Shared Neural Representations of Cognitive Conflict and Negative Affect in the Dorsal Anterior Cingulate Cortex

Luc Vermeylen[1*], David Wisniewski[1], Carlos González-García[1], Vincent Hoofs[1], Wim Notebaert[1], Senne Braem[1,2]

[1]Department of Experimental Psychology, Ghent University, Belgium

[2]Department of Experimental and Applied Psychology, Vrije Universiteit Brussel, Belgium

*Correspondence

Luc Vermeylen

Department of Experimental Psychology

Henri Dunantlaan 2

B – 9000 Ghent

BELGIUM

Email: Luc.Vermeylen@ugent.be

Abstract

Influential theories of dorsal anterior cingulate cortex (dACC) function suggest that the dACC registers cognitive conflict as an aversive signal, but no study directly tested this idea. In this pre-registered human fMRI study, we used multivariate pattern analyses to identify which regions respond similarly to conflict and aversive signals. The results show that, of all conflict- and value-related regions, only the dACC/pre-SMA showed shared representations, directly supporting recent dACC theories.

Main

The dACC has been implicated in various psychological processes such as cognitive control, somatic pain, emotion regulation, reward learning and decision making[1–3]. In the domain of cognitive control, dACC is consistently activated by cognitive conflict, that is, the simultaneous activation of mutually incompatible stimulus, task, or response representations[4]. It has been proposed that dACC generates a domain-general aversive learning signal which biases behavior away from costly information processing (e.g., conflict)[5–7]. Recent behavioral studies indeed demonstrated that humans dislike and tend to avoid conflict, and automatically evaluate conflict as aversive[8–10]. Similarly, it has been proposed that conflict and negative affect are integrated in the dACC[3,9,11]. Given these proposals and findings, one would expect conflict and negative affect to be encoded similarly in dACC ("shared representations").

One recent study tried to investigate this hypothesis using a repetition suppression procedure, and found that dACC showed an attenuated response to negative affect following cognitive conflict[12]. However, other studies failed to provide evidence for this idea. For example, a number of studies and meta-analyses demonstrated that distinct parts of the ACC are associated with

cognitive conflict and pain processing[13–16]. Similarly, a recent meta-analysis failed to observe overlap between cognitive control, pain processing, and (negative) emotion in the medial Prefrontal Cortex[17]. However, these previous studies often focus on peak activations across fMRI studies that differ in experimental control, or involve intense pain responses that could mask similarities with the arguably subtler affective evaluation of cognitive conflict.

Here, we took a different approach and developed a tightly controlled within-subjects test of shared neural representations of conflict and affect in the brain. Namely, by using multivariate cross-classification analyses, we assessed whether and where a classifier algorithm trained to discern conflict (incongruent vs congruent events) can successfully predict affect (negative vs positive events), and vice versa. Successful classification would indicate a similarity between the neural pattern response, and thus a shared representational code between these two domains[18,19]. Specifically, 38 human subjects performed a color Stroop[20] and flanker task[21] in the conflict domain, and two closely matched tasks in the affective domain (Fig. 1A). Importantly, we used two tasks in each domain in order to demonstrate an abstract representation of conflict (and affect), that is independent of conflict type (and affect source)[22]. Conflict and affect-related brain signals were used to perform a leave-one-run-out cross-classification analysis using a linear Support Vector Machine (see Methods). We performed preregistered Region of Interest (ROI) and whole brain searchlight analyses (Supplementary Table 1), and report accuracy-minus-chance values for each ROI and searchlight sphere (ROIs: Amygdala, Anterior Cingulate Cortex [ACC], dACC/pre-SMA, Anterior Insula [AI], Posterior Cingulate Cortex [PCC], Ventral Striatum [VS], and the ventromedial Prefrontal Cortex [vmPFC]).

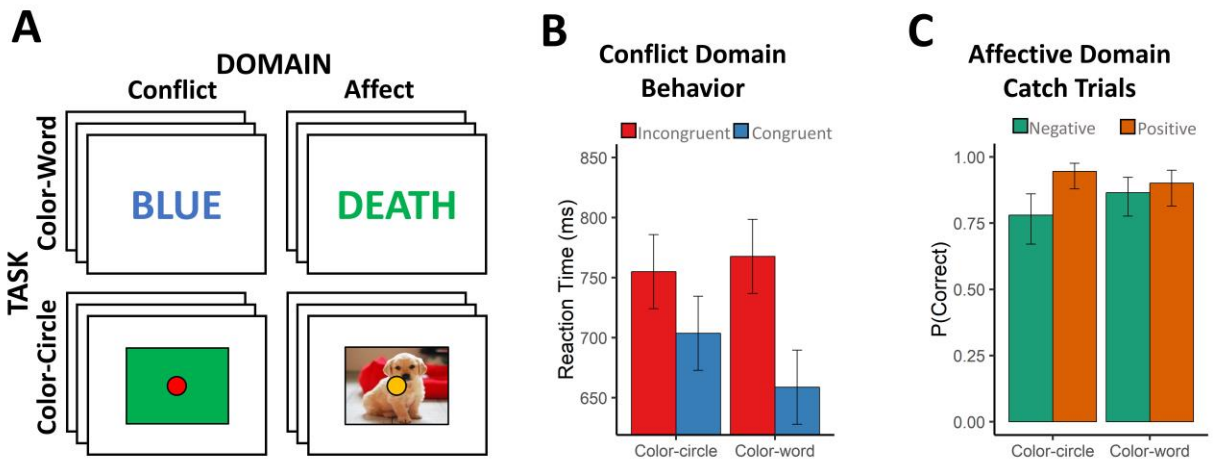

Figure 1. Task Design and Behavioral Data. **(A)** Task design. Subjects either judged the color of words or the color of circles. In the conflict domain, the color either matched or mismatched with word meaning or background color creating congruent or incongruent conditions, respectively. In the affective domain, positive or negative words and pictures were used to create the respective conditions. These four task contexts were presented block-wise. **(B)** In the conflict domain, typical congruency effects were found ($F_{(1,37)}$=148.81, p<.001, BF>100), which were larger in the color-word task ($F_{(1,37)}$=35.55, p<.001, BF>100). **(C)** On catch trials in the affective domain, subjects had to make a valence judgement (positive or negative) on the affective background stimuli.

The behavioral data (Fig. 1B and Supplementary Table 3) and univariate brain results (Supplementary Table 2) from the conflict tasks showed the typical differences between congruent and incongruent trials. In the affective tasks, catch trials (where subjects had to make a valence judgement instead of a color judgement) and a post-experiment incidental memory test were used to inform processing of the (task-irrelevant) affective stimuli (see Supplementary

Table 4 for behavioral results). We observed above-chance catch trial performance (chance level = 50%; see Fig. 1C and Methods) and successful post-experiment incidental recognition of the affective stimuli (Supplementary Figure 5), ensuring that subjects processed the affective pictures.

In a first set of multivariate pattern analyses, we trained and tested a classifier within-task (within the Stroop or flanker task; Fig. 2A, left panels; which regions respond to conflict within tasks?), as well as cross-task (train and test on different tasks; which regions respond similarly to conflict independent of low-level task features?), in each of our preregistered ROIs (for analysis details, see Method and Fig. 2B, left panels). Within-task ROI analyses in the conflict domain (congruent vs. incongruent) revealed evidence for above chance-level decoding in the dACC/pre-SMA (*Wilcoxon V*=327, *P*=.009, *BF10*=8.48), but not in in any of the other regions (all *P*>.060, *BF*<0.60) (Fig. 2A, right panel). This decoding accuracy in the dACC/pre-SMA did not differ by task ($F_{(1,37)}$=0.72, *P*=.400, *BF*=0.34). Second, the results show for the first time a conflict representation independent of conflict task as within-conflict cross-task ROI analyses revealed above-chance level conflict decoding in the dACC/pre-SMA (*V*=283, *P*=.012, *BF*=5.57). Again, decoding accuracy did not differ between cross-task combination (i.e., from flanker to Stroop or Stroop to flanker) ($F_{(1,37)}$=0.89, *P*=.352, *BF*=0.35) (Fig. 2B, right panel). These results were also replicated in an overall decoding approach where the classifier was trained and tested in the whole domain regardless of task (resulting in more samples to train the classifier; Supplementary Fig. 1A). Within the affective domain (positive vs. negative), we also performed these within- and cross-task decoding analyses. However, while these analyses showed evidence for affect information in the insula, they did not show evidence for decoding in the dACC/pre-SMA (Supplementary Fig. 2).

Finally, we evaluated our main hypothesis by training a classifier on discerning conflict (incongruent vs congruent) and testing its performance on discerning affect (negative vs positive), and vice versa. For this analysis, we focussed on the cross-domain cross-task decoding (train and test in different domains on different tasks) as this analysis also controls for more low-level shared features between the two tasks (Fig. 2C, right panel). The cross-domain cross-task ROI decoding revealed evidence for cross-classification in the dACC/pre-SMA (V=330, P=.007, BF=8.43; Fig. 2C, right panel), which did not differ by cross-task combination ($F_{(1,37)}$=0.36, P=.551, BF=0.29). None of the other ROIs reached significance (all Ps>.101). These results were replicated with the overall decoding approach in the main dACC/pre-SMA ROI (V=449, P=.021, BF=4.65; Supplementary Fig. 1C).

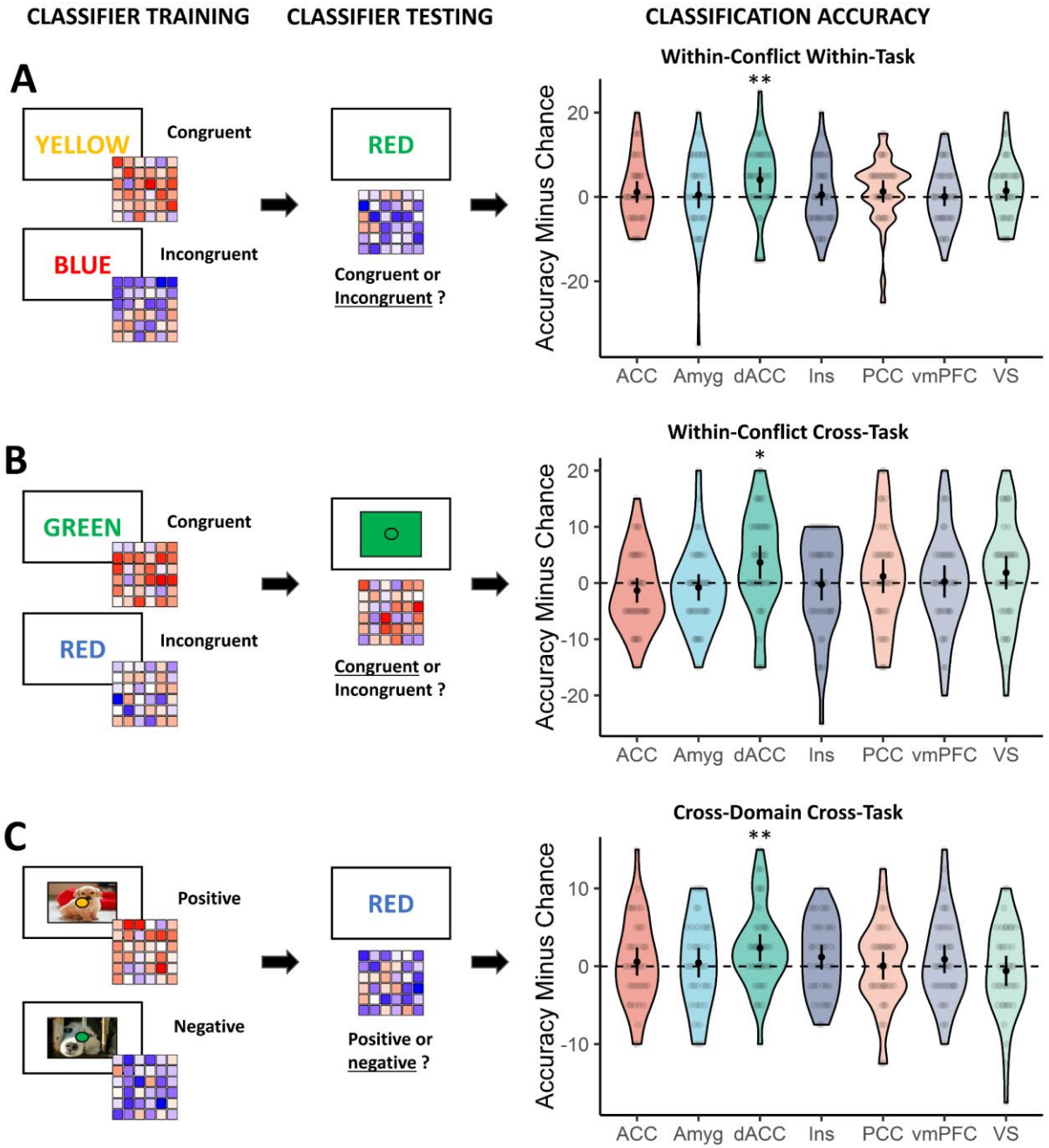

Figure 2. Main Results. **(A)** Training and testing the classifier within the conflict domain, within the same task. **(B)** Training the classifier on one conflict task and testing its performance on another conflict task. **(C)** Training the classifier to discern affect and testing its performance on classifying conflict across-tasks (and vice versa). **P*<.05; ***P*<.01; black dots and error bars

represent mean and ± 95 CI respectively; transparent dots represent individual data points; the shape of the violin shows the distribution of the data.

A number of control analyses further confirmed our main finding. First, we replicated this result using different smoothing parameters (Supplementary Fig. 3), or when using spherical ROIs instead of the Harvard-Oxford atlas ROIs (Supplementary Fig. 4). Second, also when using a set of functionally (rather than anatomically) defined conflict-sensitive ROIs based on a recent meta-analysis[23] (Supplementary Fig. 1, panel D), we again observed evidence for cross-domain cross-task classification in the dACC/pre-SMA (V=450, P=.013, BF=3.75) but not for other conflict-sensitive ROIs (left MOG, right AI, left AI, left IFG, left IPL, right IPL, left MFG), except for the left AI (V=425, P=.005, BF=8.61). The result again replicated when using the overall decoding approach in the dACC/pre-SMA (V=449, p=.001, BF=41.06), but not in the left AI (V=335, P=.260, BF=0.34).

Together, our results are the first to show that the dACC/pre-SMA shows a similar voxel pattern response to conflict and negative affect, and thereby offer important support for the popular proposal that the dACC registers conflict as an aversive signal[3,5,6], thought to bias behavior away from costly, demanding or suboptimal outcomes (as evidenced by behavioral avoidance and negative evaluation of conflict[8,9]).

Moreover, our study is also the first to show decoding of conflict across conflict tasks in the dACC, suggesting a shared component in the detection of conflict across the Stroop and flanker task[22]. The fact that we did not observe a similar (significant) above-chance decoding of affect in the dACC, but did observe cross-domain decoding, might seem surprising. However, this most

likely suggests differences in signal to noise ratio (SNR) between the two domains and does not invalidate the cross-domain decoding result[24]. A lower SNR in the affect domain can be explained by the fact that affect was not relevant for the main task.

The present findings also contradict the idea that cognitive control and affect are processed in distinct subdivisions of the ACC (e.g., dorsal-cognitive vs. ventral-emotional[14]). While the integration of cognitive control and affect in the dACC gained traction over the last two decades[3,25], direct evidence for this idea was lacking, and recent (meta-analytical) studies were more in line with the idea that both are processed in different subregions[13,17]. These studies were problematic for many theories of dACC functioning as these theories often hold the (implicit) assumption that dACC's response to suboptimal outcomes (e.g., conflict) has an evaluative component (e.g., signaling avoidance learning[3,5], expected value of control[6,7], value of the non-default option[26], evaluating action-outcome expectancies[27]). By using a tightly controlled within-subject design and multivariate analysis techniques, we now show that conflict and negative affect are indeed integrated in the dACC/pre-SMA, thereby providing important support for a more integrative view and current theories of dACC functioning.

Acknowledgements

We would like to thank Tobias Egner for valuable comments on a previous draft of the manuscript. W.N., S.B. (G.0660.17N) and L.V. (11H5619N) were supported by the FWO – Research Foundation Flanders. C.G.G. was supported by the Special Research Fund of Ghent University (BOF.GOA.2017.0002.03). D.W. was supported by the FWO (FWO.KAN.2019.0023.01), and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 665501. All procedures applied

in the present experiment were carried out with adequate understanding and written consent of the subjects and are in accordance with the Declaration of Helsinki.

Author Contributions

S.B. and W.N. developed the study concept. S.B., W.N. and L.V. contributed to the study design. Data collection was performed by L.V. and V.H.. Data analysis was performed by L.V. under the supervision of S.B., D.W. and C.G.C.. The manuscript was drafted by L.V. in cooperation with S.B., W.N., D.W. and C.G.C.. All authors approved the final version of the manuscript for submission.

Competing Interests

The authors have no competing interests to declare.

References

1. Ebitz, R. B. & Hayden, B. Y. *Nat. Neurosci.* **19**, 1278 (2016).
2. Heilbronner, S. R. & Hayden, B. Y. *Annu. Rev. Neurosci.* **39**, 149–170 (2016).
3. Shackman, A. J. *et al. Nat. Rev. Neurosci.* **12**, 154–167 (2011).
4. Botvinick, M. M. *et al. Psychol. Rev.* **108**, 624–652 (2001).
5. Botvinick, M. M. *Cogn. Affect. Behav. Neurosci.* **7**, 356–366 (2007).
6. Shenhav, A., Botvinick, M. M. & Cohen, J. D. *Neuron* **79**, 217–240 (2013).
7. Shenhav, A., Cohen, J. D. & Botvinick, M. M. *Nat. Neurosci.* **19**, 1286 (2016).
8. Dreisbach, G. & Fischer, R. **24**, 255–260 (2015).
9. Inzlicht, M., Bartholow, B. D. & Hirsh, J. B. *Trends Cogn. Sci.* **19**, 126–132 (2015).
10. Dignath, D., Eder, A. B., Steinhauser, M. & Kiesel, A. *Psychon Bull Rev.* In press.
11. Lieberman, M. D. & Eisenberger, N. I. *Proc. Natl. Acad. Sci.* **112**, 15250–15255 (2015).

12. Braem, S. *et al. J. Cogn. Neurosci.* **29**, 137–149 (2017).

13. Jahn, A., Nee, D. E., Alexander, W. H. & Brown, J. W. *J. Neurosci.* **36**, 12385–12392 (2016).

14. Bush, G., Luu, P. & Posner, M. I. *Trends Cogn. Sci.* **4**, 215–222 (2000).

15. Lieberman, M. D., Burns, S. M., Torre, J. B. & Eisenberger, N. I. *Proc. Natl. Acad. Sci.* **113**, E2476–E2479 (2016).

16. De La Vega, A., Chang, L. J., Banich, M. T., Wager, T. D. & Yarkoni, T. *J. Neurosci.* **36**, 6553–6562 (2016).

17. Kragel, P. A. *et al. Nat. Neurosci.* **21**, 283 (2018).

18. Kaplan, J. T., Man, K. & Greening, S. G. *Front. Hum. Neurosci.* **9**, 151 (2015).

19. Wisniewski, D. *Front. Psychol.* **9**, (2018).

20. Stroop, J. R. *J. Exp. Psychol.* **18**, 643 (1935).

21. Eriksen, B. A. & Eriksen, C. W. *Percept. Psychophys.* **16**, 143–149 (1974).

22. Jiang, J. & Egner, T. *Cereb. Cortex* **24**, 1793–1805 (2013).

23. Chen, T. *et al.* A *Brain Struct. Funct.* **223**, 3813–3840 (2018).

24. van den Hurk, J. & de Beeck, H. P. O. *bioRxiv* 592410 (2019).

25. Okon-Singer, H., Hendler, T., Pessoa, L. & Shackman, A. J. *Front. Hum. Neurosci.* **9**, 58 (2015).

26. Calhoun, A. J. & Hayden, B. Y. *Curr. Opin. Behav. Sci.* **5**, 24–31 (2015).

27. Brown, J. W. & Alexander, W. H. *J. Cogn. Neurosci.* **29**, 1656–1673 (2017).

Methods

Participants

The study was pre-registered with the pre-registration template from AsPredicted.org on the Open Science Framework (https://osf.io/p5frq/). As pre-registered, 40 participants participated in our study. Two participants were excluded (one due to excessive head motion [>2.5mm translation] and one aborted the scanning session). The average age of the remaining 38 participants (13 male) was 23.71 years (*SD*=3.53, min=18, max=33). Thirty-six participants were right-handed, one was left-handed and one was ambidextrous (as assessed by the Edinburgh Handedness Inventory[28]). Every participant had normal or corrected to normal vision and reported no current or history of neurological, psychiatric or major medical disorder. Every participant gave their informed written consent before the experiment, and was paid 35 euros for participating afterwards. The study was approved by the local ethics committee (University Hospital Ghent University, Belgium).

Experimental Paradigm

The experiment was implemented using Psychopy 2 version 1.85.2[29]. On each trial, participants had to judge the color of a target stimulus in the center of the screen, using two MR-compatible response boxes (each box had two buttons) to indicate one out of four possible response options (red, blue, green and yellow). The key-to-color mapping was counterbalanced between participants. The exact features of the target stimulus varied block-wise, depending on one of four different task-contexts. Specifically, participants either had to respond to the color of words ("color-word naming task") or respond to the color of circles ("color-circle naming task"), which both had a conflict and affective version.

The conflict-version of the color-word naming task was a Stroop task[20], where the meaning of the words could either be congruent or incongruent with the actual color of the word. For example, participants could see the words “BLUE”, “RED”, “GREEN” or “YELLOW” (Dutch: “ROOD”, “BLAUW”, “GROEN” or “GEEL”) presented in a blue, red, green or yellow font. The conflict version of the color-circle naming task was essentially a color-based variant on the Eriksen flanker task[21], where the irrelevant feature consisted of a colored background square which could either be congruent or incongruent with the color of the circle. Here, participants could see blue, red, green or yellow circles presented on a blue, red, green or yellow background square. In both tasks, half of the trials were congruent (e.g., “RED” in a red font; a red circle presented on a red square background) while the other half of the trials were incongruent (e.g., “RED” in a blue font; a red circle on a blue square background).

The affect-versions of the color-word naming and color-circle naming tasks made use of irrelevant affective words or pictures, respectively. In the color-word naming task, 16 positive and 16 negative words were presented[30] that were matched on arousal, power, age of acquisition, Dutch word frequency[31], word length and grammatical category (Noun, Adjective and Verbs). The affective picture distractors in the background of the color-circle naming task were retrieved from the OASIS database[32]. Sixteen positive and 16 negative pictures were presented that were matched on semantic category (Animals, Objects, People, Scenery) and arousal. This resulted in a total of eight conditions: congruent, incongruent, positive or negative trials, that either involved words or pictures/colored backgrounds. While our stimuli were matched on arousal, we also performed a control analysis where we trained a classifier to distinguish low versus high arousing stimuli (matched on valence) and tested its performance on distinguishing congruent versus incongruent stimuli (and vice versa). In contrast to our affect decoding results, this cross-

domain cross-task decoding was not significant in the dACC/pre-SMA (V=294, P=403, $BF10$=0.26).

Each trial started with a fixation sign ("+") that was presented for 3 to 6.5 seconds (in steps of 0.5 s; M=3.5 s; drawn from an exponential distribution). Next, the target stimulus was presented for 1.5 seconds (fixed presentation time regardless of RT). In order to increase the saliency of the irrelevant dimension (conflict and affect), the onset of the affective word or picture preceded the presentation of the target feature by 200 ms during which the color of the target feature (word or circle) was white.

Participants performed five scanning runs and during each run the subjects performed each of the four task contexts in separate blocks. The order of the four blocked task contexts was fixed within participant but counterbalanced between participants. Each block hosted 32 trials (16 congruent/positive and 16 incongruent/negative) which were presented in a pseudo-random fashion with the following restriction: neither relevant nor irrelevant features of the target stimulus could be repeated. This restriction was used to investigate confound-free congruency sequence effects (see [33]; but this was not the aim of the current study and will not be discussed further). In total, each participant made 640 trials (i.e., five runs of four blocks of 32 trials).

In each task context (block), we also included one catch trial (at random, but not in the first two or last two trials of each block). In these catch trials, the presentation of the task-irrelevant word, picture, or colored square would not be followed by the presentation of the target color, and remain on screen for three seconds. Participants were instructed that during these catch trials, when no color information was present in the relevant dimension, their goal was to judge the irrelevant dimension depending on the cognitive domain. In the conflict domain, participants had to respond to the meaning of the word ("RED", "BLUE", "GREEN" or "YELLOW") or to the

color of the background square (red, blue, green or yellow) by using the respective key that would be used to judge the relevant dimension. In the affective domain, participants had to judge the affective word or background picture as either positive or negative by pressing all keys once or twice (response mapping for positive and negative stimuli counterbalanced between participants). The purpose of these catch trials was to increase the saliency of the irrelevant dimension.

Before the scanning session, participants were welcomed and instructed to read the informed consent after which they started practicing the experimental paradigm. After the scanning sessions, participants performed an unannounced recognition memory test on old and new affective words and pictures. Here, participants had to indicate whether they had previously seen the word or picture in the experiment (old/new judgement). The new words were matched with the old words in terms of valence, arousal, power, age of acquisition, word length, frequency, grammatical category. The new pictures were matched on valence, arousal and semantic category. In both a behavioral (n = 20) and fMRI pilot (n = 20), we already established that participants showed adequate performance on both the main task and the recognition memory task. Finally, participants completed four questionnaires (Need for Cognition, Behavioral Inhibition/Activation Scale, Positive and Negative Affect Schedule, Barret Impulsivity Scale) and were thanked for their participation. No significant correlations between these questionnaire scales and cross-classification accuracies were found, so we do not report these results.

Behavioral Data Analysis

Behavioral analyses were performed in R (RStudio version 1.1.463, www.rstudio.com). For the reaction time (RT) analyses, we removed incorrect, premature (< 150 ms), and extreme responses (RTs outside 3 SD from each condition mean for each participant). This resulted in an

average of 94.42 % of the trials left for the RT analyses (*SD*=3.18, min=84.22, max=98.28). We conducted a repeated measures ANOVA on the reaction time and accuracy measure with the within-subject factors Condition (conflict domain: congruent vs. incongruent, affective domain: positive vs. negative) and Task (color-word naming vs. color-circle naming). We also assessed post-scanning recognition memory of affective stimuli with a probit generalized linear mixed effects model on the probability to say that the stimulus was 'old' with fixed effects for Experience (old vs. new), Valence (positive vs. negative) and Task Type (word vs. picture) and crossed random effects for Participant and Item. We also pre-registered some exclusion criteria based on behavioral performance. Participants with a mean RT outside 3 SD from the sample mean or a hit rate below 3 SD or 60 % (chance level=25 %) from the sample mean were excluded. Participants that performed poorly on the post-scanning recognition memory test, i.e., hit rate or false alarm rate outside 3 SD of the sample mean were also excluded. In the end, no exclusions based on task performance had to be made. While performance on catch trials was not a pre-registered exclusion criterion, we found that two participants responded on chance level in the catch trials of the affective domain (chance level=50 %, positive vs. negative judgement). Excluding these participants did not change our conclusions.

fMRI data acquisition

fMRI data was collected using a 3T Magnetom Trio MRI scanner system (Siemens Medical Systems, Erlangen, Germany), with a sixty-four-channel radio-frequency head coil. A 3D high-resolution anatomical image of the whole brain was acquired for co-registration and normalization of the functional images, using a T1-weighted MPRAGE sequence (TR=2250 ms, TE=4.18 ms, TI=900 ms, acquisition matrix=256 × 256, FOV=256 mm, flip angle=9°, voxel size=1 × 1 × 1 mm). Furthermore, a field map was acquired for each participant, in order to

correct for magnetic field inhomogeneities (TR=520 ms, TE1=4.92 ms, TE2=7.38 ms, image matrix=70 x 70, FOV=210 mm, flip angle=60°, slice thickness=3 mm, voxel size=3 x 3 x 2.5 mm, distance factor=0%, 50 slices). Whole brain functional images were collected using a T2*-weighted EPI sequence (TR=1730 ms, TE=30 ms, image matrix=84 × 84, FOV=210 mm, flip angle=66°, slice thickness=2.5 mm, voxel size=2.5 x 2.5 x 2.5 mm, distance factor=0%, 50 slices) with slice acceleration factor 2 (Simultaneous Multi-Slice acquisition). Slices were orientated along the AC-PC line for each subject.

fMRI data analysis

fMRI data analysis was performed using Matlab (version R2016b 9.1.0, MathWorks) and SPM12 (www.fil.ion.ucl.ac.uk/spm/software/spm12/). Raw data was imported according to BIDS standards (http://bids.neuroimaging.io/) and functional data was subsequently realigned, slice-time corrected, normalized (resampled voxel size 2 mm^3) and smoothed (full-width at half maximum of 8 mm). The preprocessed data was then entered into a first-level general linear model analysis (GLM), and subsequently into a multivariate pattern analysis (MVPA[34–37]). Results were analyzed using a mass-univariate approach. Although we pre-registered that we would not normalize and smooth the data for our classification analyses, we found that Signal-to-Noise Ratio (SNR) was significantly improved with these additional preprocessing steps (Supplementary Fig. 3A). In addition, an independent classification analysis (classifying left vs. right responses in primary motor cortex) showed that decoding accuracies were significantly higher with these additional preprocessing steps (Supplementary Fig. 3B). Knowing that decoding information in the PFC is notoriously difficult as decoding accuracies are close to chance (relative to decoding in occipitotemporal cortex[38]), and the finding that smoothing can and does often improve SNR and decoding performance[39–41], we decided to optimize our MVPA

346 analyses by decoding on normalized and smoothed data. For completeness, however, we also
347 depict the results from our main cross-classification analysis for different levels of smoothing
348 (FWHM 0, 4 and 8 mm; see Supplementary Fig. 3C).

349 First-level GLM analyses consisted of 5 identically modeled sessions (i.e., the five runs). Each
350 session consists of eight regressors of interest (for the eight conditions, see above), four block
351 regressors (to account for the blocked presentation of each combination of word versus picture
352 versions of the conflict versus affect tasks), two nuisance regressors (that model performance
353 errors and catch trials) and six movement regressors. The regressors were convolved with the
354 canonical HRF. The modeled duration of the regressors of interest (the eight conditions) and
355 nuisiance regressors (errors, catch trials) was zero, while the modeled duration of the block
356 regressors was equal to the length of the blocks.

357 Next, the beta images from the first-level GLM were submitted to leave-one-run-out decoding
358 scheme with 'The Decoding Toolbox'[42]using a linear support-vector classification algorithm
359 (C=1). We performed whole-brain searchlight decoding (sphere radius: 3 voxels; Supplementary
360 Table 1) as well as ROI decoding (see below for ROI methods). Cross-validation decoding was
361 conducted within the affective (positive vs. negative) and conflict (congruent vs. incongruent)
362 domain for each task separately ("within-domain within-task classification"). To assess the
363 generalizability of the classifier within the domain, we also conducted cross-classification
364 analyses where we trained the classifier on one task and tested its performance on the other task
365 for each task type combination (from color-circle naming to color-word naming and vice versa)
366 separately ("within-domain cross-task classification"). To investigate the generalizability of
367 these classifiers across the domain (our main hypothesis), we trained the classifier in the conflict
368 domain and tested its performance in the affective domain, and vice versa. We conducted these

analyses cross task type combinations (i.e., from color-circle naming to color-word naming, or from color-word naming to color-circle naming) to further control for low-level task features, following the same reasoning as the within-domain cross-task classification analyses. The results from these classification analyses were then averaged to return the cross-domain cross-task decoding results. For each of these three decoding analyses, we also ran ANOVAs to evaluate whether the result differed depending on the task (e.g., color-circle naming versus color-word naming) or task-to-task direction (i.e., from color-circle naming to color-word naming, or from color-word naming to color-circle naming). Finally, we also report an "overall decoding" analysis, where the classifier was trained across the two task types at once, thereby ignoring whether the event featured words or pictures/colored backgrounds.

Each classification analysis resulted in 'accuracy-minus-chance' decoding maps for each subject. These maps were then entered into a group second-level GLM analysis in SPM12. Here, a one-sample t-test determined which voxels show significant accuracy above chance level.

Next to MVPA, we also conducted classic univariate analyses. Here, we constructed a set of contrasts subtracting (A) positive from negative conditions and (B) congruent from incongruent conditions for (1) each task separately as well as across both tasks. These contrast images were then entered into a second-level analysis in which a one-sample t-test determined which voxels show significant activation for each contrast. We applied a statistical threshold of $p < 0.001$ (uncorrected) at the voxel level, and $p < 0.05$ (family-wise error corrected) at the cluster level on all analyses (Supplementary Table 2).

ROI analyses

As part of our pre-registered main analysis plan, we conducted ROI decoding analyses. We set out to study the Amygdala, Anterior Cingulate Cortex (ACC), dorsal Anterior Cingulate Cortex/pre-SMA (dACC/pre-SMA), Anterior Insula (AI), Parietal Cingulate Cortex (PCC), Ventral Striatum (VS), and the ventromedial PFC (vmPFC). All ROIs were obtained from the Harvard-Oxford cortical and subcortical structural atlases, thresholded at 25%. As the dACC ROI was not defined in the Harvard-Oxford atlas, we decided to retrieve this ROI from Neurosynth[43] by entering "dacc" as search term (returning 162 studies reporting 4547 activations). Although this ROI was based on the "dacc" search term, the peak effect of studies reporting dACC activity actually lies more dorsally than the cingulate gyrus, overlapping with the pre-SMA[11]*. Therefore, we refer to this ROI as the dACC/pre-SMA. Next, we built a 10 mm sphere around the peak activation point in this activation map (association map). Because the dACC ROI was spherical (in contrast to the other six atlas ROIs), we also re-analyzed our results from the atlas ROIs with 10 mm spherical alternatives retrieved from Neurosynth, which returned highly similar results and did not change our statistical conclusions.

In addition to the pre-registered ROI analyses which were based on anatomically determined ROIs, we also ran a second set of ROI analyses with functionally informed ROIs. Namely, we created 10 mm sphere ROIs for all conflict-sensitive regions based on the most recent and inclusive meta-analysis we could find on cognitive conflict[23].

Each ROI decoding analysis returned one accuracy-minus-chance value per ROI and participant. We tested whether these values were significantly higher than zero (one-tailed) with the non-parametric Wilcoxon signed-rank test and a Bayesian t-test (using the default priors from the BayesFactor package in R; Cauchy prior width: r=.707). We report the Bayes Factor (BF) that quantifies the evidence for the alternative hypothesis (i.e., decoding accuracy is higher than

zero). Our pre-registered stopping criterion was if the main finding was *BF*>6 (i.e., or if we had reached 40 subjects, for financial reasons), but we would like to note that, if so, this result was typically also $p<.00714$, which is the Bonferroni-corrected alpha for the main set of 7 ROIs. Finally, we investigated whether the significant cross-task cross-domain classification accuracy correlated with the following behavioral indices: post-scanning affective recognition memory (d-prime), congruency sequence effects in reaction time and error rate and congruency sequence effects in reaction time and error rates (p-values of reported correlations are Holm-corrected for five tests) (see Supplementary Figure 5).

Data Availability

The minimal data necessary to replicate the reported findings can be found on the Open Science Framework (https://osf.io/p5frq/). Raw fMRI data and preprocessing scripts will be uploaded to a repository in the near future.

Code Availability

The custom code used for the analyses of this study can be found on the Open Science Framework (https://osf.io/p5frq/).

References

28. Oldfield, R. C. *Neuropsychologia* **9**, 97–113 (1971).

29. Peirce, J. W. *J. Neurosci. Methods* **162**, 8–13 (2007).

30. Moors, A. *et al. Behav. Res. Methods* **45**, 169–177 (2013).

31. Keuleers, E., Brysbaert, M. & New, B. *Behav. Res. Methods* **42**, 643–650 (2010).

32.Kurdi, B., Lozano, S. & Banaji, M. R. *Behav. Res. Methods* **49**, 457–470 (2017).

33. Braem, S. *et al. Trends Cogn. Sci.* **23**, 769–783 (2019).

435 34. Cox, D. D. & Savoy, R. L. *NeuroImage* **19**, 261–270 (2003).

436 35. Kriegeskorte, N., Goebel, R. & Bandettini, P. *Proc. Natl. Acad. Sci.* **103**, 3863–3868 (2006).

437 36. Haxby, J. V. *NeuroImage* **62**, 852–855 (2012).

438 37. Haynes, J.-D. *Neuron* **87**, 257–270 (2015).

439 38. Bhandari, A., Gagne, C. & Badre, D. *J. Cogn. Neurosci.* **30**, 1473–1498 (2018).

440 39. Kamitani, Y. & Sawahata, Y. *NeuroImage* **49**, 1949–1952 (2010).

441 40. Hendriks, M. H. A., Daniels, N., Pegado, F. & Op de Beeck, H. P. *Front. Neurol.* **8**, (2017).

442 41. Op de Beeck, H. P. *NeuroImage* **49**, 1943–1948 (2010).

443 42. Hebart, M. N., Görgen, K. & Haynes, J.-D. *Front. Neuroinformatics* **8**, (2015).

444 43. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Essen, D. C. V. & Wager, T. D. *Nat. Methods* **8**,

445 665–670 (2011).

446