5 form the largest angle of 90°, and others are in between. Conflict similarity is defined by the cosine value of θ.

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). 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 conflict 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(35), (36). It has been shown that the CSE diminishes if the two consecutive trials have different conflict types(37)-(39). 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 design, 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.

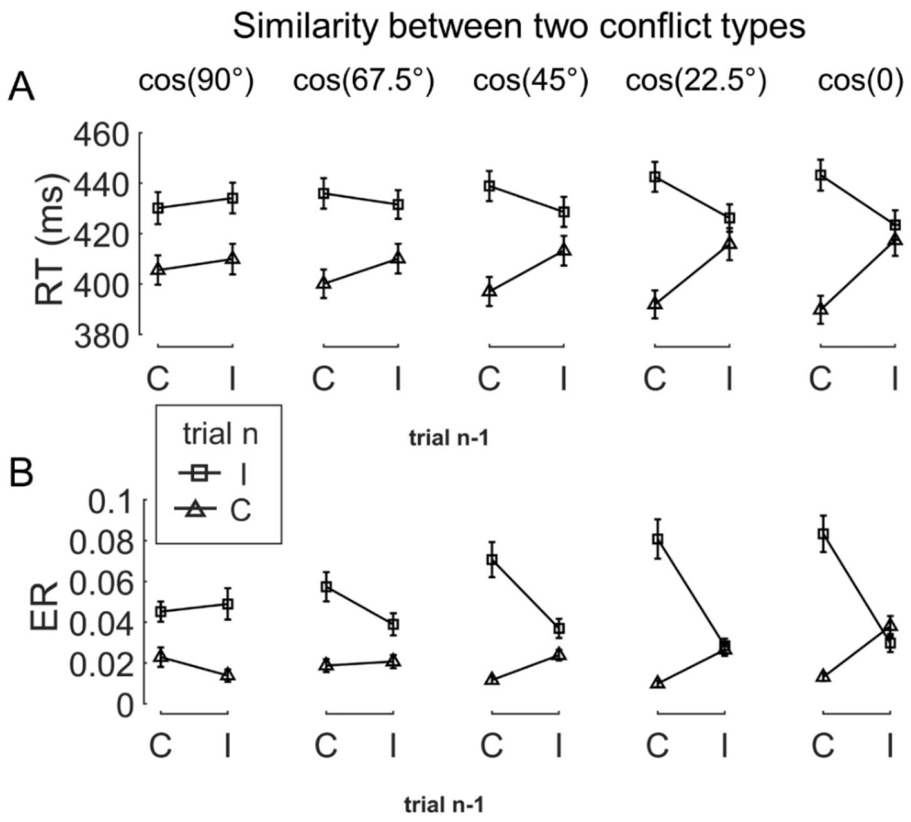


Fig. 2.

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.

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: $\beta = 0.10 \pm 0.01$, $t(1978) = 15.82$, $p < .001$, $\eta_p^2 = .120$; ER: $\beta = 0.15 \pm 0.02$, $t(1978) = 7.84$, $p < .001$, $\eta_p^2 = .085$, 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[28]. The two-stage analysis also showed a strong effect of conflict similarity (RT: $\beta = 0.58 \pm 0.04$, $t(493) = 14.74$, $p < .001$, $\eta_p^2 = .383$; ER: $\beta = 0.36 \pm 0.05$, $t(493) = 7.01$, $p < .001$, $\eta_p^2 = .321$, 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.

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).

Brain activations modulated by conflict type dissimilarity 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(20),(40), this effect was observed in the pre-supplementary motor area (pre-SMA) and anterior cingulate cortex (ACC) areas (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, in the left middle frontal gyrus (MFG), fMRI activation scaled with the increase in spatial Stroop conflict, whereas the right inferior parietal sulcus (IPS) and the right dorsomedial prefrontal cortex (dmPFC) displayed positive correlation between fMRI activation and Simon conflict (Fig. 3, Fig. S3, Table S1).

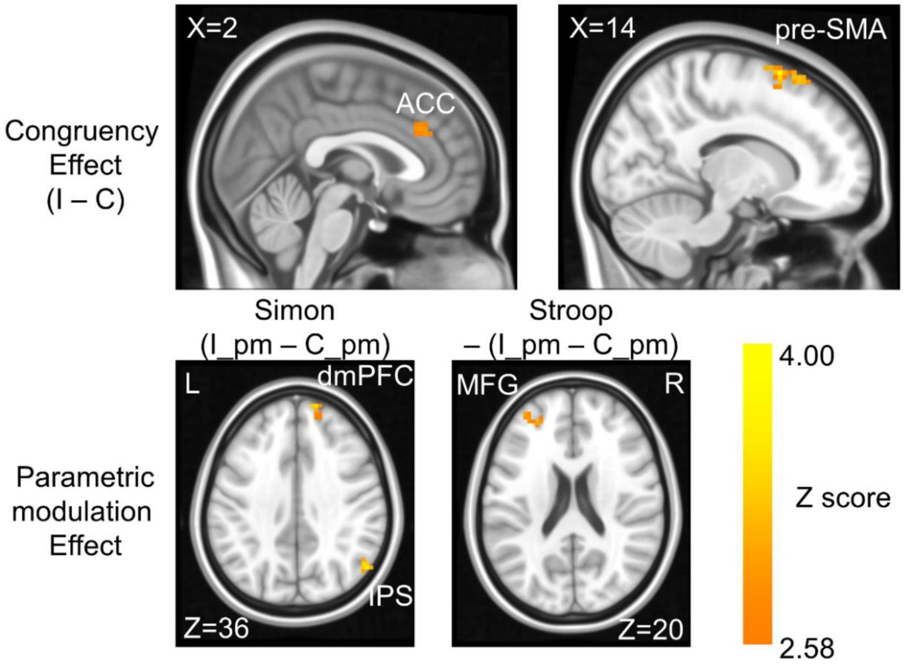


Fig. 3.

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

Results displayed are thresholded with voxel-wise one-tailed $p < .005$ and cluster-size > 20 voxels. The congruency effect denotes the higher activation in incongruent than congruent condition. 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 (bottom left panel). The negative parametric modulation effect [converted to positive with – (I_pm – C_pm)] denotes the higher activation when the conflict type contained a higher ratio of spatial Stroop conflict component (bottom right panel). I = incongruent; C = congruent; pm = parametric modulator.

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 p_{FDR} >.762, 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 (conflict 1, 2, 3, 4 and 5, as shown in 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[(41),(42)]. 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. 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. S5). 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 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 8C area (a subregion of dlPFC[(42)]), a47r, TPOJ3 and V3CD in the right hemisphere, and the 6r, 7Am, 24dd, VMV1, VMV2, 7Pl, 23c and 25 areas in the left hemisphere (p_{FDR}s < 0.05, with raw *ps* < 0.001, 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 lateral area 7P (7P1), left ventromedial visual area 1 (VMV1), left dorsal area 24d (24dd), right Brodmann area 8C (8C), and right V3CD met this criterion, p_{FDR}s < .01, one-tailed (Table 1, Fig. 4B), suggesting that the representation of conflict type was strengthened when conflict was present (e.g., Fig. 4D). 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* = 0.43, p_{FDR} = .027, one-tailed, Fig. 4C) but not in the other regions (all p_{FDR} > .632, one-tailed). In addition, we did not find evidence supporting the encoding of congruency in the right 8C area (see Note S5), suggesting that the right 8C area specifically represents conflict similarity. 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.

Table 1.

Summary statistics of regions showing larger encoding strength in incongruent than congruent conditions for the conflict type and orientation effects.

Region name	$t(34)$	β(SD)	Cohen's d	p_{FDR}
		Conflict type effect		
left 7P1	3.13	0.0049 ± 0.0016	0.53	.011
left VMV1	3.96	0.0077 ± 0.0019	0.67	.002
left 24	7.82	0.0094 ± 0.0012	1.32	< .001
right 8C	3.15	0.0073 ± 0.0023	0.53	.011
right V3CD	2.86	0.0057 ± 0.0020	0.48	.017
		Orientation effect		
left V2	3.20	0.0107 ± 0.0033	0.54	.007
left FEF	2.97	0.0066 ± 0.0022	0.50	.010
left IP2	5.73	0.0129 ± 0.0022	0.97	.001
right V1	2.70	0.0060 ± 0.0022	0.46	.014
right V2	3.26	0.0083 ± 0.0025	0.55	.007
right H	2.79	0.0037 ± 0.0013	0.47	.014
right PF	5.31	0.0097 ± 0.0018	0.90	< .001

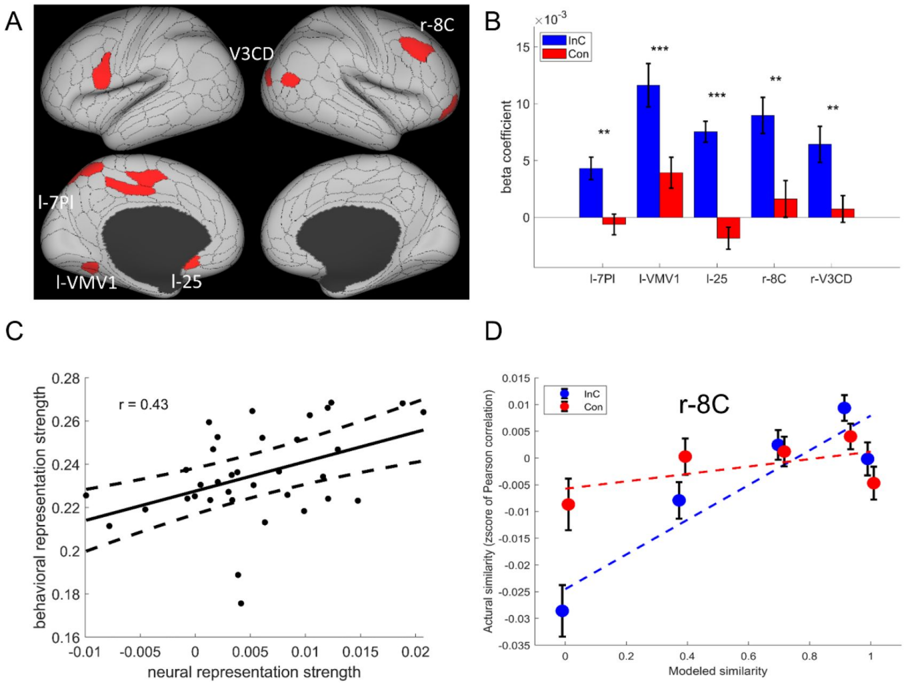


Fig. 4.

The conflict type effect.

(A) Brain regions surviving the FDR-correction ($p_{FDR} < 0.05$ and $p < 0.001$) across the 360 regions (criterion 1). Labeled regions are those meeting the criterion 2. (B) The regions showing stronger encoding of conflict type in the incongruent than congruent conditions (criterion 2). ** $p_{FDR} < .01$, *** $p_{FDR} < .001$. (C) The brain-behavior correlation of the right 8C (criterion 3). (D) Illustration of the different encoding strength of conflict type similarity in incongruent versus congruent conditions of right 8C. l = left; r = right.

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, bilateral V2 and V3, right V4, left area temporoparietooccipital junction 3 (TPOJ3) and right PHT areas (FDR corrected *p*s < 0.05 and raw *p*s < 0.001). We also found the same effect in several oculomotor related regions, including the left frontal eye field (FEF), anterior 6m (6ma), area intraparietal 2 (IP2), right parietal area F (PF) and bilateral 5m, as well as other regions (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, bilateral V2, left FEF, left IP2, right hippocampus (H) and right PF encoded stronger orientation effect in the incongruent than the congruent condition, p_{FDR}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 p_{FDR} > .675, 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 S5). 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 conflict was present.

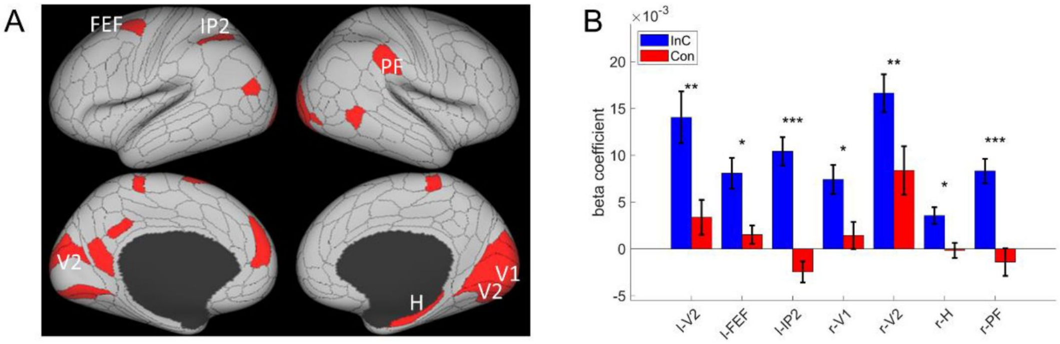


Fig. 5.

The axial orientation effect.

(A) Brain regions surviving the FDR-correction ($p_{\mathrm{FDR}} < 0.05$ and $p < 0.001$) across the 360 regions (criterion 1). Labeled regions are those meeting the criterion 2. (B) The regions showing stronger encoding of orientation in the incongruent than congruent conditions (criterion 2). * $p_{\mathrm{FDR}} < .05$, ** $p_{\mathrm{FDR}} < .01$, *** $p_{\mathrm{FDR}} < .001$.

To explore the relation between conflict type and orientation representations, we conducted representational connectivity (i.e., the similarity between two RSMs of two regions)[43] analyses and found that among the orientation effect regions, the right V1 and bilateral 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 three 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 left V3, β = 0.1447 ± 0.0069), we found higher connectivity in the left V2, β = 0.1645 ± 0.0060, $t(34)$ = 4.86, right V1, β = 0.1628 ± 0.0065, $t(34)$ = 4.54, and right V2, β = 0.1678 ± 0.0074, $t(34)$ = 5.65, all $p_{FDR} < .001$, one-tailed (Fig. S6).

Discussion

Understanding how different types of conflict 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(15),(21). Reconciliation of the two views also suffers from the inability to fully address how infinite conflict can be resolved by a limited set of cognitive control processes. In this study, we hypothesized that this issue can be addressed if conflict is organized as a cognitive space. Leveraging the well-known dissociation between the spatial Stroop and Simon conflict(44)-(46), 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(28). 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 three brain regions (right IPS, right dlPFC and left MFG) 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 the conflicts are organized in a cognitive space that enables a limited set of cognitive control processes to resolve infinite possible types of conflict.

Conventionally, the domain-general view of control suggests a common representation for different types of conflict (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 conflict(21) (i.e., each domain only has one conflict type) and thus is not suitable to test the cognitive space hypothesis. Here, we parametrically manipulated the similarity of conflict in different conflict types and demonstrated that the two theories can be reconciled as a cognitive space(22) (Fig. 6, middle). Specifically, the cognitive space provides a solution to use a single cognitive space organization to encode different types of conflict 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). Moreover, geometry can also emerge in the cognitive space(20), 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).

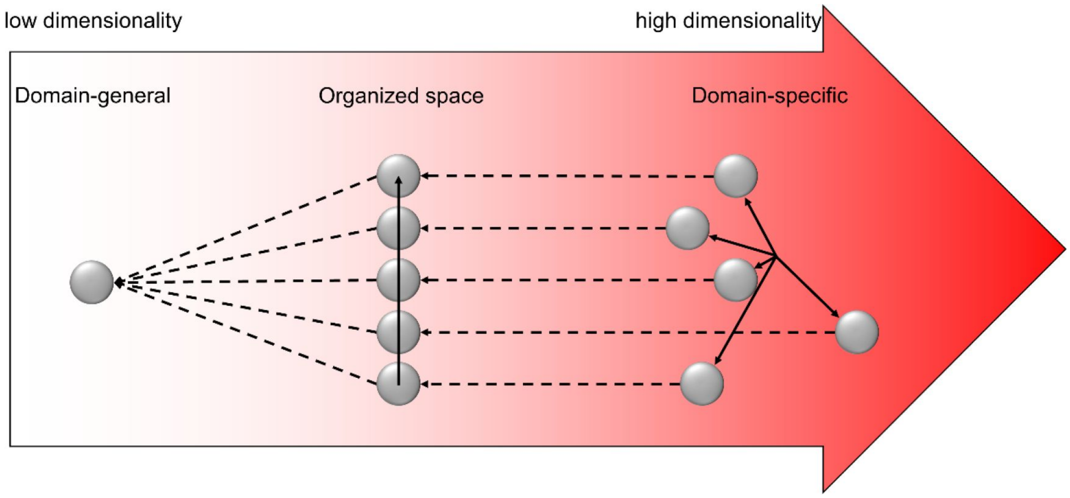


Fig. 6.

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.

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 conflict 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)(47). 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(26).

Although our finding of cognitive space in the right dlPFC differs from other cognitive space studies(24),(25),(48) 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(49)-(51) and is widely engaged in multiple task demands(12),(52). However, it is not until recently that the multivariate representation in this region has been examined. For instance, Vaidya et al.(29) reported that frontal regions presented latent states that are organized hierarchically. Freund et al.(32) 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 S6), 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(53),(54), which is consistent with our findings. Moreover, the right PFC also represents a composition of single rules(55), which may explain how the spatial Stroop and Simon types can be jointly encoded in a single space.

We found that participants with stronger conflict representation as cognitive space in right dlPFC 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 conflict 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(56). Similarly, we propose that the cognitive space could serve as a mental model to assist fast learning and efficient organization of cognitive control settings. 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. On the other hand, without a cognitive space, there would be no measure of similarity between conflict 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 bilateral V2 encoded the spatial orientation of the target location (Fig. 5) and showed strong representational connectivity with the right dlPFC (Fig. S6), 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(57), which is consistent with the idea of a low-dimensional representation of cognitive control(26),(31). The compression and abstraction processes might be why the frontoparietal regions are the top of hierarchy of information processing(58) and why the frontoparietal regions are widely engaged in multiple task demands(59).

With conventional univariate analyses, we observed that the overall congruency effect was located at the medial frontal regions (i.e., pre-SMA and ACC), which is consistent with previous studies(20),(40). Beyond that, we also found regions that can be parametrically modulated by conflict type difference, including right IPS, right dlPFC (modulated by Simon difference) and left MFG (modulated by spatial Stroop difference). The lateralization of these regions is consistent with a previous finding(19), which highlighted the difference of Stroop and Simon types with brain activities at different hemispheres. The scaling of brain activities based on conflict difference is potentially important to the representational organization of different types of conflict. 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(60). Future studies may adopt approaches such as repetition suppression induced fMRI adaptation(26) to test the role of univariate activities in task representations.

One limitation of this study needs 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(61). Future studies may replicate our findings with a multiple-choice design with confound-free trial sequences(62).

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(28). 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(28) (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 conflict: 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(33),(63), which contains the dimensional overlap between task-relevant stimulus and task-irrelevant stimulus(1); 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(33),(34), which contains the dimensional overlap between task-irrelevant stimulus and response(1). 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 conflict. 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 conflict.

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(64) conducted by the NeuroDesign package(65) (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(66) (RRID:SCR_016216), which is based on Nipype 1.5.1(67) (RRID:SCR_002502). Specifically, BOLD runs were slice-time corrected using 3dTshift from AFNI 20160207(68) (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[28]. 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(7),(19), 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 Conf 1, Conf 2, Conf 3, Conf 4, and Conf 5 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)(41). 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 conflict 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 thresholded by 3dclust function in AFNI (69) with voxel-wise p < .005 and cluster-size > 20 voxels, which was supposed to produce a desirable balance between Type I and II error rates(70). 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[42]. 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. S4)[60],[71]. 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[72]. 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. Similar to the behavioral analyses, we assumed the conflict similarity between two trials is commutive and hence collapsed the RSM along the diagonal and converted the lower triangle 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 three regressors denoting the similarity of Run (i.e., whether two conditions are within the same run), Subject (i.e., whether two conditions are within the same subject), and 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. The intercept and slopes of the regressors were set as random effects at the subject level. Individual effects for each regressor were also extracted from the model for statistical inference and 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 determined with one-sample t-tests (one-tailed). Multiple comparison correction was applied with false discovery rate (FDR) approach[73] across all cortical regions ($p_{\mathrm{FDR}} < 0.05$), together with a threshold of 0.001 for each region. To test if the representation strengths are different between congruent and incongruent conditions, we also conducted the RDM analyses using only congruent and incongruent trials separately. Individual effects were extracted from each model and tested using a paired t-test. 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).

Representational connectivity analysis

To explore the possible relevance between the conflict type and the orientation effects, we conducted representational connectivity[43] between regions showing evidence encoding conflict similarity and orientation similarity. Similar to the RSA mentioned above, 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., bilateral V2) as the predictor. Intercept and the slope of the regressor were set as random effects at the subject level, and individual coefficients of the slope were extracted for further statistical analysis. 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., $q_{\mathrm{FDR}} > 0.05$ 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 higher connectivity slope than any of the control regions with paired-t tests.

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.

Supplementary Notes

Note S1. Behavioral congruency effects

To test the congruency effects for the five conflict types, we conducted 5 (conflict type) × 2 (congruency) repeated-measure ANOVAs with RT and ER from both experiments. The results are displayed in Supplementary Fig. 1.

Experiment 1

For the RT, we observed a significant main effect of Congruency, $F(1, 32) = 407.70$, $p < .001$, ${\eta_p}^2 = .93$, a significant main effect of Conflict Type, $F(4, 128) = 6.32$, $p < .001$, ${\eta_p}^2 = .16$, and an interaction between Conflict Type and Congruency, $F(4, 128) = 27.86$, $p < .001$, ${\eta_p}^2 = .47$. Simple effect analyses showed that participants responded more slowly in incongruent conditions than in congruent conditions for all conflict types, p_{FDR}s < .001. Additionally, the congruency effect of the Type 2, 3 and 4 were significantly larger than that of the Type 1, and the congruency effect of the Type 2 and 3 were significantly larger than that of the Type 5, p_{FDR}s < .05.

Similar results were found with the ER. We observed a significant main effect of Congruency, $F(1, 32) = 56.83$, $p < .001$, ${\eta_p}^2 = .64$, a significant main effect of Conflict Type, $F(4, 128) = 6.29$, $p < .001$, ${\eta_p}^2 = .16$, and an interaction between Conflict Type and Congruency, $F(4, 128) = 13.23$,

$p < .001$, $\eta_p^2 = .29$. Simple effect analyses showed that participants were more error-prone in incongruent conditions than in congruent conditions for all conflict types, p_{FDR}s < .001. The congruency effect of the Type 2, 3 and 4 were significantly larger than that of the Type 1, and the congruency effect of the Type 3 and 4 were significantly larger than that of the Type 5, p_{FDR}s < .05.

Experiment 2

For the RT, We observed a significant main effect of Congruency, $F(1, 34) = 149.71$, $p < .001$, $\eta_p^2 = .81$, a significant main effect of Conflict Type, $F(4, 136) = 10.11$, $p < .001$, $\eta_p^2 = .23$, and an interaction between Conflict Type and Congruency, $F(4, 136) = 7.63$, $p < .001$, $\eta_p^2 = .18$. Simple effect analyses showed that participants responded more slowly in incongruent conditions than in congruent conditions for all conflict types, p_{FDR}s < .001. The congruency effect of the Type 4 condition was larger than that of Type 1, and Type 3 and Type 4 were significantly larger than that of the Type 5, p_{FDR}s < .05.

For the ER, we only observed a significant main effect of Congruency, $F(1, 34) = 29.80$, $p < .001$, $\eta_p^2 = .47$. All the types showed a larger error rate in incongruent than congruent conditions (p_{FDR}s < .001), except that the Type 1 only showed a marginal significance ($p_{FDR} = .062$).

In sum, we observed strong behavioral congruency effects in both experiments. The findings indicate that these conflict conditions indeed engaged cognitive control[1].

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., Conf 2−Conf 3 and Conf 3−Conf 2) 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: $\beta = 0.09 \pm 0.01$, $t(7812) = 13.74$, $p < .001$, $\eta_p^2 = .025$; ER: $\beta = 0.09 \pm 0.01$, $t(7812) = 7.66$, $p < .001$, $\eta_p^2 = .018$) and Experiment 2 (RT: $\beta = 0.21 \pm 0.02$, $t(3956) = 9.88$, $p < .001$, $\eta_p^2 = .043$; ER: $\beta = 0.20 \pm 0.03$, $t(4201) = 6.11$, $p < .001$, $\eta_p^2 = .038$). Thus, the observed modulation of conflict similarity on behavioral CSEs cannot be explained by physical proximity.

Note S3. The fMRI sequence generation approach

Two sequences of 170 trials each were generated independently with the NeuroDesign package [2]. Each sequence was initialized as 10 consecutive sub-blocks of each condition (incongruent and congruent) for each conflict type (Conf 1, Conf 2, Conf 3, Conf 4 and Conf 5). The contrasts of interest were the main effect of congruency (i.e., [1 −1 1 −1 1 −1 1 −1 1 −1]) and the parametric effect (i.e., [−2 −2 −1 −1 0 0 1 1 2 2]). The order was optimized after 5000 cycles of crossover, mutation, immigration, fitness, and natural selection. The final number of trials for different conflict types varied from 64 to 73.

Note S4. fMRI data preprocessing

Results included in this manuscript come from preprocessing performed using fMRIPrep 20.2.0 (RRID:SCR_016216)[3], which is based on Nipype 1.5.1 (RRID:SCR_002502)[4].

Anatomical data preprocessing

The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection[5], distributed with ANTs 2.3.3 (RRID:SCR_004757)[6], and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, RRID:SCR_002823)[7]. Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: ICBM 152 Nonlinear Asymmetrical template version 2009c [RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym][8].

Functional data preprocessing

For each of the 5 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Susceptibility distortion correction (SDC) was omitted. The BOLD reference was then co-registered to the T1w reference using flirt (FSL 5.0.9)[9] with the boundary-based registration[10] cost-function. Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9)[11]. BOLD runs were slice-time corrected using 3dTshift from AFNI 20160207 (RRID:SCR_005927)[11]. The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying the transforms to correct for head-motion. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Several confounding time-series were calculated based on the preprocessed BOLD: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following Power (absolute sum of relative motions)[11] and Jenkinson (relative root mean square displacement between affines, Jenkinson et al.[9]). FD and DVARS are calculated for each functional run, both using their implementations in Nipype (following the definitions by Power et al.[11]). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (CompCor)[12]. Principal components are estimated after high-pass filtering the preprocessed BOLD time-series (using a discrete cosine filter with

128s cut-off) for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 2% variable voxels within the brain mask. For aCompCor, three probabilistic masks (CSF, WM and combined CSF+WM) are generated in anatomical space. The implementation differs from that of Behzadi et al. in that instead of eroding the masks by 2 pixels on BOLD space, the aCompCor masks are subtracted a mask of pixels that likely contain a volume fraction of GM. This mask is obtained by thresholding the corresponding partial volume map at 0.05, and it ensures components are not extracted from voxels containing a minimal fraction of GM. Finally, these masks are resampled into BOLD space and binarized by thresholding at 0.99 (as in the original implementation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each([12]). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardised DVARS were annotated as motion outliers. All resamplings can be performed with a single interpolation step by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels([13]). Non-gridded (surface) resamplings were performed using mri_vol2surf (FreeSurfer).

Many internal operations of fMRIPrep use Nilearn 0.6.2 (RRID:SCR_001362)([14]), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in fMRIPrep's documentation.

Note S5. The multivariate representations of conflict type and orientation are different from the congruency effect

An explanation to the stronger encoding of conflict type in incongruent than congruent condition (Fig. 3B/D) in right 8C area may be the encoding of congruency. To test this possibility, we first tested the univariate congruency effect (incongruent minus congruent) using the parametric modulating GLM1 that was used to estimate fMRI activation levels of conflict type × congruency conditions. We observed no univariate congruency effect in the right 8C region, $t(34) = -0.03$, $p = .513$, one-tailed. We further tested the possibility that the congruency effect may be manifested in behavioral relevance. To this end, we extracted the congruency effect (incongruent minus congruent) on encoding strength of conflict similarity for each subject from the mixed-effect model based on the cross-subject RSA (see the *Representational similarity analysis* of Methods in the main text) and correlated it with the behavioral congruency effect, averaged across the five conflict types (i.e., the main effect reported in the Note S1). No significant correlation was observed ($r = 0.14$, $p = .380$, one-tailed). Taken together, these results suggested that the neural encoding strength of conflict type does not reflect the level of cognitive control engagement, but the dynamic adjustment of cognitive control instead.

Similarly, we tested whether those regions with stronger encoding of orientation in incongruent than congruent condition (i.e., bilateral V2, left FEF, left IP2, right V1, right H and right PF) reflects the congruency effect. We observed no uni-voxel congruency effect in any of these regions, all $p_{FDR} > .998$, one-tailed. In addition, the orientation effect was not correlated to the behavioral congruency in any of the regions, all $p_{FDR} > .608$, one-tailed. Together with our finding that there was no correlation between the strength of orientation encoding and the conflict similarity modulation on behavioral CSEs in any of these regions (see the *Multivariate patterns of visual and oculomotor areas encode stimulus orientation* of Results in the main text), these results indicate that the encoding of orientation effect did not reflect the encoding of congruency or conflict type. Instead, we speculate that the encoding of orientations provides perceptual information to determine the conflict type.

Note S6. 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[(15)]. We found that none of these regions show the representation of conflict type, all $p_{FDR} > .99$. These results indicate that the conflict type is specifically represented in the right dlPFC.

Supplementary Figures

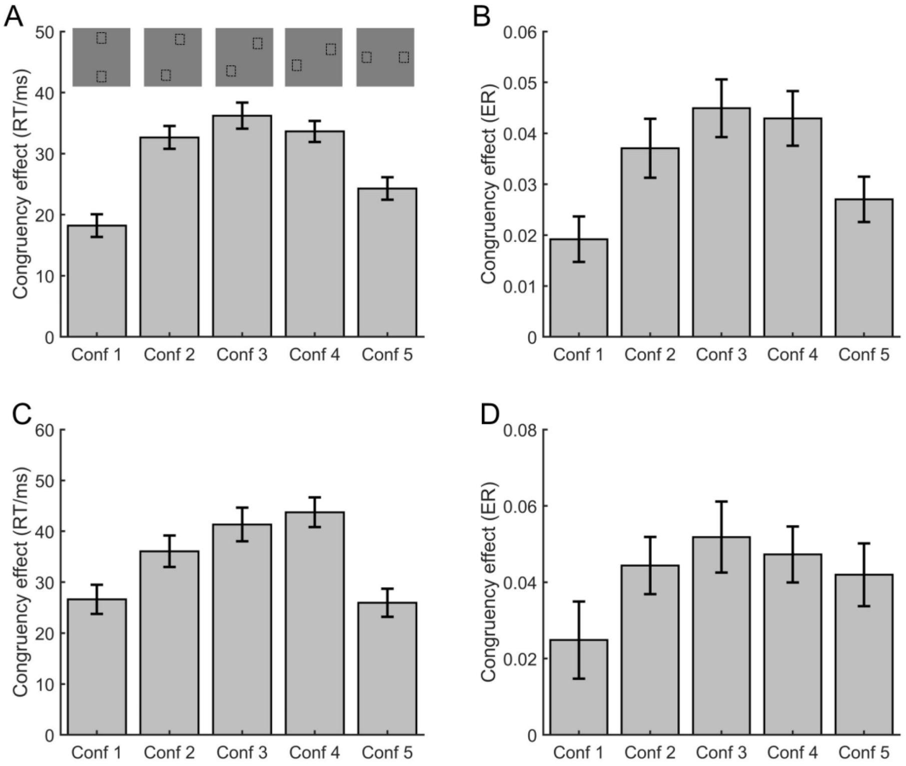


Fig. S1.

The congruency effects of Experiment 1 (A and B) and Experiment 2 (C and D). Error bars denote the standard errors of mean. Conf 1 to 5 denotes the five conflict types. Small insets on top of panel A denote an example of stimuli positions for each conflict type. RT = reaction time; ER = error rate.

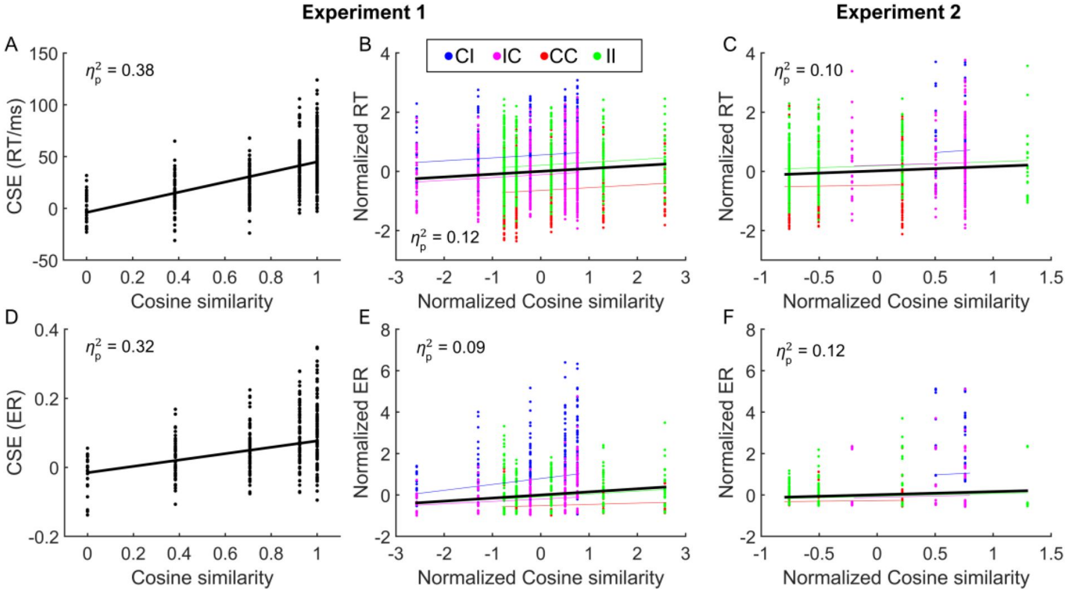


Fig. S2.

The conflict similarity modulation on performance of Experiment 1 (A, B, D and E) and Experiment 2 (C and F), respectively. A and D are scatter plots of CSE [i.e., (CI−CC) − (II−IC)] for RT and ER as a function of the cosine similarity, respectively. In B, C, E and F, the cosine similarity and RT / ER are normalized across conflict similarity levels within each of the four CSE conditions (i.e., CC, II, CI and IC). Conflict similarity for CC and II conditions are reversed (multiplied by −1), such that for all the four CSE conditions, higher conflict similarity is expected to be associated with worse performance (see *Behavioral analysis* in Methods). Each dot represents a subject. The thin colored lines in B, C, E and F are the fitted lines for each of the four CSE conditions, and the thick black lines are the fitted lines collapsing across all CSE conditions. For panel C and F some similarity levels are missing because of the limited trial numbers in the experimental design in Experiment 2. CSE = congruency sequence effect; RT = reaction time; ER = error rate; CI = congruent (trial n−1)-incongruent (trial n); IC = incongruent-congruent; CC = congruent-congruent; II = incongruent-incongruent.

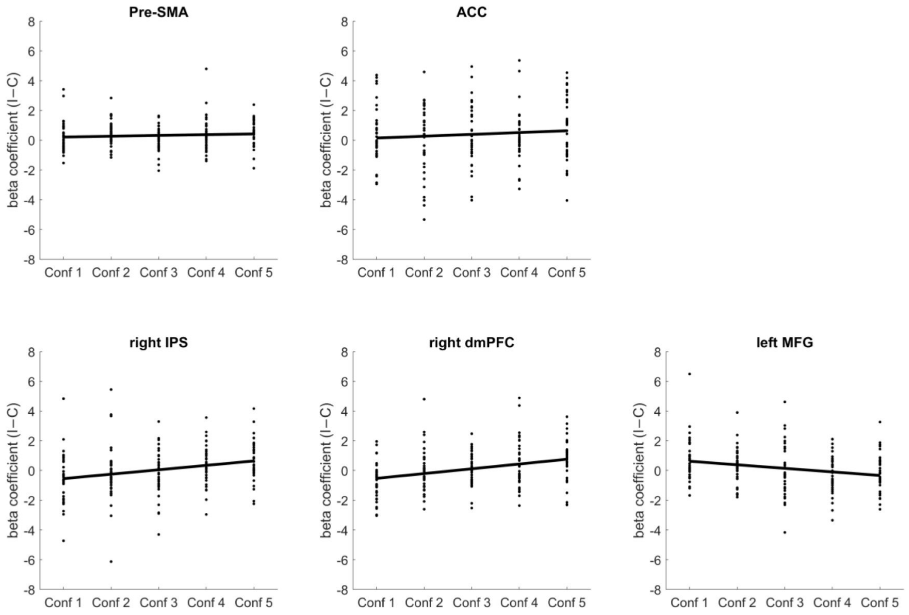
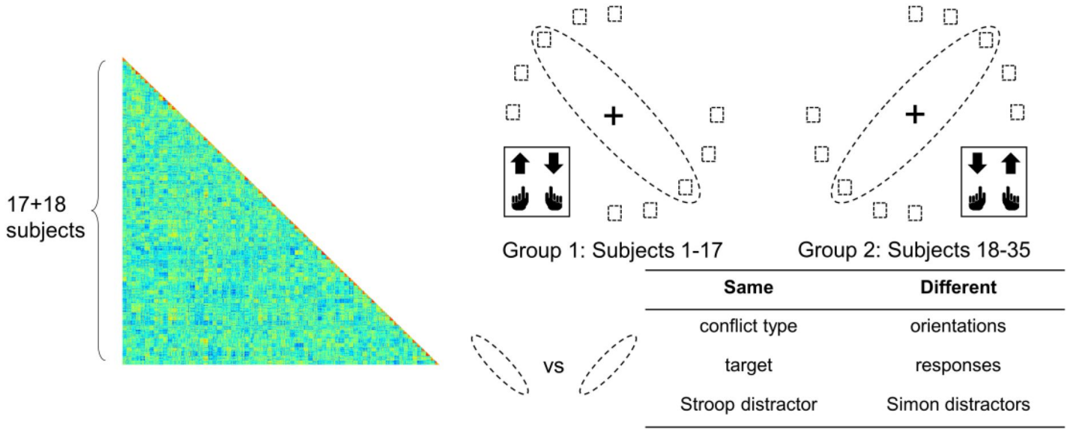


Fig. S3.

Neural congruency effect (I−C) by GLM2 [see the *Estimation of fMRI activity with univariate general linear model (GLM)* of Methods in the main text], plotted as a function of conflict type in different cortical ROIs. The ROIs were selected because they show a statistically significant congruency effects or parametric modulation effects when analyzed using the univariate GLM1.The pre-SMA and ACC showed overall congruency effects regardless of the conflict type (upper panel); the right IPS and right dmPFC were positively modulated by the conflict type and the left MFG was negatively modulated by the conflict type (lower panel). Conf 1 to 5 denotes the five conflict types, from the spatial Stroop to the Simon. Pre-SMA = pre-supplementary motor area; ACC = anterior cingulate cortex; IPS = inferior parietal sulcus; dmPFC = dorsomedial prefrontal cortex; MFG = middle frontal gyrus.



Same	Different
conflict type	orientations
target	responses
Stroop distractor	Simon distractors

$$RSM_{brain} = \beta_0 + \beta_1 RSM_{ConflictType} + \beta_2 RSM_{Orientation} + \beta_3 RSM_{Congruency} + \beta_4 RSM_{Target} + \beta_5 RSM_{Response} + \beta_6 RSM_{StroopDistractor} + \beta_7 RSM_{SimonDistractor} + \beta_8 RSM_{Run} + \beta_9 RSM_{Subject} + \beta_{10} RSM_{Group} + \beta_{11} RSM_{UnivoxelRow} + \beta_{12} RSM_{UnivoxelColumn} + \varepsilon$$

Fig. S4.

The cross-subject RSA model and the rationale. The RSM is calculated as the Pearson's correlation between each pair of conditions and the 35 subjects. For 17 subjects, the stimuli were displayed on the top-left and bottom-right quadrants, and they were asked to respond with left hand to the upward arrow and right hand to the downward arrow. For the other 18 subjects, the stimuli were displayed on the top-right and bottom-left quadrants, and they were asked to respond with left hand to the downward arrow and right hand to the upward arrow. Within each subject, the conflict type and orientation regressors were perfectly covaried. For instance, the same conflict type will always be on the same orientation. To de-correlate conflict type and orientation effects, we conducted the RSA across subjects from different groups. For example, the dashed ellipses highlight the conditions that are orthogonal to each other on the orientation representation, response, and Simon distractor, when their conflict type, target and spatial Stroop distractor are the same. The dashed boxes show the possible target locations for different conditions. RSM = representational similarity matrix.

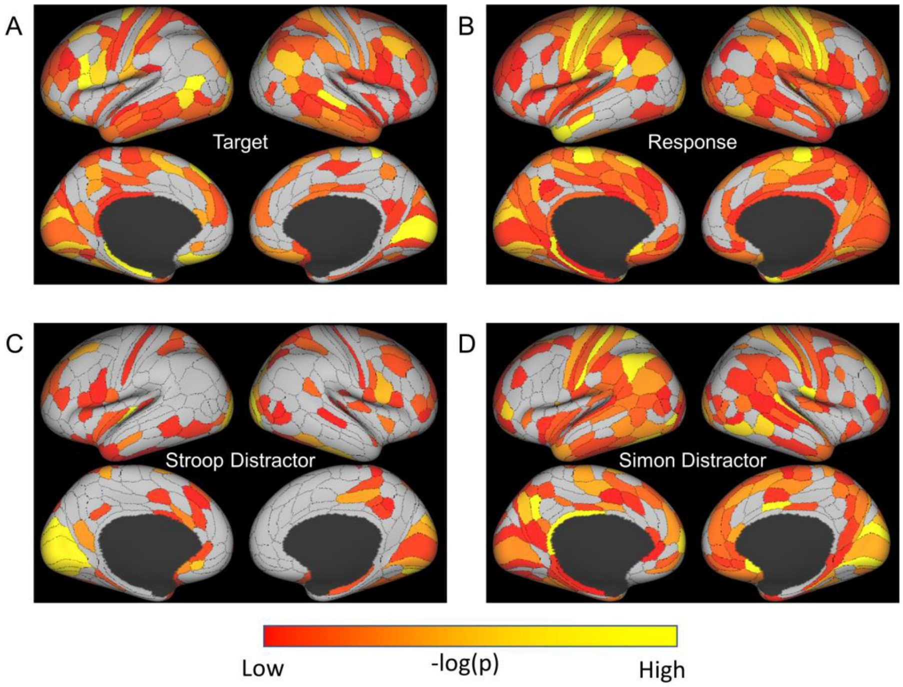


Fig. S5.

The cortical regions showing different effects in the main RSA. (A) The target effect reflects the above chance encoding of upward and downward arrow directions, and is most strongly encoded in the visual, memory and semantic regions, possibly because producing a goal-direct response to the stimulus require processing in all these regions. (B) the response effect reflects the above chance encoding of left and right responses, and is most strongly encoded in motor regions. (C) the spatial Stroop distractor effect reflects the above chance encoding of vertical location of the stimulus, and is most strongly encoded in left visual regions. (D) the Simon distractor effect reflects the above chance encoding of horizontal locations of the stimulus, and is most strongly encoded at the right visual regions, among others. All *p*-values are FDR-corrected ($p_{FDR} < 0.05$ and raw $p < 0.001$) across the 360 cortical ROIs. Brighter colors denote stronger effects as indicated by the opposite of log-transformed *p* values.

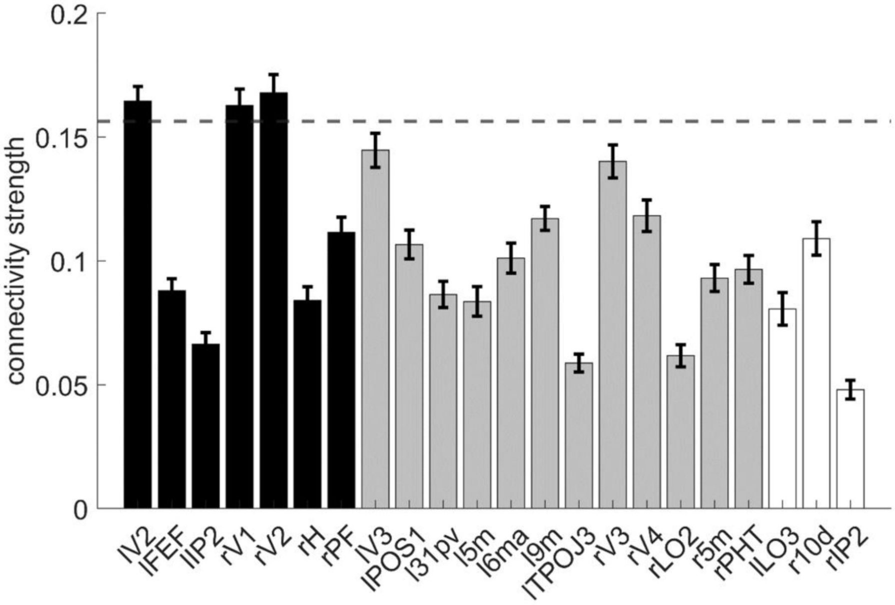


Fig. S6.

The representational connectivity between the right 8C area and the cortical regions showing significant encoding of orientation. The black bars represent regions showing both the overall orientation effect and higher encoding of orientation in incongruent than congruent conditions; the grey bars are regions showing only the overall orientation effect but not higher encoding of orientation in incongruent than congruent conditions; and the white bars are regions not showing any of the effects of interest (i.e., $q_{FDR} > 0.05$ for all the conflict type, orientation, congruency, target, response, spatial Stroop distractor and Simon distractor effects). The grey and white bars show controlled regions. Error bars are the standard error of the mean. The dashed line indicates the 95% confidence interval of the highest connectivity of controlled regions (i.e., left V3). l = left, r = right.

Supplementary Tables

Table S1.

Brain activations for the uni-voxel parametric analysis in GLM1 (voxel-wise one-tailed $p < .005$, cluster > 20)

Region	L/R	MNI coordinate (mm)			Volume (No. of voxels)	MaxZ	BA
		x	y	z			
incongruent > congruent							
Pre-supplementary motor area	R	12	12	73	60	4.07	6
Anterior cingulate cortex	L	−3	26	34	24	3.17	24
Positive parametric modulator (linear Simon effect)							
Inferior parietal sulcus	R	52	−64	33	68	3.06	39
Dorsomedial prefrontal cortex	R	15	57	42	43	2.96	9
Negative parametric modulator (linear spatial Stroop effect)							
Middle frontal gyrus	L	−27	45	25	29	3. 06	46

Notes. L = left; R = right; BA = Brodmann area.

References

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

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

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

4. 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** *J. Exp. Psychol. Hum. Percept. Perform* **43:**807–818
https://doi.org/10.1037/xhp0000351

5. Hazeltine E. , Lightman E. , Schwarb H. , Schumacher E.H. (2011) **The boundaries of sequential modulations: evidence for set-level control** *J. Exp. Psychol. Hum. Percept. Perform* **37:**1898–1914
https://doi.org/10.1037/a0024662

6. 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

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

8. 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

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

10. 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** *Cereb. Cortex* **30:**6336–6349
https://doi.org/10.1093/cercor/bhaa189

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

12. 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** *Nat. Neurosci* **16:**1348–1355
https://doi.org/10.1038/nn.3470

13. 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

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

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

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

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

18. 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** *Psychol. Res* **79:**143–162
https://doi.org/10.1007/s00426-014-0540-1

19. 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** *Neurosci. Biobehav. Rev* **83:**123–131
https://doi.org/10.1016/j.neubiorev.2017.09.032

20. 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* **376:**
https://doi.org/10.1126/science.abm9922

21. Braem S. , Abrahamse E.L. , Duthoo W. , Notebaert W. (2014) **What determines the specificity of conflict adaptation? A review, critical analysis, and proposed synthesis** *Front. Psychol* **5:**1134
https://doi.org/10.3389/fpsyg.2014.01134

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

23. 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

24. 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

25. 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

26. 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

27. Grahek I. , Leng X. , Fahey M.P. , Yee D. , Shenhav A. (2022) **Grahek, I., Leng, X., Fahey, M.P., Yee, D., and Shenhav, A. (2022). Empirical and Computational Evidence for Reconfiguration Costs During Within-Task Adjustments in Cognitive Control. In44.** *Empirical and Computational Evidence for Reconfiguration Costs During Within-Task Adjustments in Cognitive Control. In* **44:**

28. 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** *J. Exp. Psychol. Learn. Mem. Cogn* **47:**1705–1719
https://doi.org/10.1037/xlm0001054

29. 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

30. 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

31. 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:**102606
https://doi.org/10.1016/j.conb.2022.102606

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

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

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

35. Egner T. (2007) **Congruency sequence effects and cognitive control** *Cogn. Affect. Behav. Neurosci* **7:**380–390
https://doi.org/10.3758/cabn.7.4.380

36. 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

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

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

39. 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

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

41. 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:**2053
https://doi.org/10.1038/s41467-020-15928-z

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

43. Kriegeskorte N. , Mur M. , Bandettini P. (2008) **Representational similarity analysis - connecting the branches of systems neuroscience** *Front Syst Neurosci* **2:**4
https://doi.org/10.3389/neuro.06.004.2008

44. 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

45. 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

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

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

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

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

50. Milner B. (1963) **Effects of Different Brain Lesions on Card Sorting - Role of Frontal Lobes** *Arch Neurol-Chicago* **9:**90
https://doi.org/10.1001/archneur.1963.00460070100010

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

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

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

54. Friehs M.A. , Klaus J. , Singh T. , Frings C. , Hartwigsen G. (2020) **Perturbation of the right prefrontal cortex disrupts interference control** *NeuroImage* **222:**117279
https://doi.org/10.1016/j.neuroimage.2020.117279

55. 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

56. 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:**105103
https://doi.org/10.1016/j.cognition.2022.105103

57. 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

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

59. 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

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

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

62. 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 Cogn. Sci* **23:**769–783
https://doi.org/10.1016/j.tics.2019.07.002

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

64. 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

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

66. Esteban O. , Markiewicz C.J. , Blair R.W. , Moodie C.A. , Isik A.I. , Erramuzpe A. , Kent J.D. , Goncalves M. , DuPre E. , Snyder M. , et al. (2019) **fMRIPrep: a robust preprocessing pipeline for functional MRI** *Nat. Methods* **16:**111–116
https://doi.org/10.1038/s41592-018-0235-4

67. 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** *Front. Neuroinform* **5:**13
https://doi.org/10.3389/fninf.2011.00013

68. 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

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

70. Lieberman M.D. , Cunningham W.A. (2009) **Type I and Type II error concerns in fMRI research: re-balancing the scale** *Soc. Cogn. Affect. Neurosci* **4:**423–428
https://doi.org/10.1093/scan/nsp052

71. van Baar J.M. , Chang L.J. , Sanfey A.G. (2019) **The computational and neural substrates of moral strategies in social decision-making** *Nat. Commun* **10:**1483
https://doi.org/10.1038/s41467-019-09161-6

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

73. Genovese C.R. , Lazar N.A. , Nichols T. (2002) **Thresholding of statistical maps in functional neuroimaging using the false discovery rate** *NeuroImage* **15:**870–878
https://doi.org/10.1006/nimg.2001.1037

1. Freund MC , Etzel JA , Braver TS. (2021) **Neural Coding of Cognitive Control: The Representational Similarity Analysis Approach** *Trends Cogn Sci* **25:**622–638

2. Durnez J , Blair R , Poldrack RA. (2018) **Neurodesign: Optimal Experimental Designs for Task fMRI** *BioRxiv* **119594:**

3. Esteban O , et al. (2019) **fMRIPrep: a robust preprocessing pipeline for functional MRI** *Nat Methods* **16:**111–116

4. Gorgolewski K , et al. (2011) **Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python** *Front Neuroinform* **5:**13

5. Tustison NJ , et al. (2010) **N4ITK: improved N3 bias correction** *IEEE Trans Med Imaging* **29:**1310–1320

6. Avants BB , Epstein CL , Grossman M , Gee JC. (2008) **Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain** *Med Image Anal* **12:**26–41

7. Zhang Y , Brady M , Smith S. (2001) **Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm** *IEEE Trans Med Imaging* **20:**45–57

8. Fonov VS , Evans AC , McKinstry RC , Almli CR , Collins DL. (2009) **Unbiased nonlinear average age-appropriate brain templates from birth to adulthood** *NeuroImage* **47:**

9. Jenkinson M , Smith S. (2001) **A global optimisation method for robust affine registration of brain images** *Med Image Anal* **5:**143–156

10. Greve DN , Fischl B. (2009) **Accurate and robust brain image alignment using boundary-based registration** *NeuroImage* **48:**63–72

11. 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

12. Behzadi Y , Restom K , Liau J , Liu TT. (2007) **A component based noise correction method (CompCor) for BOLD and perfusion based fMRI** *NeuroImage* **37:**90–101

13. Lanczos C. (1964) **Evaluation of Noisy Data** *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis* **1:**76–85

14. Abraham A , et al. (2014) **Machine learning for neuroimaging with scikit-learn** *Front Neuroinform* **8:**14

15. Freund MC , Bugg JM , Braver TS. (2021) **A Representational Similarity Analysis of Cognitive Control during Color-Word Stroop** *J Neurosci* **41:**7388–7402

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

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 of central theoretical importance to cognitive control, and they make an interesting an interesting contribution to this question

Concerns
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.

2. 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.

3. 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).

4. 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.

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).

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).

5. 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).

Reviewer #2 (Public Review):

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 task-irrelevant 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. 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.

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

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.

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.

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.

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 voxel-level 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).

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).

A key source of potential bias comes from the fact that the off-diagonal cells are not independent (e.g., the correlation between subject A and B is strongly dependent on the correlation between subject A and C). For appropriate degrees of freedom calculation, the model must take this into account somehow. As fitted, the current models do not seem to handle this appropriately. That being said, it may be possible to devise an appropriate test via mixed-level models. In fact, Chen et al. have a series of three recent Neuroimage articles that extensively explore this question (all entitled "Untangling the relatedness among correlations") - adopting one of the methods described in the papers, seems much safer, if possible.

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.

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.

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.

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. However, several potential caveats need to be considered.

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.

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.

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.