Dynamic Transitions between Neural States Are Associated with Flexible Task Switching during a Memory Task

Wei Liu1,2, Nils Kohn2, and Guillén Fernández2

Abstract

Flexible behavior requires switching between different task conditions. It is known that such task switching is associated with costs in terms of slowed RT, reduced accuracy, or both. The neural correlates of task switching have usually been studied by requiring participants to switch between distinct task conditions that recruit different brain networks. Here, we investigated the transition of neural states underlying switching between two opposite memory-related processes (i.e., memory retrieval and memory suppression) in a memory task. We investigated 26 healthy participants who performed a think/no-think task while being in the fMRI scanner. Behaviorally, we show that it was more difficult for participants to suppress unwanted memories when a no-think was preceded by a think trial instead of another no-think trial. Neurally, we demonstrate that think–no-think switches were associated with an increase in control-related and a decrease in memory-related brain activity. Neural representations of task condition, assessed by decoding accuracy, were lower immediately after task switching compared with the nonswitch transitions, suggesting a switch-induced delay in the neural transition toward the required task condition. This suggestion is corroborated by an association between condition-specific representational strength and condition-specific performance in switch trials. Taken together, we provided neural evidence from the time-resolved decoding approach to support the notion that carryover of the previous task set activation is associated with the switching cost, leading to less successful memory suppression.

INTRODUCTION

In everyday life, we are continuously switching between different tasks (Monsell, 2003). Transitions between task conditions have often been studied using task-switching paradigms in which participants are required to switch between two or more distinct tasks (Meiran, 2010). Usually, participants perform less accurately and/or more slowly immediately after switches (i.e., switch costs; Goschke, 2000; Rogers & Monsell, 1995; Spector & Biederman, 1976; Jersild, 1927). Results from univariate fMRI studies suggested the involvement of prefrontal parietal regions in task switching (Richter & Yeung, 2014; Gruber, Karch, Schlueter, Falkai, & Goschke, 2006; Braver, Reynolds, & Donaldson, 2003; Dove, Pollmann, Schubert, Wiggins, & von Cramon, 2000). Two studies (Loose, Wisniewski, Rusconi, Goschke, & Haynes, 2017; Waskom, Kumaran, Gordon, Rissman, & Wagner, 2014) used multivariate fMRI methods (Cohen et al., 2017; Haynes, 2015) to investigate how task switching modulates neural representations of the current task condition but reported mixed results. Waskom et al. (2014) reported stronger task representations after a shift in task conditions, whereas Loose et al. (2017) found no difference in task representations between the switch and the nonswitch condition. Therefore, it is still unclear whether task switching will strengthen or weaken the task representation and how the altered task representation links to switch costs.

Previous experiments investigating task switching were typically designed to minimize the perceptual differences between conditions but not to maximize differences in underlying task conditions (Loose et al., 2017; Waskom et al., 2014; Kiesel et al., 2010; Braver et al., 2003). For example, one task could be to judge the relative size of the presented object compared with a computer monitor (large, e.g., truck; small, e.g., carrot). The other task would be to judge whether the object is manmade (e.g., truck) or natural (e.g., carrot; Braver et al., 2003). However, behavioral and neural correlates of task switching between two opposite tasks within one cognitive domain remain largely unexplored. We reasoned that switching between two opposite task conditions within one cognitive domain is a special case of task switching that needs to be investigated because the same set of networks must reconfigure swiftly in how they interact with each other (e.g., cooperation or competition between the very same networks). In this study, we investigated task switching between memory retrieval and suppression and its associated neural processes in a memory task. Cognitive and neural models of memory retrieval and suppression suggest that successful retrieval could be the result of “cooperation” between an inhibitory control network and an episodic retrieval network (Rugg & Vilberg, 2013), whereas effective suppression depends on top–down control of the inhibitory control network.
network upon an episodic retrieval network (i.e., competition; Anderson & Hanslmayr, 2014). Previous task fMRI studies of memory suppression supported this idea by showing that, compared with think trials, no-think trials are associated with stronger activation in control-related regions, including the dorsolateral pFC (DLPFC), ventrolateral pFC, inferior parietal lobule (IPL), and SMA (W. Liu, Peeters, Fernández, & Kohn, 2020; Guo, Schmitz, Mur, Ferreira, & Anderson, 2018; Anderson et al., 2004). At the same time, these activity increases are accompanied by reduced activity in memory-related areas in the medial-temporal lobe, including the hippocampus (Anderson & Hanslmayr, 2014). A recent resting-state fMRI study showed that individual differences in memory suppression ability could be predicted by the Internetwork communication of the inhibitory control network during the task-free condition (Yang et al., 2021).

Here, we used a modified think–no-think (TNT) paradigm (Levy & Anderson, 2012; Anderson & Green, 2001) to probe task switching between memory retrieval and suppression. Specifically, participants were instructed to switch between memory retrieval and memory suppression according to trial-specific instructions. We asked whether we can find behavioral “switch costs” (i.e., less optimal memory performance) when participants switch between two opposite memory tasks. If switch costs exist in our memory task, we would like to detect the neural source of switch costs. Previous cognitive theories of task switching propose that switch costs could result from the carryover of previous task set activation and depends on cognitive resources required to reconfigure the task set (Monsell, 2003). These cognitive theories cannot be directly tested without the development of a series of multivariate methods to probe neural representation in noninvasive human brain imaging data (Kriegeskorte & Diedrichsen, 2019). Here, we used a time-resolved multivariate decoding approach to capture the dynamic transitions between neural states (i.e., “think”: cooperation between memory and control network; “no-think”: competition between memory and control network) during task switching in fMRI data. We hypothesized that a delayed transition between neural states that represent task conditions could be the neural underpinning of behavioral switch costs, because failing to update neural states on time could result in a neural state that is optimal for the opposite (e.g., retrieval), but not the current (e.g., suppression), task condition. This assumption is built on the idea that the human brain can demonstrate diverse brain states during different cognitive tasks or environmental demands and whether the brain can properly reconfigure its state is behavioral relevant (Cocuzza, Ito, Schultz, Bassett, & Cole, 2020; Shine et al., 2016, 2019; Shine & Poldrack, 2018; Westphal, Wang, & Rissman, 2017; Gonzalez-Castillo et al., 2015; Sadaghiani, Poline, Kleinschmidt, & D’Esposito, 2015; Hermans et al., 2011). The task-switching paradigm is suitable to study such a rapid neural reconfiguration process, because it allows us to compare directly how different task conditions are represented in neural states and how transitions of neural states are associated with human behaviors (i.e., switch costs).

**METHODS**

**Participants**

In total, 32 right-handed, healthy young participants recruited from the Radboud Research Participation System finished all of the experimental procedures. All of them are native Dutch speakers. Six participants were excluded from data analyses because of low memory performance (i.e., lower chance level [25%] during the final memory test; \( n = 2 \)) or excessive head motion (\( n = 4 \)). We used the motion outlier detection program within the FMRIB Software Library (i.e., *FSLMotionOutliers*) to detect time points with large motion (threshold = 0.9). There are at least 20 spikes detected in these excluded participants, with the largest displacement ranging from 2.6 to 4.3, whereas included participants had fewer than 10 spikes. Finally, 26 participants (15 women, age = 19–30 years, mean = 23.51 years, SD = 3.30 years) were included in the behavioral and neuroimaging analysis reported in this study. According to a recent meta-analysis of TNT studies, usually, 20–30 participants were analyzed (W. Liu, Peeters, et al., 2020). The sample size of this study was determined by previous fMRI studies using the TNT paradigm, and the final number of participants analyzed was comparable to previous studies. Because of the reconstruction error during the data acquisition, one run of one participant is not complete (20–30 images were missing). Therefore, that run was not included in our analysis of time series. However, unaffected acquired images of that run were used in our univariate activation analysis. No participants reported any neurological and psychiatric disorders. We further used the Dutch version of the Beck Depression Inventory (Roelofs et al., 2013) and State–Trait Anxiety Inventory (van der Bij, de Weerd, Cikot, Steegers, & Brespenning, 2003) to measure the participants’ depression and anxiety level during scanning days. No participant showed a sign of emotional problems (i.e., their Beck Depression Inventory and State–Trait Anxiety Inventory scores are within the normal range). The experiment was approved by and conducted in accordance with the requirements of the local ethics committee (Commissie Mensgebonden Onderzoek, Regio Arnhem-Nijmegen, The Netherlands) and the declaration of Helsinki, including the requirement of written informed consent from each participant before the beginning of the experiment. Each participant received €10/hr for their participation.

**Experiment Design**

This experiment is a 2-day fMRI study, with a 24-hr delay between two sessions (Appendix Figure A1). fMRI data of
the Day 2 final memory test have been published in another publication (W. Liu, Kohn, & Fernández, 2020), and the comprehensive reports of the experimental materials and design can be found there. Because all of the behavioral and neuroimaging data included in this study came from the Day 2 session, we just presented a brief description of the Day 1 session. On Day 1, we instructed participants to memorize a series of sequentially presented location–picture associations, for which 48 distinct photographs were presented together with 48 specific locations on two cartoon maps. In our experiment, we used picture (i.e., location)–picture pairs as to-be-remembered materials instead of word–word or word–picture pairs to keep visual processes during scanned tasks largely consistent. More specifically, whole maps were presented sequentially, highlighting specific locations by colored frames as memory cues. Therefore, at the perceptual level, participants were always processing the same two maps. All photographs can be assigned into one of the four categories, including animal, human, scene (e.g., train station), and object (e.g., pen and notebooks). Therefore, objective memory performance could be assessed within the scanner by instructing participants to indicate the picture’s category when cued by the map location. During this study phase, each location–picture association was presented twice, and the learning was confirmed by two typing tests outside the scanner. During the typing tests, participants were required to describe the photograph associated with the memory cue in one or two sentences. Immediately after the study phase (Day 1), 88.01% of the associated pictures were described correctly (SD = 10.87%, range = 52%–100%).

On Day 2, participants first performed the second typing test and still recalled 82.15% of all associations (SD = 13.87%, range = 50%–100%). Then, they performed the TNT task and final memory test inside the scanner. We used the TNT task with a trial-by-trial performance rating to monitor the retrieval or suppression of each trial. Compared with the original TNT task (Anderson et al., 2004), the additional self-report did not affect the underlying memory suppression process and also was used in a neuroimaging experiment before (Levy & Anderson, 2012). Forty-eight picture–location associations were divided into three conditions (i.e., “think or retrieval,” “no-think or suppression,” and “baseline or control” conditions) in a counterbalanced way. Therefore, for each association, the possibility of belonging to one of the three conditions is equal. For each map, 24 locations that were distributed evenly across the map were paired with six pictures from each category. One third of associations (eight associations; two pictures from each category) on that map were retrieval associations (i.e., “think” associations), one third of associations were suppression associations (i.e., “no-think” associations), and the remaining one third were control associations. The spatial distribution of the three conditions was determined manually by the experimenter to prevent clusters of specific conditions on certain parts of the maps. Control associations only presented during the final memory test but not during the TNT phase. During the retrieval condition of the TNT phase, locations were highlighted with the green frame for 3 sec, and participants were instructed to recall the associated picture quickly and actively and to keep it in mind until the map disappeared from the screen. By contrast, during the suppression condition, locations were highlighted with the red frame for 3 sec, and our instruction for participants was to prevent the potential memory retrieval and try to keep an empty mind. We gave additional instructions for the suppression condition: “When you see a location, highlighted with a RED frame, you should NOT think about the associated picture. Instead, you should try to keep an empty mind during this stage. It is a difficult task, and it is totally fine that sometimes you still think about the associated picture. But please do NOT close your eyes, focus on something outside the screen, or think about something else in your life. These strategies, although useful, could negatively affect the brain activity that we are interested in….” After each trial, participants had a maximum of 3 sec to press the button on the response box to indicate whether and how often the associated picture entered their mind during think or no-think trials. Specifically, they rated their experience from 1 to 4, representing from no recall (i.e., never) to always recall. Responses during think trials were used as retrieval frequency ratings, whereas responses during no-think trials were regarded as intrusion frequency ratings. The TNT task included five functional runs, with 32 retrieval trials and 32 suppression trials per run. All “retrieval” or “suppression” associations were presented twice within one run but not next to each other. Therefore, they were presented 10 times during the entire TNT task. Between each trial, fixation was presented for 1–4 sec (mean = 2 sec, exponential model) as the intertrial intervals.

To investigate the task switching within the TNT task, for each run of each participant, we predefined the sequence of task conditions to form “blocks” of memory retrieval or suppression with the length range from 1 to 4 trials (mean = 1.9 trials, std = 1.01 trials, P_{one-trial block} = 46.875%, P_{two-trials block} = 25%, P_{three-trials block} = 18.75%, P_{four-trials block} = 9.375%). In this sequence, the task condition of the current trial can be the same as the previous trial (“non-switch” trial) or differ from the previous trial (“switch” trial). Within one run of a total of 64 trials, 31 trials were “non-switch” trials, 32 trials were “switch” trials, and the first trial cannot be labeled as a “non-switch” or “switch” trial because it has no predecessor. The “non-switch” trials and “switch” trials both accounted for around 50% of the “retrieval” and “suppression” trials. After determining the sequence of task conditions, specific location–picture associations from retrieval or suppression condition were randomly selected for each trial.

After the TNT task, a final memory test was performed by participants within the scanner to evaluate the effect of different modulations on memory. All 48 memory cues
(i.e., locations), including retrieval, suppression, and control conditions, were presented again with the duration of 4 sec by highlighting a certain part of the map with a blue frame. Participants were instructed to recall the associated picture as vividly as possible during the presentation and then give the responses on two multiple-choice questions within 7 sec (3.5 sec for each question). The first one is the measure of subjective memory: “How confident are you about the retrieval?” Participants had to rate from 1 to 4, representing “cannot recall, low confident, middle confident, and high confident” separately. The second one is the measure of objective memory: “Please indicate the category of the picture you were recalling.” They needed to choose from four categories (i.e., animal, human, scene, and object). It is notable that we only analyzed the behavioral data from this within-scanner memory test; the neural activity during this test is not the focus of this study.

Behavioral Data Analysis

Behavioral results of this project were comprehensively reported in another study of our laboratory with the focus on the final memory test (W. Liu, Kohn, et al., 2020). No results of task switching (i.e., switch costs) were reported in that study, and task switching is the central scientific question of this study. First, we analyzed the behavioral performance during the TNT task. Trial-by-trial performance reports from each participant were used to calculate the percentage of successful recall chosen across 160 retrieval trials and successful suppression across 160 suppression trials. Following previous studies (W. Liu, Kohn, et al., 2020; Levy & Anderson, 2012), performance reports from suppression trials were used to quantify individual differences in memory suppression efficiency ("intrusion slope score"). To account for the individual differences in memory performance before the TNT, we restricted the analysis of suppression into the associations for which participants can still remember during the second typing test (“remembered associations”). We used linear regression to model the relationship between intrusion frequency ratings of “remembered associations” and the number of repetitions of suppression at the individual level. Participants with more negative slope scores are better at down-regulating memory intrusions than those with less negative slope scores. Furthermore, we labeled each trial as the “nonswitch” trial or “switch” trial based on whether the task condition of the current trial is the same as the previous trial. Trial-by-trial performance between “switch” and “nonswitch” trials during retrieval or suppression was compared using paired t tests.

We also quantified the individual differences in “suppression-induced forgetting effect” based on two types of participants’ performance (i.e., recall accuracy and confidence rating) during the final memory test. Memory performance for associations that belong to the control condition was regarded as the baseline to quantify the suppression-induced forgetting effect. For each participant, recall accuracy (objective memory measure) and confidence rating (subjective memory measure) were calculated for no-think associations and control associations separately. Then objective and subjective suppression scores were computed separately by subtracting the accuracy and confidence of no-think associations from the control associations. The more negative a suppression score is, the stronger the suppression-induced forgetting effect is. The memory suppression score was used to correlate with the “intrusion slope score” and transition of neural states during the TNT.

MRI Data Acquisition and Preprocessing

We used a 3.0-T Siemens PrismaFit scanner (Siemens Medical) and a 32-channel head coil system at the Donders Institute Centre for Cognitive Neuroimaging in Nijmegen, The Netherlands, to acquire MRI data. For each participant, MRI data were acquired on two MRI sessions (around 1 hr for each session) with 24-hr intervals. In this study, we only used the data from the Day 2 session. Specifically, we acquired a 3-D magnetization-prepared rapid gradient echo anatomical T1-weighted scan for the registration purpose with the following parameters: 1 mm isotropic, echo time (TE) = 3.03 msec, repetition time (TR) = 2300 msec, flip angle = 8°, field of view = 256 × 256 × 256 mm. All functional runs were acquired with EPI-based multiband sequence (acceleration factor = 4) with the following parameters: 68 slices (multislice mode, interleaved), voxel size = 2 mm isotropic, TR = 1500 msec, TE = 39 msec, flip angle = 75°, field of view = 210 × 210 × 210 mm. In addition, to correct for distortions, magnitude and phase images were also collected (voxel size = 2 × 2 × 2 mm, TR = 1020 msec, TE = 12 msec, flip angle = 90°).

We used FEAT (FMRI Expert Analysis Tool) Version 6.00, part of FSL (FMRIB’s Software Library, www.fmrib.ox.ac.uk/fsl; Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012), together with Automatic Removal of Motion Artifacts (ICA-AROMA; Pruim et al., 2015), to perform our preprocessing. This pipeline was based on procedures suggested by Mumford and colleagues (mumfordbrainstats.tumblr.com) and the article that introduced the ICA-AROMA (Pruim et al., 2015). Specifically, we first removed the first four volumes of each run from the 4-D sequences for the stabilization of the scanner and then applied the following prestatistics processing: (1) motion correction using MCFLIRT (Jenkinson, Bannister, Brady, & Smith, 2002), (2) field inhomogeneities were corrected using B0 Unwarping in FEAT, (3) nonbrain removal using BET (Smith, 2002), (4) grand mean intensity normalization of the entire 4-D data set by a single multiplicative factor, and (5) spatial smoothing (6-mm kernel). ICA-AROMA was used to further remove motion-related spurious noise. We chose to conduct “nonaggressive denoising” and applied high-pass temporal filtering (Gaussian-weighted
least-squares straight-line fitting with sigma = 50.0 sec) before the following analyses.

All of the mentioned preprocessing steps were performed in native space. We used the following steps to perform the registration between native space, the participant’s high-resolution T1 space, and standard space. First, we used the boundary-based registration (Greve & Fischl, 2009) to register functional data to the participant’s high-resolution structural image. Next, registration of high-resolution structural to standard space was carried out using FLIRT (Jenkinson et al., 2002; Jenkinson & Smith, 2001) and was then further refined using FNIRT nonlinear registration (Andersson, Jenkinson, & Smith, 2007). The resulting parameters were used to align processed functional images from native space to standard space for the following signal extraction.

Univariate General Linear Model Analyses

We ran the voxel-wise general linear model analyses of the TNT task to identify brain regions that are more active during memory suppression compared with memory retrieval (i.e., no-think vs. think). All time-series statistical analysis was carried out using FILM with local autocorrelation correction (Woolrich, Ripley, Brady, & Smith, 2001) using FEAT. In total, three regressors were included in the model. We modeled the presentation of memory cues (locations) as two kinds of regressors (duration = 2 TR; i.e., suppression trials and retrieval trials). For each participant, based on the second typing test, which was immediately before the TNT phase, we separately modeled the location–picture associations, which participants cannot describe (i.e., unlearned associations) as a separate regressor. Because these location–picture associations were unlearned/forgotten, there were no memories to be recalled or suppressed during the TNT phase. We conducted the two contrasts of interest (i.e., no-think vs. think and think vs. no-think) first at the native space and then aligned resulting statistical maps to Montreal Neurological Institute (MNI) space using the parameters from the registration. These aligned maps were first used for participant-level averaging across five TNT runs and then the group-level analyses. The group-level statistical map was corrected for multiple comparisons using default cluster-level correction within FEAT (voxelwise Z > 3.1, cluster-level p < .05 family-wise error rate (FWER)-corrected).

Networks-of-interest Identification

To identify our networks of interest (i.e., inhibitory control network and memory retrieval network), we performed several term-based meta-analyses using Neurosynth (https://neurosynth.org/; Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). “Inhibitory control” and “memory retrieval” were used as terms separately to search for all studies in the Neurosynth database whose abstracts include the input term at least once. Then, all identified studies were combined separately for each term to generate the corresponding statistical map. We used uniformity test maps in our study. This method tested whether the proportion of studies that report activation at a given voxel differs from the rate that would be expected if activations were uniformly distributed throughout the gray matter. voxel-wise Z score from the one-way ANOVA testing was saved in a statistical map. Each map was thresholded to correct for multiple comparisons using a false discovery rate (FDR; p < .01). It is notable that because of the continuous update of the Neurosynth database, the number of studies included in the analyses could be slightly different for each search; the maps we used can be found in our Neurovault repository (https://identifiers.org/neurovault.collection:7731). Similar network identification was also performed using BrainMap (Laird, Lancaster, & Fox, 2005) as a confirmation. The two methods of meta-analysis yielded similar maps of network of interests (Appendix Figure A2), and we used the maps generated by Neurosynth in our main text.

We used the thresholded (p_{FWER} < .01) spatial maps of “inhibitory control” and “memory retrieval” to general three masks of networks of interest. The areas that belong to both the “inhibitory control” and “memory retrieval” masks were labeled as the “overlap network,” the areas that only belong to the “inhibitory control” mask were labeled as “inhibitory control network,” and the areas that only belong to the “memory retrieval” masks were labeled as “memory retrieval network.”

Brain Parcels for the Extraction of Time Series

We combined a parcellation of cerebral regions (n = 300; mean size = 440.1 voxels, SD = 188.9 voxels; Schaefer et al., 2018) and all subcortical regions (n = 14; mean size = 491.57 voxels, SD = 334.7 voxels) from the probabilistic Harvard–Oxford Subcortical Structural Atlas (Desikan et al., 2006) as a whole-brain parcellation. The parcellation of cerebral regions was based on a gradient-weighted Markov random field model, which integrated local gradient and global similarity approaches (Schaefer et al., 2018). Based on both task fMRI and resting-state fMRI acquired from 1489 participants, parcels with functional and connectional homogeneity within the cerebral cortex were generated. Each parcel is one of the seven large-scale functional brain networks, including visual, somatomotor, dorsal attention, ventral attention, limbic, frontoparietal, and default network (Buckner, Krienen, Castellanos, Diaz, & Yeo, 2011). Subcortical regions included bilateral thalamus, caudate, putamen, globus pallidus, hippocampus, amygdala, and ventral striatum. Details of each parcel (e.g., name, coordinates, and hemisphere) within the whole-brain parcellation can be found in our OSF folder (https://osf.io/c9q6b/). Some may argue that the size of parcels differs across brain regions, which could lead to different levels of signal-to-noise ratio for
extracted time series after averaging across all voxels within these parcels. We agreed with this possibility but held the idea that it is more important to extract signals from biological-valid and meaningful parcels instead of parcels with externally forced equal size. The latter may cause the issue that averaged neural signals are the combinations of two or more rather separate signals, which associate with different (cognitive) functions.

For each of the 314 parcels of the whole-brain parcellation, we compared it with the mask of overlap network, inhibitory control network, and memory retrieval network and identified the mask in which the parcel shared the highest percentage of common voxels. The parcel was assigned to that category if the highest percentage is higher than 10%. If the highest percentage of common voxels is lower than 10%, the parcel was not assigned to any category. After this procedure, 110 out of the 314 parcels were assigned to one of the categories. Specifically, 71 parcels were considered as memory-related regions, 29 parcels were categorized as control-related regions, and 10 parcels were labeled as overlap regions in our following analysis.

**Extraction of Time Series from Parcels**

We additionally removed nuisance time series (cerebrospinal fluid signals, white matter signals, motion, and event-related activity) using a method based on a projection on the orthogonal of the signal space (Lindquist, Wager, & Caffo, 2019; Friston et al., 1994). We generated confounding time series (cerebrospinal fluid, white matter, the six rigid-body motion parameters [three translations and three rotations], and framewise displacement [FD]) for each run of each participant. Event-related activity time series were estimated by a finite impulse response function. A recent study has shown that the removal of event-related activity based on finite impulse response modeling is an important step for the preprocessing of time series during a task (Cole et al., 2019). The signal from each parcel was extracted and z-scored, and all nuisance time series were removed simultaneously using the nilearn.signal.clean function. All cleaned time series were shifted three TRs (4.5 sec to account for the hemodynamic response function (HRF) delay and then aligned with the task condition (i.e., retrieval or suppression) at that moment. We shifted the time series by 4.5 sec because the peak of the HRF is between 4 and 6 sec from the triggering event (Lindquist, Meng Loh, Atlas, & Wager, 2009), and we reasoned that shifting four TRs (6 sec) could already miss the peak. This TR-wise data shifting is a standard practice in time-resolved decoding of fMRI data (https://brainhack-princeton.github.io/handbook/content_pages/05-02-mvpa.html#).

**The Transition of Neural States Analysis**

First, we characterized the transition of neural states at the group level. Extracted time series from each run of each participant were split according to the task instruction (i.e., memory retrieval or memory suppression) and concatenated. Second, two kinds of time series were further concatenated across five TNT runs within that participant (except for one participant, only four complete TNT runs were included). Third, time series were concatenated across all participants. Fourth, two time series were averaged across all time points separately to represent the mean activity intensity for that parcel during retrieval or suppression.

To estimate the relative dominance of each parcel during two neural states (i.e., think and no-think), we ranked the mean activity intensity of each parcel (the highest activity was ranked first). We then calculated the changes in ranks when the task switched from think to no-think by subtracting the rank during think from the rank during no-think. The same analyses were conducted with raw signal intensity and Z values. Results can be found in Appendix Figure A3. The negative change suggested an increase in relative dominance, whereas the positive change represented the opposite. We calculated two neural indexes ("state transition index" and "state transition index Version 2 [V2]") to quantify the transition of neural states at the individual level and associate this individual difference with the subsequent suppression-induced forgetting effect. The state transition index was calculated by adding up the averaged relative decreases (absolute values for negative values) in rank values of memory-related regions and the averaged relative increase in rank values of all control-related regions. The calculation and results of "state transition index V2" can be found in the Appendix: An Alternative Method to Quantify Individual Differences in Neural State Transitions. It is notable that although transition index and transition index V2 were calculated using different methods, they were based on the same set of data. Therefore, the data analyses of index V2 should not be regarded as independent analyses. To explore whether the neural state transition during the TNT relates to the suppression-induced forgetting effect, we correlated two state transition indices with the objective suppression score and subjective suppression score. Suppression scores were calculated based on memory performance during the final memory test after TNT.

**Neural States Decoding Analysis**

Before the decoding analysis, we generated the labels of task conditions for each time point within the trial based on its instruction (i.e., think or no-think). For example, if the trial is a think trial, time points started from the presentation of memory cues of this trial to the presentation of memory cues of the next trial were labeled as "think." We performed the time-resolved multivariate decoding analysis based on the brain activity of all 110 regions and corresponding labels of task conditions during each time point. This decoding analysis allowed us to generate the predicted label of task condition for
each time point, thus revealing the fast dynamics of the neural state transition induced by the switch of task conditions. Specifically, we performed the decoding analysis via the linear support vector classification (SVC), the C-support vector machine within the scikit-learn package (https://scikit-learn.org/stable/). We used the following parameters of the function: kernel = linear, regularization (C) = 1, radial basis function kernel with degree = 3. The classification of neural states was performed separately for each time point using a leave-one-run-out cross-validation approach within each participant. This procedure resulted in a decoded task condition for each time point of each participant. These predictions were evaluated by comparing these decoded task conditions with actual task conditions. To separate all types of correct and incorrect classification for the following analyses, we generated the confusion matrix for each participant. This confusion matrix contained the percentage of all four situations based on the task condition and if the prediction matches the task instruction (i.e., think–correct classification, think–incorrect classification, no-think–correct classification, and no-think–incorrect classification). We extracted all SVC discriminating weights assigned to the features during the participant-specific decoding and averaged them across all participants to generate the neural state-predictive pattern. The brain parcels with higher absolute values contributed more to decoding models.

To test for possible differences in neural representations of task conditions induced by the task switching, we performed the described decoding analyses for switch time points and nonswitch time points separately. The switch time points were defined as the presentation time (two TRs; 3 sec) of the first memory cue after the switch of task condition. The two decoding analyses yielded decoding accuracies for switch time points and for nonswitch time points for each participant. We compared these two types of decoding accuracies using the paired t test. Less accurate decoding was described as the evidence for the weaker representation of the current task condition in the literature (Loose et al., 2017; Waskom et al., 2014). Also, because we only have two task conditions, less accurate decoding reflects the unsuccessful transition from the previous demand to the current demand according to the instruction. After the general comparison between decoding accuracies between the switch and nonswitch conditions, we further analyzed them as the function of time (i.e., TR) and separately for think and no-think. This approach allowed us to explore the demand-specific neural dynamics of task condition representation within the different phases of one trial.

Next, we aimed to investigate the behavioral relevance of the mismatch moments (i.e., incorrect classification) between task conditions and the underlying neural states. Because we were mainly interested in the switch-induced mismatch, we first restricted our analyses to these switch time points and then extended them to the nonswitch time points as an exploratory analysis. For each participant, we averaged the trial-by-trial behavioral performance during the TNT task based on whether the actual task condition matches with decoded task conditions. This yielded retrieval performance and suppression performance for match and mismatch conditions. Paired t tests were performed to examine the effect of mismatch on the performance of memory retrieval and memory suppression separately. The performance calculations and comparisons described above were repeated for nonswitch time points as well.

### Alternative Classifier Training Procedures and Effects on Our Results

During our main neural state decoding analysis, fMRI time points regardless of their specific phase (e.g., memory cue, response, or fixation) and corresponding performance were all labeled as one of the two states (i.e., think or no-think) and used for classifier training and testing. We investigated the effects of applying two alternative classifier training procedures on our reported results. We described the procedures for alternative classifier training A and B below: (A) Only time points during the memory cue phase were used for classifier training and testing; during classifier training, only the data from trials that participants reported successful retrieval and suppression were used for classifier training to further increase the specificity of the classifier. Successful retrieval trials were defined as trials with “often” or “always” retrieval reports, and successful suppression trials were defined as trials with “never” or “sometimes” intrusion reports. (B) A subsampling balancing step was used to make sure that, for each participant, an equal number of data points from the “think” and “no-think” state was used for the classifier training; to prevent classifiers from using any sequential information during training, we reordered fMRI data (together with their state labels) not to reflect its original temporal sequence. For both procedures (A) and (B), after training, classifiers were used to generate the neural states (i.e., predicted state labels) only for time points during the memory cue phase. These labels were then analyzed to replicate three key findings of our decoding analysis: (1) higher-than-chance level neural state decoding accuracies; (2) less accurate decoding after switching compared with nonswitch trials; and (3) when predicted state labels (i.e., neural state) cannot match with the actual task condition, trial-by-trial behavioral performance was compromised.

### Relationship between Head Motion, Neural State Transitions, and Behaviors

To explicitly assess how head motion could potentially affect our results, we derived a time point-by-time point measure of head motion, FD (Power, Barnes, Snyder,
Schlaggar, & Petersen, 2012), during the TNT task. FD is defined as the sum (in mm) of rotational and translational displacements from the current volume to the next volume. We aligned the time series of FD with task structure and behaviors in a way similar to the analyses of the time series of fMRI signals but did not consider the HRF. The following contrasts were performed to compare head motion between conditions: (1) difference in FD between think trials and no-think trials, (2) difference in FD between correct neural state decoding and incorrect neural state decoding, and (3) difference in FD between the switch and nonswitch conditions. Correlations analyses were performed between individual differences in head motion between think and no-think trials (i.e., \( \text{FD}_{\text{Think}} - \text{FD}_{\text{No-think}} \)), state transition index, and objective/subjective suppression score.

**Data and Code Availability**

All research data of this study were uploaded to the Donders repository (https://data.donders.ru.nl/) and are publicly available. The project was named “Tracking the involuntary retrieval of unwanted memory in the human brain with functional MRI” in the repository (https://doi.org/10.34973/5afg.7ri1). Data, such as statistical maps and brain parcels of interest, were shared via the Neurovault repository (https://identifiers.org/neurovault.collection:7731).

Behavioral data were analyzed by JASP (https://jasp-stats.org/). For the term-based meta-analysis of neuroimaging studies, we used Neurosynth (https://neurosynth.org/) and BrainMap (www.brainmap.org/). Preprocessing of neuroimaging data was performed by FSL (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki), ICA-AROMA (https://github.com/maartenmennes/ICA-AROMA), and fMRIPrep (https://fmriprep.readthedocs.io/en/stable/). Python packages, including Nilearn (https://nilearn.github.io/), Nistats (https://nistats.github.io/), Pandas (https://pandas.pydata.org/), and NumPy (https://numpy.org/), were used for the analyses of time series. Machine learning algorithms were based on scikit-learn (https://scikit-learn.org/) and implemented via Nilearn (https://nilearn.github.io/). Anaconda (https://www.anaconda.com/) Python 3.6 was used as the platform for all the programming and statistical analyses. Custom Python scripts were written to perform all analyses described based on the mentioned Python packages; all code is available from the authors upon request and will be released via our OSF repository (https://osf.io/ccp96h/) upon publication.

**RESULTS**

**Behavioral Results**

Our study used a modified TNT paradigm (Figure 1A) with trial-by-trial reports of (in)voluntary memory retrieval (i.e., retrieval/intrusion frequency rating; Levy & Anderson, 2012). At the cue phase of each trial, the participant received a trial-specific instruction to either retrieve the memory that is associated with the cue (i.e., think trials) or suppress the tendency to recall the memory (i.e., no-think trials). Then, during the subsequent report phase, participants reported how well they just retrieved (i.e., retrieval rating) or suppressed the memory (i.e., intrusion rating). As intended, during the entire experiment (without considering the repetitions of presenting memory cues), most of the associations were successfully recalled in think trials (1 − mean \( p_{\text{Never}} \) = 84.05%, \( SD = 11.79\% \), range = 56.25%–100%; Appendix Figure A4A), whereas participants suppressed memory retrieval successfully in no-think trials in about half of the trials (mean \( p_{\text{Never}} \) = 50.62%, \( SD = 25.35\% \), range = 4%–92.5%; Appendix Figure A4B). Participants successfully retrieved associated memories during most of the trials (i.e., 84%), whereas they suppressed unwanted memories for only half of the trials (i.e., 50%). The successful retrieval and suppression percentage differences may suggest the unbalanced difficulty levels between the two task conditions involved (i.e., memory retrieval is easier than memory suppression). The difficulty could be the confound of our experimental design and raise the alternative interpretation for our results beyond memory-specific effect. In addition, we investigated the learning/practicing effect by analyzing participants’ memory retrieval and suppression performance as the function of repetition: for memory retrieval trials, the percentage of reporting “always” (i.e., index of more successful retrieval) increased, \( F(9, 234) = 5.3, p < .001, \eta^2 = .02 \); for memory suppression trials, the percentage of reporting “never” (i.e., index of more successful suppression) increased, \( F(9, 234) = 5.4, p < .001, \eta^2 = .04 \), from the 1st to the 10th repetition. These results together suggest that participants were getting more successful at retrieving or suppressing memory traces throughout the experiment.

Our central aim was to determine whether there were behavioral switch costs in the TNT task and to reveal their neural underpinnings. We defined each trial as a “switch” or “nonswitch” trial considering both the task condition of the current trial and its predecessor (Figure 1B). Specifically, we identified “switch” trials if a preceding trial had the “opposite” task condition (e.g., previous trial: think; current trial: no-think; T → NT). By contrast, if the current trial and the preceding trial had the “same” task condition, then the current one was a “nonswitch” trial. Within each run, the number of “switch” and “nonswitch” trials are almost identical (i.e., 32 vs. 31). We separately compared the trial-by-trial performance between “switch” and “nonswitch” trials for the think and no-think conditions. Participants showed comparable performance for “switch” and “nonswitch” trials in the think condition, \( t(25) = 0.348, p = .731, \) Cohen’s \( d = 0.068 \) (Figure 1C), whereas they reported more memory intrusions for “switch” trials compared with “nonswitch” trials in the no-think condition, \( t(25) = 3.19, p = .004, \) Cohen’s \( d = 0.627 \) (Figure 1D), suggesting “switch costs” when the task condition switched from a previous think trial to a current
Figure 1. (A) After learning 48 location–picture associations, participants performed a TNT task while brain activity was measured by fMRI. During think trials, participants were instructed to retrieve associated pictures based on the highlighted locations as memory cues. By contrast, during no-think trials, participants were required to suppress the tendency to retrieval the associated pictures. (B) The sequence of trials was designed to probe task switching between two task conditions (i.e., think and no-think). When the task condition of the current trial was the same as the previous trial, it was defined as the “nonswitch” trial. By contrast, although the task condition of the current differed from the previous trial, it was defined as the “switch” trial. (C) During think trials, participants demonstrated comparable memory retrieval performance ($p = .73$, Cohen’s $d = 0.068$) for both “switch” (no-think–think; NT → T) and “nonswitch” (think–think; T → T) trials. (D) During no-think trials, participants reported worse memory suppression performance, indexed by more memory intrusions for “switch” trials (think–no-think; T → NT) compared with “nonswitch” trials (no-think–no-think; NT → NT; $p = .004$, Cohen’s $d = 0.627$). (E) Brain regions showed increased activation during no-think trials compared with think trials. (F) Brain regions showed increased activation during think trials compared with no-think trials. Whole-brain brain imaging was thresholded at voxelwise $Z > 3.1$, cluster-level $p < .05$ FWER corrected. For C and D, bar charts demonstrated the mean and 95% confidence interval of the switch and nonswitch conditions. For E and F, unthresholded statistical maps can be found in the corresponding Neurovault repository (see Data and Code Availability section) for 3-D visualization.
no-think trial (T → NT). As we mentioned above, the no-think condition may be cognitively more difficult compared with the think condition. Therefore, the asymmetric switch costs may be interpreted as that the effect of task switching on behaviors is more pronounced during the cognitively challenging environment. However, our time-resolved decoding analysis below could potentially explain the asymmetric effect in the context of memory manipulation. Also, the link between trial-by-trial decoding accuracy and memory performance further suggest that the observed effects should be largely attributed to memory retrieval and suppression.

After the TNT task, participants performed a final memory test in which memory performance of think trials, no-think trials, and control trials (i.e., associations that were learned but not presented during the TNT task) was assessed. Behavioral results from the final memory test had been reported in another publication in detail (W. Liu, Kohn, et al., 2020) and Appendix Figure A5 of this study. In this study, we used memory performance during the final memory test to quantify individual differences of suppression-induced forgetting effects. The “forgetting effect” was defined as the differences between memory performance between the no-think and control trials in the final memory test. More specifically, we calculated individual differences in both subjective and objective suppression-induced forgetting effects based on subject and objective memory measures and correlated them with fMRI measures of neural state transitions.

fMRI Results

To replicate the univariate neural signature of memory suppression reported in prior studies (Anderson & Hanslmayr, 2014; Levy & Anderson, 2012; Anderson et al., 2004), we first conducted a univariate analysis to contrast brain regions engaged in memory suppression and memory retrieval (i.e., no-think vs. think). We found an increased activity for no-think trials in regions that are consistently involved in memory suppression, including the bilateral DLPFC, bilateral insula, bilateral IPL, SMA, and middle cingulate gyrus (voxelwise Z > 3.1, cluster-level p < .05 FWER corrected; Figure 1E and Appendix Table A1). Additionally, we found higher activity in ventral visual areas and the right thalamus during no-think compared with think trials. Next, we contrasted the think condition with the no-think condition and found the increased activity for the think condition in a set of regions, including the medial pFC, posterior cingulate cortex, hippocampus, IPL, precuneus, angular gyrus, and cerebellum (voxelwise Z > 3.1, cluster-level p < .05 FWER corrected; Figure 1F and Appendix Table A2). Together with the behavioral results from the final memory test, these results confirmed that participants in our experiment followed task instructions, leading to univariate neural signatures of memory retrieval and suppression consistent with prior findings (Anderson & Hanslmayr, 2014; Levy & Anderson, 2012; Anderson et al., 2004), as well as recent meta-analyses of memory suppression (W. Liu, Peeters, et al., 2020; Guo et al., 2018).

The Transition of Large-scale Neural States from Memory Retrieval to Memory Suppression

Based on neurocognitive models of memory suppression (Anderson & Hanslmayr, 2014), we focused on the neural dynamics within the inhibitory control network and the memory retrieval network. First, we used Neurosynth (https://neurosynth.org/), an automatic meta-analysis tool of neuroimaging data (Yarkoni et al., 2011), to identify the inhibitory control network and memory retrieval network independently from our fMRI data. Using the term “inhibitory control” and “memory retrieval,” we performed term-based meta-analyses to reveal two distinct brain networks of inhibitory control (Appendix Figure A6A) and memory retrieval (Appendix Figure A6B) separately. The two meta-analytic maps have “overlapping” areas, including the inferior frontal gyrus (IFG), insular, SMA, IPL (Appendix Figure A6C). Interestingly, the latter areas are highly similar to a “task-switching” map generated by Neurosynth using the term “task switching” (Appendix Figure A6D).

In the next step, we tried to identify individual brain regions within the inhibitory control, memory retrieval, and overlapping networks. Based on the combination of a connectivity-based neocortical parcellation (number of parcels = 300; Schaefer et al., 2018) and subcortical regions (number of regions = 14; for details, see Methods section), we identified 71 regions (i.e., memory-related regions) within the memory retrieval network, 29 regions (i.e., control-related regions) within the inhibitory control network, and 10 regions (i.e., overlapping regions) within the “overlapping” network (Figure 2A). Finally, for each of the 110 regions, the BOLD time series were extracted from each voxel, averaged within each region, and further processed.

Using these time series, we characterized the group-averaged transition of neural states when the task condition changed from think to no-think trials (Figure 2B). Based on the task instruction, the time series were first split for the think and no-think conditions separately and then concatenated across all runs of all participants. For each task condition, all regions were ranked (the highest activity was ranked first) based on their state-specific averaged neural activity across runs to represent their relative dominance during that neural state (i.e., think or no-think). A Kruskal–Wallis test showed that, during think trials, memory-related regions, control-related regions, and overlapping regions differed in their ranks, H(2) = 40.48, p < .001. Post hoc Mann–Whitney tests using a Bonferroni-adjusted alpha level of .017 (.05/3) were used to compare all group pairs. Memory-related regions (mean_memory = 40.22, SD_memory = 27.39) ranked higher than control-related regions (mean_control = 82.54, 2568 Journal of Cognitive Neuroscience Volume 33, Number 12
$SD_{\text{control}} = 22.28$) and overlapping regions ($mean_{\text{overlap}} = 75.10, SD_{\text{overlap}} = 19.08$; memory-related vs. control-related: $U = 263, p < .001$; memory-related vs. overlapping: $U = 108, p < .001$). Control-related regions and overlapping regions did not differ significantly in their ranks ($U = 104, p = .096$). Three types of regions also differed in their ranks during no-think trials, $H(2) = 36.60, p < .001$. Memory-related regions ($mean_{\text{memory}} = 67.96, SD_{\text{memory}} = 27.94$) ranked lower than control-related regions ($mean_{\text{control}} = 27.24, SD_{\text{control}} = 23.13$) and overlapping regions ($mean_{\text{overlap}} = 37.80, SD_{\text{overlap}} = 21.10$; memory-related vs. control-related: $U = 238, p < .001$; memory-related vs. overlapping: $U = 144, p = .001$). Control-related regions and overlapping regions did not differ significantly in their ranks ($U = 101, p = .081$). All comparisons between memory-related and control-related/overlapping regions were significant after Bonferroni adjustment (all $ps \leq .001$). Furthermore, we performed additional analyses of neural state transition by dividing all regions into three groups (i.e., increased group, stable group, and decreased group) based on their relative changes in rank (see Appendix: Additional Analyses of TNT Neural State Transition). In short, when the task condition changed from think to no-think, memory-related regions showed decreases whereas control-related regions demonstrated an increase in their activity ranks.

For control purposes, the same analysis was repeated for raw signal intensities (Appendix Figure A3B) and their $Z$ values (Appendix Figure A3C) and yielded similar patterns. Here, we present parametric paired $t$ tests of $Z$ values to validate our findings based on ranks: Control-related regions demonstrated increased (think > no-think: $t(28) = 4.79, p < .001$, Cohen’s $d = 0.89$) whereas memory-related regions showed decreased (no-think > think: $t(70) = 4.00, p < .001$, Cohen’s $d = 0.48$) neural activity during the TNT transition. Furthermore, we analyzed the average signal within the default mode network (DMN), a functional network that is largely involved in memory retrieval, and found that DMN was less activated during the no-think compared with the think condition, $t(25) = −3.24, p = .003$, Cohen’s $d = 0.63$.

The Transition of Neural States during the TNT Task Is Associated with Subsequent Suppression-induced Forgetting

We already demonstrated that the activity of control-related regions increased and the activity of memory-related regions decreased when the task conditions switched from think to no-think. To assess if these changes in neural states are associated with the behavioral consequence of memory suppression (i.e., suppression-induced forgetting effect), we quantified this neural state transition at the individual level and examined whether individual differences in the neural state transition predict individual subsequent suppression-induced forgetting measures (i.e., subjective and objective suppression score). The subjective and objective suppression score was calculated by subtracting the memory measure (i.e., confidence rating or recall accuracy, respectively) of suppression associations (i.e., “no-think” items) from control associations separately.

Based on the group-level fMRI results, we calculated a “state transition index” to represent the degree of neural transition. The index was calculated by subtracting the rank of memory-related regions from the rank of control-related regions (or overlapping regions if applicable) during the TNT transition. This index was then correlated with the suppression-induced forgetting scores (i.e., subjective and objective scores). The correlation coefficients were $r = .36$ and $r = .20$ for subjective and objective scores, respectively. These results suggest that the transition of neural states during the TNT task predicts individual differences in suppression-induced forgetting.

Figure 2. (A) Memory retrieval network (green) and inhibitory control network (red) was defined using Neurosynth independent of fMRI data analyzed in this study. The overlap between the two brain networks was defined as the overlapping network (blue). (B) When the task condition switched from think to no-think, the activity of brain regions within the inhibitory control network increased whereas the activity of brain regions within the memory retrieval network decreased. $y$ axis is the rank of neural activity among 110 regions. The top of the axis represents the highest rank (i.e., strongest activity). Each dot is a brain region. (C) Individual differences in the neural state transition tended to correlate with the objective suppression-induced forgetting effect during the subsequent final memory task ($r = .36, p = .06$). (D) The same index also tended to correlate with the subjective suppression-induced forgetting effect ($r = .38, p = .05$).
transition during the TNT task for each participant. The state transition index was calculated by adding up the averaged relative decreases (absolute values for decreased values) in ranks of all memory-related regions and the averaged relative increase in the rank of all control-related regions during the TNT transition. The larger the state transition index represents a larger decrease for memory-related regions and a larger increase for control-related regions. The state transition index tended to be positively correlated with individual differences in objective suppression scores \((r = .36, p = .06;\) Figure 2C) and subjective suppression scores \((r = .38, p = .05;\) Figure 2D). For validation purposes (not an independent analysis), we used an alternative method (i.e., state transition index V2) to measure neural state transitions for each participant. This method was based on additional analyses of TNT neural state transition (see Appendix for details of state transition index calculation and results): All regions were divided into three groups (i.e., increased group, stable group, and decreased group) based on their relative changes in ranks. The state transition index V2 was calculated as the sum of the percentage of memory-related regions within the decreased group and the percentage of control-related regions within the increased group. Similarly, a larger transition index V2 suggests the stronger TNT neural transition (i.e., decreasing activity memory-related regions while increasing activity for control-related regions). We also found the same significant correlations between state transition index V2 and both objective and subjective suppression scores (Appendix Figure A7). These results suggested that the transition of neural states during the TNT task is relevant for the subsequent suppression-induced forgetting effects measured in the final memory test.

**Switch of Task Condition Is Accompanied by the Delayed Transition between Neural States**

The large-scale neural state transition described above was based on all neural data available, which means non-switch trials were also included and analyzed. To reveal how neural representations of task conditions change during task switching, we used a multivariate decoding method to track the dynamics of neural state transitions on a time point-by-time point basis (Figure 3A). By doing this, each time point can be labeled as “switch” or “non-switch,” and therefore, the effect of task switching on the neural representation of task condition can be examined at the temporal resolution of each fMRI volume. Linear SVC was used to classify the underlying neural states (i.e., think vs. no-think) based on the fMRI activity intensity of all 110 regions at each given time point. In addition, the same decoding analysis was performed based on the on neural activities from all 314 parcels (i.e., all parcels within the whole-brain atlas). In short, network-specific decoding and whole-brain decoding yielded similar results, leaving open the possibility that our result may not be specific to predefined brain networks. Participant-specific classifiers were fitted on neural and task condition data from \(N - 1\) runs (i.e., four runs) and tested on the one remaining test run. Then, the decoding accuracy was evaluated for each TNT run by comparing the decoded task conditions with the actual demands. Averaged across runs, we were able to decode task conditions based on multivariate regional neural activity with a mean accuracy of 59.5% \((SD = 3.9\% , range = 52.5\% -67.1\%;\) Figure 3B). This accuracy is significantly higher than chance level (i.e., 50%), \(t(25) = 12.5, p < .001,\) Cohen’s \(d = 2.453\). Because we identified the learning/practicing effect at the behavioral level (i.e., participants were getting better at retrieving/suppressing memories with repetitions), here, we asked whether such behavioral effects could affect the accuracies of our neural state decoding. We found that although participants’ behavioral performance improved from the first run to the last run, decoding accuracies of five runs did not differ from each other, \(F(4, 96) = 2.19, p = .075, \eta^2 = .08\). Admittedly, the effect of repetition on the neural decoding is close to being significant (i.e., decreasing, not as expected increasing, decoding accuracies throughout the experiment). This tendency raised two things to be noted: First, because we were using the leave-one-run-out cross-validation, which assumes each run is an identical replication, the close-to-be-significant repetition effect here partly violated the assumption. Second, we suspected that the tendency of decreasing decoding accuracies might suggest that the learning/practicing process led to a less typical neural state for each task condition and therefore allows more flexible neural state transition corresponding to the switching of external task conditions. That is the reason why we observed improved behavioral performance but a nonsignificant tendency of less accurate neural decoding.

We further generated the confusion matrix of our decoding analysis to quantify all types of correct and incorrect classifications (Figure 3C): 57.9% \((SD = 4.1\%, range = 50.6\% -65.7\%)\) of think time points were correctly classified as think. Among all no-think time points, 61.1% \((SD = 3.9\%, range = 53.9\% -68.4\%)\) of them were correctly classified as no-think. To reveal the relative contribution of each region to this decoding performance, we visualized the neural state-predictive pattern (i.e., SVC discriminating weights) in Figure 3D, which revealed a frontoparietal network of strong task condition representation, including the dorsal ACC (dACC), DLPFC, IFG, superior parietal lobule, and IPL (Appendix Table A3). These regions were largely similar to the overlapping network (Appendix Figure A8).

To reveal how the switches of task conditions affected underlying neural state transitions, we calculated the decoding accuracy for “switch” and “nonswitch” time points separately. Higher decoding accuracy represented a timely update of neural states according to the current task condition, thus having a stronger neural representation of task condition. Compared with “switch” time
points, the task condition of “nonswitch” time points can be decoded more accurately, $t(25) = 7.1, p < .001$, Cohen’s $d = 1.39$ (Figure 3E). That is to say, time points within no-think trials following a think trial were more often misclassified as think trials compared with a no-think trial following another no-think trial. This pattern of results was also observed for think trials.

**Stronger Post Switch Adaptive Representation Mitigates Switch Costs during the No-think–Think Transition**

At the behavioral level, we found the switch costs during the TNT transition (i.e., T → NT), but not during the no-think–think transition (i.e., NT → T). Here, we asked whether this behavioral asymmetry can be explained neurally. Because our neural state decoding was performed on individual fMRI time points, we explored how task switching differentially affects the temporal dynamics of task representation within no-think and think trials by analyzing decoding accuracies as the function of time (i.e., TR) separately for no-think and think trials (Figure 4A and B). It is notable that our decoding analysis did not consider the variability in the shape of HRF across brain regions and participants. Therefore, it should be treated with caution. We found an asymmetric effect of task switching on the temporal dynamics of task representations for the think and no-think conditions. Specifically, a stronger post switch adaptive task representation was selectively found after the no-think–think transition. Although the task representations are weakly represented at the beginning (TR = 1: $t(25) = −6.26, p < .001$, Cohen’s $d = 1.22$; i.e., “delay” effects), it was better represented halfway during the trials (i.e., “adaptation” effects), mainly the response phase, compared with the no-switch trials (TR = 3: $t(25) = 5.79, p < .001$, Cohen’s $d = 1.13$; TR = 4: $t(25) = 2.5, p = .019$, Cohen’s $d = 0.49$; TR = 5: $t(25) = 2.96, p = .007$, Cohen’s $d = 0.58$; Figure 4A). By contrast, for the no-think condition, although there was also a “delay” effect immediately after the switch (TR = 1: $t(25) = −3.33, p = .003$, Cohen’s $d = 0.65$), post switch adaptation did not exist: Task representations only differed at the beginning and then became comparable during the response and fixation phase (Figure 4B).

Furthermore, we calculated the differences in decoding accuracies by subtracting accuracies of the switch condition from the nonswitch condition (i.e., Accuracies_{nonswitch} − Accuracies_{Switch}) and plotted them as the function time for demonstration purposes (Figure 4C). To identify potential interactions, we compared these differences of accuracies between think and no-think trials using a $2 \times 2$ ANOVA (Think/Nothink × Switch/Nonswitch). There was a
significant interaction effect only when the TR equals 3,
\(F(1, 25) = 12.53, p = .002, \eta^2 = .07\): Decoding accuracies for the switch and nonswitch conditions did not differ in no-think trials \(t = -1.49, p_{holm} = .28\), whereas decoding accuracies for the switch condition was higher than the nonswitch condition in think trials \(t = 6.76, p_{holm} < .001\). To further support the idea that that higher decoding accuracies during think trials reflect stronger task representation and therefore can facilitate memory retrieval, we correlated individual differences in decoding accuracies (when there were marked “adaptation” effects) with memory retrieval/suppression performance. We showed that individuals who demonstrated stronger “adaptation” effects, indexed by higher decoding accuracies, performed better at memory retrieval (switch condition: \(r = .44, p = .02\); nonswitch condition: \(r = .43, p = .02\)), but not memory suppression (switch condition: \(r = -.09, p = .64\); nonswitch condition: \(r = -.16, p = .43\); Appendix Figure A9). Although this correlation needs to be interpreted with caution and only regarded as preliminary evidence because our sample size is modest \((n = 26)\), it suggests the task-specific association between adaptive representational strength and memory retrieval.

**Mismatch between Task Condition and Underlying Neural State Relate to Trial-by-Trial Memory Performance**

We have already shown the association between decoding accuracies and behaviors in cross-participant analyses. Next, we sought to further link the underlying neural state to memory performance on a trial-by-trial basis. Unlike most of the decoding analyses, which usually focused on the accuracy of the classifier, here, we were particularly interested in the relationship between misclassified moments and their behavioral consequences (i.e., switch costs). We already demonstrated that these misclassifications were largely induced by task switching, and we

Figure 4. Time-revolved neural decoding of the task representation and the behavioral relevance of misclassification. (A) Decoding accuracies of time points during the think condition. Lower decoding accuracies (i.e., “delay” effects) were found immediately after the switch (i.e., no-think → think, NT → T) compared with the nonswitch condition, whereas higher decoding accuracies (i.e., “adaptation” effects) were found for switch compared with nonswitch condition during the middle of the trials (i.e., response phase). (B) Decoding accuracies of time points during the no-think condition. Lower decoding accuracies (i.e., “delay” effects) were found immediately after the switch (i.e., think → no-think, T → NT). Decoding accuracies were comparable between switch and nonswitch during the response and fixation phase. (C) Comparison of differences in decoding accuracies as the function of time between think and no-think condition. Differences were calculated by subtracting the accuracies of the switch condition from the nonswitch condition. Therefore, positive values (i.e., nonswitch > switch) represented the switch-induced delay in the neural state transition, whereas negative values (i.e., switch > nonswitch) signified the post switch adaptive representation. Note that this figure was made only for demonstration. The 2 × 2 ANOVA (Think/No-think × Switch/Nonswitch) was performed to test the group differences of differences statistically. (D) Memory retrieval data from think trials. After the task switching, when the decoded neural state did not match with the task condition (i.e., neural state = no-think; task condition = think), participants reported worse memory retrieval performance \((p < .001, Cohen’s d = -0.70)\). (E) Memory suppression data from no-think trials. When the neural decoder misclassified no-think moments as think (i.e., neural state = think; task condition = no-think), participants reported more memory intrusions during no-think trials \((p = .005, Cohen’s d = 0.606)\).
predicted that this mismatch could be the neural source of behavioral switch costs. To test this idea, we averaged the trial-by-trial performance measures (i.e., retrieval frequency rating for think trials and intrusion frequency rating for no-think trials) for four situations at issue (i.e., think–correct classification, think–incorrect classification, no-think–correct classification, and no-think–incorrect classification) within switch trials.

We found that participants’ behavioral performance was impaired during these mismatch moments (i.e., incorrect classifications; mean\textsubscript{incorrect} = 43.9%, SD\textsubscript{incorrect} = 2.6%, range = 38.5%–49.8% of all classifications) immediately after task switches. Specifically, during the think condition, when neural states were mistakenly classified as no-think, the retrieval frequency rating was lower, \( t(25) = -3.57, p = .001, d = 0.701 \) (Figure 4D) compared with the situation in which task conditions matched with the neural state. During no-think trials, if neural states were erroneously decoded as think, participants reported higher intrusion frequency rating, \( t(25) = 3.08, p = .005, d = 0.606 \) (Figure 4E) compared with the situation in which classifications were correct.

In our exploratory analyses, we found that such mismatch moments not only occurred during the task switching but were also observed (but less frequently) during nonswitch trials. Using the same decoding method but focusing on nonswitch time points, we found a similar detrimental effect of mismatch on behavioral performance (Appendix Figure A10). These findings suggested that spontaneous, uninstructed neural state fluctuations that do not fit current task conditions also have behavioral impacts.

Applying Alternative Classifier Training Procedures Did Not Change Our Findings

For the neural state decoding results presented above, we trained the classifier based on all individual fMRI time points, including not only memory-related phases (i.e., retrieval or suppression) but also the response and fixation phase. This was based on our assumption that the neural representation of the current task condition remained active during the entire trial and only changed when task conditions switched. One may argue that including time points beyond the mere memory-related phases (e.g., response phase) may bias the decoding analyses. To rule out this possibility, we retrained our classifiers using an alternative training procedure (A) and tried to replicate our key findings. More specifically, (1) we restricted our classifier training and testing to time points during memory-related phases only (i.e., retrieval or suppression), and (2) based on trial-by-trial memory performance, we further excluded time points when memory retrieval or suppression was unsuccessful. Classifiers that were trained based on the described alternative procedure can still decode task conditions significantly higher than the chance level, \( t(25) = 4.92, p < .001, \) Cohen’s \( d = 0.96 \). When we compared decoding accuracies between “switch” and “nonswitch” trials, we found a switch-induced delay in neural state transitions again: lower decoding accuracy for switch trials compared with no-switch trials, \( t(25) = -2.42, p = .02, \) Cohen’s \( d = 0.47 \). Lastly, we also revealed the behavioral relevance of the misclassified time points: When the no-think trials were misclassified as think trials, participants reported more memory intrusions, \( t(25) = 3.10, p = .005, \) Cohen’s \( d = 0.60 \); when the think trials were mistakenly labeled as no-think trials, participants tended to report lower rates of memory recall, \( t(25) = -1.98, p = .058, \) Cohen’s \( d = 0.39 \).

Another potential bias in our classifier training relates to the fact that we used more no-think time points for the training than think time points, \( t(25) = 4.56, p < .001, \) Cohen’s \( d = 0.89 \). Therefore, classifiers may be better suited to identify a “no-think” state (i.e., suppression state). To control this factor, we included a subsampling balancing step before classifier training: For each participant, we randomly excluded no-think data points to make sure that an equal number of think and no-think time points were used for the model training. At the same time, the original temporal sequence of the training data/label was reorganized. The latter ensures that the classifier was never aware of the actual order of the trial presentation, which should minimize the potential effects of temporal autocorrelation in the fMRI data on time-resolved decoding analyses. We retrained classifiers and again using the described alternative training procedure (B) and replicated our three key findings: (1) Classifiers can decode task condition better than chance level, \( t(25) = 4.97, p < .001, \) Cohen’s \( d = 0.97 \); (2) but the decoding tended to be less accurate after task switches, \( t(25) = -1.83, p = .07, \) Cohen’s \( d = 0.35 \); and (3) participants’ performances was worse during misclassified time points compared with the correctly classified time points: lower memory recall, \( t(25) = -1.72, p = .09, \) Cohen’s \( d = 0.33 \), and more memory intrusions, \( t(25) = 2.99, p = .006, \) Cohen’s \( d = 0.89 \).

Lastly, we asked whether our decoding analyses are contingent on using the specifically defined memory and inhibition networks from Neurosynth. A recent study attempted to understand neural state transitions during task performance using whole-brain patterns of activity (Shine et al., 2019). To investigate if the delay in the neural state transitions is the network-specific or whole-brain representation, we trained our classifier based on neural activities from all 314 parcels. We found that (1) whole-brain representation can still allow us to decode task conditions at a similar level of network-specific decoding (mean\textsubscript{whole-brain} = 59.9%, SD\textsubscript{whole-brain} = 4.7%, range = 44.5%–68.4%; mean\textsubscript{network-specific} = 59.1%, SD\textsubscript{network-specific} = 4.7%, range = 45.7%–67.1%); (2) whole-brain representation was delayed after task switches, \( t(25) = -7.83, p < .001, \) Cohen’s \( d = 1.53 \); and (3) misclassified moments were also associated with impaired memory performance: lower
memory recall, $t(25) = -4.91, p < .001$, Cohen’s $d = 0.96$, and more memory intrusions, $t(25) = 4.69, p < .001$, Cohen’s $d = 0.92$.

In summary, we retrained our decoding models to investigate whether our results are robust to three alternative classifier training procedures. Although some of the statistical tests failed to reach significance, they showed numerical trends toward our original findings, and our key decoding results can be replicated across different classifier training procedures. This outcome suggests that our conclusions are independent of specific methodological choices during classifier training.

**Neural State Transitions Are Not the Results of Differences in Head Motion**

We observed large-scale neural state transitions during the TNT task. Individual differences in these transitions were associated with the subsequent suppression-induced forgetting effect. There is a possibility that these neural state transitions are based on artifacts caused by different levels of participants’ head motion (Huang et al., 2018; Siegel et al., 2017) between think and no-think trials because more inhibitory control resource was required for no-think trials compared with think trials (Anderson & Hanslmayr, 2014; Anderson & Green, 2001). Therefore, we examined the relationship between head motion, neural state transitions, and behaviors to rule out this alternative explanation.

We analyzed the time series of head motion (i.e., FD; Power et al., 2012) during the TNT task. First, there is no significant difference in mean FD values between think and no-think trials ($FD_{\text{think}} = 0.149, SD = 0.047; FD_{\text{no-think}} = 0.149, SD = 0.046$), $t(25) = 0.30, p = .76$, Cohen’s $d = 0.06$. Second, for each participant, we calculated differences between the head motion of think and no-think trials (i.e., $FD_{\text{think}} - FD_{\text{no-think}}$) and found no correlation between these differences and state transition indices ($r = -.3, p = .12$), objective suppression scores ($r = -.14, p = .46$), or subjective suppression scores ($r = -.23, p = .25$). Third, we asked whether head motion could affect our neural state decoding analysis. The head motion level tended to be lower, $t(25) = -1.96, p = .06$, $d = 0.38$, for correct decoding ($FD_{\text{correct}} = 0.147, SD = 0.045$), compared with incorrect decoding ($FD_{\text{incorrect}} = 0.151, SD = 0.048$). This difference raised the question of whether the lower decoding accuracy for switch compared with nonswitch condition resulted from higher head motion instead of differences in the neural representation related to task conditions. Therefore, we also compared head motions between the switch and nonswitch conditions. In fact, we found that head motion is even lower in the switch condition ($FD_{\text{switch}} = 0.145, SD = 0.048$; $FD_{\text{nonswitch}} = 0.150, SD = 0.047$), $t(25) = 3.35, p = .003$, $d = 0.65$. This result ruled the head motion out as an alternative explanation for the lower decoding accuracy for the switch condition. If the lower decoding accuracy during switching was dominantly driven by excessive head motion, we would observe relatively higher instead of lower head motion during switching. In summary, analyses of head motion suggest that our neuroimaging results are not likely to be the consequence of variations in head motions between task conditions.

**DISCUSSION**

Task switching is a crucial cognitive ability that has been intensively studied using behavioral and neuroimaging methods (Richter & Yeung, 2014; Ruge, Jamadar, Zimmermann, & Karayanidis, 2013; Meiran, 2010). Here, we investigated the task switching process between memory retrieval and suppression and demonstrated that memory suppression is more difficult when the task condition for the participants just switched from retrieval to suppression (but not vice versa). Applying multivariate time-resolved decoding methods to human fMRI data, we revealed that immediately after the switch, task conditions were weakly represented by the inhibitory control and memory retrieval networks, indexed by the lower decoding accuracy, compared with nonswitch trials. Importantly, during the switching, when the neural representation of task condition cannot be updated in time to match the current condition (i.e., the mismatch between task condition and neural state), participants reported more memory intrusions in no-think trials and less memory retrieval in think trials. Together, we provided neural evidence based on the decoding approach to support the previously proposed cognitive theories of switch costs: Delayed transition of task-related neural states is associated with behavioral switch costs. That is to say, if the neural state cannot be timely updated after the switch of task condition, behavioral performance is compromised.

In the current study, participants were instructed to perform one of two opposite memory-related tasks (i.e., memory retrieval and memory suppression), with the task condition staying the same or switching between consecutive trials. Similar to what was reported in the classical task-switching paradigms, which focused on RT and/or accuracy (Spector & Biederman, 1976; Jersild, 1927), we found switch costs in the memory performance. We interpreted the switch costs during memory suppression as the carryover effect of task representation of memory retrieval on the current memory suppression. The carryover effects of the previous task set activation on the performance of the following task have been widely studied in the task-switching literature (Monsell, 2003), but not within the context of memory. Hulbert, Henson, and Anderson (2016) reported a carryover effect of memory suppression on the subsequent memory formation: When healthy participants suppressed unwanted memories, they were more likely to fail to encode information that was presented after a suppression trial. It was proposed that memory suppression created an amnesic time window, preventing the experience within the window...
from being transformed into long-term memory. Evidence from fMRI supported this model by showing the reduction of hippocampal activity during memory suppression trials and the positive correlation between individual differences in decreased hippocampal activity and the extent of memory impairment across participants (Hulbert et al., 2016).

Our finding of more memory intrusions during the no-think trials that followed a think trial could result from a similar mechanism: The preceding think trials create a time window in which the hippocampus-centered memory network remains active to support retrieval. However, if the transition of the neural state is delayed, the following no-think trials are still located within this window, and therefore, more prefrontal control resources are needed to down-regulate hippocampal activity. We tested this prediction beyond the hippocampus: Large-scale neural activity of the inhibitory control and memory retrieval networks were analyzed by multivariate decoding methods to track the adaptive neural state transitions. We first characterized the transitions in neural states of memory retrieval and inhibitory control networks between think and no-think trials. Consistent with previous models of memory suppression (Anderson & Hanslmayr, 2014), our results showed that when the task condition switched from retrieval to suppression, memory retrieval-related regions, mainly including the hippocampus and regions of DMN, decreased their neural activity, whereas inhibitory control-related regions, such as dACC and lateral pFC, increased their activity. We also examined the relationship between individual differences in the efficiency of neural state transitions and the suppression-induced forgetting effect measured in the subsequent final memory test and found a positive correlation between them. There are two possible explanations for how neural state transitions are related to the effect of memory suppression: Either larger or smaller state transitions are associated with the stronger suppression effect. The more intuitive explanation is that larger transitions are beneficial for suppression. However, our data suggested the opposite: Participants who demonstrated less neural reconfiguration showed a stronger memory suppression effect in the following final memory test. This finding is nevertheless consistent with a previous study, which demonstrated that higher intelligence is associated with less task-related neural reconfiguration (Schultz & Cole, 2016). Our data, together with this study, may suggest that less neural reconfigurations could reflect optimization for efficient (i.e., less) state updates, reducing processing demands (Schultz & Cole, 2016). This optimal task-related neural reconfiguration could then be beneficial for memory suppression.

Interestingly, switch costs in our study were specific to memory suppression: Participants reported more memory intrusions when the current no-think trial followed a think trial (i.e., TNT transition), suggesting a higher demand for cognitive control over the tendency to retrieve during switch trials compared with nonswitch trials. At the same time, we did not find the effect of task switching on the no-think–think transition. One possible explanation for the asymmetric effect is the unbalanced difficulty levels between memory retrieval and suppression. Asymmetric switch costs were studied as the results of sequential difficulty effects during task switching, although the results and predictions from previous studies are mixed. Results from our study can contribute to this field if we hold the assumption that the no-think condition is more difficult than the think condition because of the larger executive control requirement. One previous task-switching study reported greater switch costs when switching from the easy to the difficult task (Arbuthnott, 2008), consistent with the switch costs only in the TNT transition. But some studies proposed a higher switch cost for the easy task than for the difficult task caused by the need for inhibition (Mosbacher, Brunner, & Grabner, 2020; Schneider & Anderson, 2010). Despite the fact that participants reported successful memory retrieval more often than successful suppression, we were reluctant to directly compare these trial-by-trial ratings across task conditions because of the intrinsic differences in the underlying cognitive processes and task instructions. Potential differences in task difficulties could confound neural decoding analysis (Todd, Nystrom, & Cohen, 2013). Therefore our decoding results could reflect the nonspecific difficulty differences instead of memory-specific effects. Our neuroimaging analyses suggested the opposite: First, our univariate contrast analyses identified the activation/deactivation of the board memory network, including but not limited to the hippocampus, whereas conventional task-switching studies with difficulty manipulation usually only involved an executive control network; second, our trial-by-trial decoding analysis linked the mismatch between neural state and task requirement and memory performance; third, our time-resolved neural decoding analysis revealed the potential neural source of these asymmetry switch costs could be the stronger post switch adaptive representation, which is specific to memory retrieval. Taken together, although the difficulty could be the confound of our experimental design and raised the alternative interpretation for our results, evidence from the neuroimaging data provided more substantial support for interpreting our results in the context of memory. Recent human fMRI studies revealed task representations using multivariate decoding methods. Brain regions such as the parietal cortex, medial, and lateral pFC encode the current task conditions (Etzel, Cole, Zacks, Kay, & Braver, 2016; Wisniewski, Reuter, Tusche, & Haynes, 2015; Waskom et al, 2014; Mommennejad & Haynes, 2013; Cole, Etzel, Zacks, Schneider, & Braver, 2011; Gilbert, 2011; Woolgar, Thompson, Bor, & Duncan, 2011; Bode & Haynes, 2009), and our study provided further support for this idea by showing that neural activity patterns of these regions largely contributed to successful discrimination.
between two kinds of visually highly similar trials with opposite task conditions (i.e., memory retrieval and suppression). These identified regions have been previously associated with cognitive processing such as retrieval, maintenance, the process of rules, or demands during task switching (Reverberi, Görgen, & Haynes, 2012; Gilbert, 2011; Woolgar et al., 2011; Bunge, Kahn, Wallis, Miller, & Wagner, 2003; Sakai & Passingham, 2003). Beyond that, memory-related areas such as the hippocampus and regions within the DMN also contributed to the successful decoding in our study because the retrieval demand and its associated neural activity significantly differed between think and no-think trials. However, whether these task representations can be modulated by external experimental manipulations and detected by fMRI signals is an ongoing debate. Task representations are modulated by factors including rule complexity (Woolgar, Afshar, Williams, & Rich, 2015), rewards (Etzel et al., 2016), difficulty (Wisniewski et al., 2015), and skill acquisition (Jimura, Cazalis, Stover, & Polkdrack, 2014), but not by variables such as task novelty (Cole et al., 2011) or intention (Wisniewski, Goschke, & Haynes, 2016; Zhang, Kriegeskorte, Carlin, & Rowe, 2013).

Two studies directly investigated whether and how cognitive control processes during task switching modulate the neural representation of the current task condition. Waskom et al. (2014) found that task representations are enhanced after switches, indexed by the higher decoding accuracy. However, they did not find evidence for behavioral switch costs in their sample; thus, the behavioral relevance of the reported stronger task representation (i.e., higher decoding accuracy) is unclear. Loose et al. (2017) did find the behavioral switch costs, but no modulation effect in the task representation (i.e., comparable decoding accuracy between the switch and nonswitch trials), and therefore, Loose and colleagues proposed the switch-independent neural representations of the current task condition. Compared with the two studies mentioned above, our study found asymmetrical behavioral switch costs (i.e., only for TNT transition, but not for no-think–think transition). To investigate the potential neural source of these asymmetric switch costs, we further analyzed decoding accuracies as a function of time separately for memory retrieval and suppression. For both tasks, decoding accuracy was lower immediately after task switching. However, we found the post switch adaptive representation of the current demand for memory retrieval. Specifically, although the representation of the current demand (i.e., retrieval) was weaker compared with the previous demand (i.e., suppression) immediately after switching, representations of retrieval demand were quickly increased, and then they became even stronger than the previous demand halfway during trials. Interestingly, we also found preliminary correlational evidence that participants who showed stronger post switch adaptive representation performed better at the memory retrieval task. Therefore, we propose that this retrieval-specific adaptation may explain why there were no obvious switch costs during memory retrieval. Also, this interpretation would be consistent with the combination of stronger task representation and no behavioral switch costs reported by Waskom and colleagues. Such demand-specific neural dynamics of adaptive task representation have never been reported before, potentially because our time-resolved decoding approach has not been used in the previous task-switching studies. Future studies with such a combination of time-resolved decoding approach and EEG or MEG may reveal even more details of this adaptive task representation during task switching. Critically, our neural state decoding analysis revealed the relationship between current task representation and behavioral performance on a trial-by-trial basis. Specifically, we showed that, in switch trials, if the underlying neural state matched the external task condition, behavioral performance remained intact, whereas if the neural state was incorrectly represented, task performance was compromised. This pattern of results may further explain why higher decoding accuracy was reported together with limited behavioral switch costs in Waskom et al.’s study. As the adaptive coding hypothesis suggests (Waskom et al., 2014; Duncan, 2001, 2010), our findings demonstrated the dynamic adjustment of task representations can be tracked by large-scale neural activity from task-relevant brain networks on a trial-by-trial basis and provided direct evidence to associate delayed neural transitions with behavioral switch costs.

Our study has implications for a better understanding of both task switching and memory suppression. First, task switching is one of the central elements of executive control. Here, we studied task switching in the context of memory (i.e., switch between memory retrieval and suppression). We found that switch-induced delays in transitions of neural states can explain compromised memory performance after task switching. This principle could be tested and used to explain switch costs in other contexts: Using neural activity-based decoding approaches, we can track whether the neural state (of the previous task) lingers on and probe its influences on the performance (of the current task). Second, memory suppression is an experimental paradigm to mimic attempts to intrusive traumatic memories. Understanding how we can better suppress intrusive traumatic memories might pave the way for new interventions and treatments for posttraumatic stress disorder and other affective disorders (Mary et al., 2020). However, memory suppression is usually challenging in daily life. Our results suggest that task switching and inherent switch cost may explain this challenge. Usually, there are far more moments when we try to retrieve the information instead of suppressing memories, and it means that we need to switch more often from a more habitual retrieval state to a less often implemented suppression state. According to our empirical results, to minimize the negative effects of switch costs and maximize suppression-induced forgetting, potential
suppression-based interventions may benefit from “long blocks” of suppression trials to prevent the carryover effect of switching from retrieval to suppression.

Limitation and Future Directions
We used distinct locations on maps as memory cues to pair with pictures as to-be-remembered associations (i.e., picture–picture pairs) instead of word–word or word–picture pairs, which were more frequently used in previous TNT studies (Levy & Anderson, 2012; Anderson et al., 2004; Anderson & Green, 2001). Our material (1) could potentially induce eye movements as noise and (2) lead to accidental memory retrieval/suppression based on uncued locations during visual search for memory cues. These explanations can only be investigated in future studies in which eye tracking and neural responses are collected simultaneously. Furthermore, the fast sequential reactivation of memory traces during eye movement can be potentially detected based on EEG/MEG signals, whose temporal resolution is much higher than fMRI. However, despite these potential limitations, we have a reason to think that the utilization of picture–picture pairs was not problematic because we replicated traditional TNT effects on both the behavioral and neural levels (Levy & Anderson, 2012; Anderson et al., 2004; Anderson & Green, 2001). We propose that using locations on maps as memory cues is important for neural state decoding analyses presented in this study and decoding for episodic memory trace presented in another publication of our laboratory (W. Liu, Kohn, et al., 2020) because visual and semantic processing can in this way be better controlled and more consistently compared with distinct picture/word cues. Furthermore, using maps as memory cues in the TNT paradigm opens exciting opportunities to probe how the organization principles of spatial memory (e.g., cognitive map, cognitive graph, and grid-like representation; Peer, Brunec, Newcombe, & Epstein, 2021; Behrens et al., 2018; Bellmund, Gärdenfors, Moser, & Doeller, 2018) affect the effects of retrieval practice (i.e., think) and suppression (i.e., no-think) on memory traces, which are located at the same map. In this study, we intentionally prevented the colocalizing of the same manipulation on certain parts of the map. In future studies, the same manipulation (e.g., no-think) can be assigned to specific locations on a particular map to see how memory suppression affects changes as a function of Euclidean distance and whether it can be generalized by a grid-like network. Lastly, because of the different levels of specificity in the task condition instructions (i.e., think instruction to recall specific memory whereas no-think instruction to suppress all related memories and thoughts), the think versus no-think contrast is not optimal, and related overall BOLD activity levels and behaviors are difficult to compare. For the neuroimaging analysis, to replicate previous memory suppression findings, we have no choice but to contrast think trials and no-think trials. Consequently, differences in overall BOLD activity levels between conditions may reflect the differences in difficulty or levels of specificity. For the behavioral data, we intentionally analyze retrieval and intrusion rating separately instead of performing the interaction analysis like the conventional task-switching studies. Lastly, contrary to our prediction, the additional whole-brain decoding analysis demonstrated the similar ability to track the processing dynamics during task switching as the network-specific decoding. These results may indicate that our decoding results may not be specific to processing dynamics within the memory retrieval and inhibitory control networks or that it is the whole-brain representation being delayed. Future studies could further investigate the relationship between task state and whole-brain patterns of activity (Shine et al., 2019).

In summary, our results provide neural insights into the flexible task switching between memory retrieval and memory suppression. We found evidence for switch costs in memory suppression: It is more difficult to suppress unwanted memories immediately after memory retrieval. During switching between retrieval and suppression, we observed delayed transitions of neural states that each of them separately represents the current task condition. Delayed neural transitions were associated with switch costs (i.e., unsuccessful suppression and retrieval), which directly support previously proposed cognitive theories that switch costs could be the result of the carryover of previous task set activation. These results provide insight into the critical role of dynamically adjusted neural reconfigurations in supporting flexible memory suppression and the broader neural mechanisms by which humans can flexibly adjust their behavior in ever-changing environments.

APPENDIX
Behavioral Performance during the Final Memory Test
During the final memory test, each memory cue was presented again, and the participant was instructed to rate the confidence of their memory for this association and then classified the category of the associated picture. We examined the effect of memory retrieval and suppression during the TNT on the subsequent subjective (confidence rating) and objective (if they selected the correct category) memory. Three kinds of associations (i.e., retrieval association, suppression association, and control association) did not differ in their objective recall accuracy, \(F(2, 26) = 0.524, p = .595, \eta^2 = .02\) (Figure A2A). Our experiment design may explain the lack of the main effect of modulation on objective memory at the group level: All associations underwent overnight consolidation, thus difficult to be modulated (W. Liu, Kohn, et al., 2020; Y. Liu et al., 2016). Crucially, replicating the previous study (Levy & Anderson, 2012), we found the suppression-induced
forgetting at the individual level, and this effect was determined by individual differences in the efficiency of suppression during the TNT task. Specifically, participants who were more effective in suppressing intrusions (more negative intrusion slope score) during the TNT phase were the ones who show larger suppression-induced forgetting effects \( r = .411, p = .03 \); Figure A2C). Next, analyzing the subjective memory, we found a significant effect of modulation on subjective memory, \( F(2, 26) = 5.928, p = .005, \eta^2 = .186 \) (Figure A2B). Participants reported higher confidence for retrieval associations compared with control associations \( t = 3.35, p_{holm} = .007 \) and a trend toward higher confidence compared with suppression associations \( t = 2.172, p_{holm} = .07 \). Finally, we asked if modulation affected retrieval speed indexed by the RT during the final test. Even though we did not find a significant main effect of modulation, \( F(2, 26) = 2.905, p = .06, \eta^2 = .03 \) (Figure A2D), recall of “retrieval associations” was faster compared with the recall of “control associations,” \( t(26) = -2.486, p = .02 \), Cohen’s \( d = -0.47 \).

Additional Analyses of TNT Neural State Transition

To describe further the reconfiguration, we divided all ROIs into three groups (increased group, stable group, and decreased group) based on their relative changes in rank. When the task condition changed from think to no-think, 47.88% of the memory-related regions showed the top one third decrease in relative rank values and therefore belonged to the decreased group. Another 39.43% of the memory-related regions did not change extensively during the transition (stable group), and

![Figure A1. Schematic of the experiment design.](image)
12.67% of the regions showed increases in their ranks (increase group). Control-related regions demonstrated the opposite neural changes: 75.86% of them belonged to the increased group, whereas 17.24% and 6.89% of their regions belonged to a stable group or decreased group separately. For overlap regions, 50% of them belonged to the increased group, 40% of them belonged to the stable group, and only one region belonged to the decreased group. We further looked at the proportion of memory-related regions, control-related regions, and overlap regions within increased, decreased, and stable groups separately. A chi-square test of independence

**Figure A2.** Comparison between brain networks generated by Neurosynth and BrainMap. (A) Memory retrieval network. The network generated by Neurosynth using the term “memory retrieval” (left). The network generated by BrainMap was based on activation-based mapping in healthy participants who perform the paradigm of “episodic recall” (right). (B) Inhibitory control network. The network was generated by Neurosynth using the term “inhibitory control” (left). The network generated by BrainMap based on activation-based mapping in healthy participants who perform the paradigm of “go/no-go” (right).

**Figure A3.** Neural state reconfiguration during the TNT transition. (A) Visualized based on the rank of activity intensity among ROIs. (B) Visualized based on raw activity intensities of ROIs. (C) Visualized based on the Z score of activity intensity of ROIs.
Figure A5. Behavioral performance during the final memory test. (A) There is no effect of retrieval or suppression on the accuracy of the categorization during the final test ($p = .595$). (B) For retrieval associations, participants reported higher subjective confidence compared with suppression associations, $t(26) = 2.172$, $p_{holm} = 0.07$, Cohen’s $d = 0.41$, and control associations, $t(26) = 3.35$, $p_{holm} = 0.007$, Cohen’s $d = 0.64$. (C) Participants who are more effective in reducing suppression failures (more negative the intrusion slope score) were the ones who show more evidence of suppression-induced forgetting (more negative the suppression score). (D) For retrieval associations, participants spent less time during categorization compared with the control associations, $t(26) = -2.486$, $p = .02$, Cohen’s $d = -0.47$, and the effect between three conditions tend to be significant, $F(2, 26) = 2.905$, $p = .06$, $\eta^2 = .03$, a.u. = arbitrary unit.
was performed to examine the relations between their functions (i.e., memory-related, control-related, or overlap) and which change group they belong to. The relation between these variables was significant ($\chi^2 = 41.38$, $p < .001$). Control-related regions were more likely to be allocated to the increased group, whereas memory-related regions were more likely to be assigned to the decreased group. Specifically, among the increased group with a total of 36 regions (around 33.3% of all 110 ROIs analyzed), 61% of the regions were control-related regions (expected percentage = 26.4%, $p < .001$, one-side binomial test). By contrast, within the decreased group, 92% of the regions were memory-related regions (expected percentage = 64.5%, $p < .001$, one-side binomial test). Taken together, we found that, during the think condition, memory-related regions showed relatively high neural activity compared with control-related regions and overlap regions. When the task condition changed from think to no-think, memory-related regions showed decreases in their relative contribution, whereas control-related regions demonstrated an increase in their activity ranks. These patterns of changes were not only presented when we analyzed the rank of activity among ROIs (Figure A4A) but also existed when we analyzed their

Table A1. Significant Activated Clusters during No-think Trials Compared with Think Trials

| Contrast                  | Brain Region       | Hemisphere | MNI Coordinates | Cluster Size | Peak Voxel Value |
|---------------------------|--------------------|------------|-----------------|--------------|-----------------|
| No-think > Think          | IFG/insula         | L          | −38 22 8        | 920          | 5.5             |
|                           | IFG/insula         | R          | 46 18 12        | 2397         | 5.5             |
|                           | DLPFC              | R          | 22 46 24        | 1681         | 4.5             |
|                           | DLPFC              | L          | −22 44 18       | 160          | 4.5             |
|                           | IPL                | R          | 56 −42 34       | 1305         | 5.5             |
|                           | IPL                | L          | −58 −54 42      | 226          | 4.4             |
|                           | Thalamus           | R          | 16 −22 2        | 677          | 6.6             |
|                           | Precuneus          | R/L        | 14 −60 56       | NA           | 5.4             |
|                           | Postcentral gyrus  | R          | 45 −20 55       | NA           | 7.4             |
|                           | SMA                | R/L        | 8 0 53          | NA           | 5.5             |
|                           | dACC               | R/L        | 8 20 37         | NA           | 4.4             |

R = right; L = left.

Table A2. Significant Activated Clusters during Think Trials Compared with No-think Trials

| Contrast                  | Brain Region       | Hemisphere | MNI Coordinates | Cluster Size | Peak Voxel Value |
|---------------------------|--------------------|------------|-----------------|--------------|-----------------|
| Think > No-think          | mPFC               | R/L        | 0 −44 −10       | 6573         | 4.03            |
|                           | Insula             | L          | −42 −16 18      | 7.49         |                 |
|                           | ITG                | L          | −42 −42 −10     | 178          | 4.54            |
|                           | STG/IPL/precuneus  | R          | 54 −30 10       | 2925         | 5.36            |
|                           | Hippocampus        | R          | 30 −40 4        | 3.74         |                 |
|                           | Precuneus/PCC/SMA  | L          | −4 −40 34       | 6220         | 4.85            |
|                           | Precentral/postcentral gyrus | L | −40 −36 60 | 7.32         |                 |
|                           | Posterior cerebellum | R     | 22 −60 −46      | 169          | 5.30            |
|                           | Anterior cerebellum | R          | 16 −54 −20      | 1065         | 7.28            |

mPFC = medial pFC; ITG = inferior temporal gyrus; STG = superior temporal gyrus; PCC = posterior cingulate cortex; R = right; L = left.
raw (Figure A4B) or Z score (Figure A4C) of activity intensity.

**An Alternative Method to Quantify Individual Differences in Neural State Transitions**

In the main text, we presented significant correlations between the state transition index and the objective/subjective suppression effect (see Figure 2C and D). However, statistical tests were just around a significant level of $p = .05$. To further validate the relationship between neural state transitions during the TNT and the subsequent forgetting effect, we used an alternative method (i.e., state transition index V2) to quantify individual differences in neural state transitions and performed the correlation again. This method is based on the additional analysis of the TNT neural state transition above. For each participant, all 110 ROIs were divided into three groups (i.e., increased group, stable group, and decreased group) based on their relative changes in rank. The state transition index V2 was defined as the sum of the percentage of memory-related nodes within the decreased group and the percentage of control-related nodes within the increased group. As shown in Figure A5 (right), state transition index V2 was positively associated with the individual differences in objective suppression score ($r = .43, p = .02$) and tended to associate with subjective suppression score ($r = .36, p = .06$).
Table A3. Top 10 Brain Parcels with High Classification Weights during the Neural State Prediction

| Anatomical Label        | x    | y    | z    | Hemisphere | Network       | Classification Weight |
|-------------------------|------|------|------|------------|-------------------|----------------------|
| SPL                     | −34  | −48  | 46   | L          | DorsAttn         | 0.189181458           |
| LOC                     | −16  | −72  | 54   | L          | DorsAttn         | −0.358755616          |
| Middle frontal gyrus    | −32  | −4   | 52   | L          | DorsAttn         | 0.151391694           |
| dACC                    | −6   | 10   | 40   | L          | SalVentAttn      | 0.169217997           |
| IPL                     | −34  | −66  | 48   | L          | Cont             | 0.1866069             |
| IPL                     | −44  | −42  | 46   | L          | Cont             | 0.255550192           |
| Middle frontal gyrus    | −42  | 12   | 34   | L          | Cont             | −0.229290806          |
| IFG                     | −54  | 20   | 12   | L          | Default          | −0.179694589          |
| SPL                     | 16   | −72  | 54   | R          | DorsAttn         | −0.190728793          |
| Putamen                 | NA   | NA   | NA   | R          | Subcortical      | −0.161437796          |

DorsAttn = dorsal attention network; SalVentAttn = salience ventral attention network; Cont = frontal parietal control network; Default = default network; SPL = superior parietal lobule; LOC = lateral occipital cortex; L = left; R = right.

Figure A7. Individual differences in two kinds of state transition index are correlated with both objective and subjective suppression scores. The results of the state transition index (i.e., left two panels) were presented in the main text. The results of state transition index V2 (i.e., right two panels) were presented in the Appendix: An Alternative Method to Quantify Individual Differences in Neural State Transitions section.
Figure A8. Neural state-predictive pattern and overlapping network share similar spatial patterns. (A) The contribution of different brain regions during the decoding. The map was visualized using an arbitrary threshold of 0.10. Higher absolute values of voxelwise statistical results represent a larger contribution during decoding. (B) Voxels that belong to both memory retrieval network and inhibitory control network (Neurosynth-based).

Figure A9. Individual difference analyses between decoding accuracies (ACC) and behavioral performances. (A) Stronger “adaptation” effects (i.e., higher decoding ACC during the time window between TR = 3 and TR = 5) correlated with better memory retrieval performance when it was a switch trial \( r = .44, p = .02 \). (B) Stronger “adaptation” effects correlated with better memory retrieval performance when it was a nonswitch trial \( r = .43, p = .02 \). (C, D) Decoding ACC during the same time window did not associate with memory suppression performance during either switch \( (p = .64) \) or nonswitch trial \( (p = .43) \).
Acknowledgments

This research was supported by the Chinese Scholarship Council (201606990020, W. L.). We thank Joyce van Arendonk and Joost Verchuren for the assistance of data acquisition and memory performance evaluation, Sjoerd Meijer for the memory performance evaluation, and Yingjie Shi for commenting on data analysis and proofreading the final manuscript.

Reprint requests should be sent to Wei Liu, School of Psychology, Central China Normal University, No. 152 Luoyu Road, Hongshan District, Wuhan 430079, Hubei Province, China, or via e-mail: weiliu1991@mail.ccnu.edu.cn.

Author Contributions

Wei Liu: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing—Original draft; Writing—Review & editing. Nils Kohn: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Supervision; Validation; Visualization; Writing—Original draft; Writing—Review & editing. Guillén Fernández: Conceptualization; Formal analysis; Funding acquisition; Project administration; Resources; Supervision; Writing—Review & editing.

Funding Information

Wei Liu, Chinese Scholarship Council (https://www.chinesescholarshipcouncil.com/), grant number: 201606990020.

Figure A10. Behavioral consequences of the mismatch between neural state and current task condition. When the decoded neural state did not match with the task condition (i.e., think decoded as no-think), participants reported worse memory retrieval performance during think trials. When the neural decoder misclassified no-think moments as think, participants reported more memory intrusions during no-think trials. These two effects can be detected during the “switching” period (i.e., left), “nonswitching” period (i.e., middle), and all time points within the “entire” TNT task (i.e., right).

Diversity in Citation Practices

A retrospective analysis of the citations in every article published in this journal from 2010 to 2020 has revealed a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .408, W(oman)/M = .335, M/W = .108, and W/W = .149, the comparable proportions for the articles that these authorship teams cited were M/M = .579, W/M = .243, M/W = .102, and W/W = .076 (Fulvio et al., *JoCN*, 33:1, pp. 3–7). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article’s gender citation balance.

REFERENCES

Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature, 410*, 366–369. https://doi.org/10.1038/35066572, PubMed: 11268212
Anderson, M. C., & Hanslmayr, S. (2014). Neural mechanisms of motivated forgetting. *Trends in Cognitive Sciences, 18*, 279–292. https://doi.org/10.1016/j.tics.2014.03.002, PubMed: 24747000
Anderson, M. C., Ochsner, K. N., Kuhl, B., Cooper, J., Robertson, E., Gabrieli, S. W., et al. (2004). Neural systems underlying the suppression of unwanted memories. *Science,*
Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, 2, 115–128. https://doi.org/10.1038/35077255

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends in Cognitive Sciences*, 14, 172–179. https://doi.org/10.1016/j.tics.2010.01.004, PubMed: 20171926

Etzel, J. A., Cole, M. W., Zacks, J. M., Kay, K. N., & Braver, T. S. (2016). Reward motivation enhances task coding in frontoparietal cortex. *Cerebral Cortex*, 26, 1647–1659. https://doi.org/10.1093/cercor/bhu327, PubMed: 25601237

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2, 189–210. https://doi.org/10.1002/hbm.460020402

Gilbert, S. J. (2011). Decoding the content of delayed intentions. *Journal of Neuroscience*, 31, 2888–2984. https://doi.org/10.1523/JNEUROSCI.5363-10.2011, PubMed: 21414910

Gonzalez-Castillo, J., Hoy, C. W., Handwerker, D. A., Robinson, M. E., Buchanan, L. C., Saad, Z. S., et al. (2015). Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns. *Proceedings of the National Academy of Sciences, U.S.A.*, 112, 8762–8767. https://doi.org/10.1073/pnas.1501242112, PubMed: 26124112

Goschke, T. (2000). Intentional reconfiguration and involuntary persistence in task set switching. In S. Monsell & J. S. Driver (Eds.), *Control of cognitive processes: Attention and Performance* (Vol. 18, pp. 331–355). Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/481.003.0023

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, 48, 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.000, PubMed: 19573611

Graber, O., Karch, S., Schlueter, E. K., Falkai, P., & Goschke, T. (2006). Neural mechanisms of advance preparation in task switching. *Neuroimage*, 31, 887–895. https://doi.org/10.1016/j.neuroimage.2005.12.043, PubMed: 16490365

Guo, Y., Schmitz, T. W., Mur, M., Ferreira, C. S., & Anderson, M. C. (2018). A supramodal role of the basal ganglia in memory and motor inhibition: Meta-analytic evidence. *Neuropsychologia*, 108, 117–134. https://doi.org/10.1016/j.neuropsychologia.2017.11.033, PubMed: 29199109

Haynes, J.-D. (2015). A primer on pattern-based approaches to fMRI: Principles, pitfalls, and perspectives. *Neuron*, 87, 257–270. https://doi.org/10.1016/j.neuron.2015.05.025, PubMed: 26182413

Hermans, E. J., van Marle, H. J. F., Ossewaarde, L., Henckens, M. J. A. G., Qin, S., van Kesteren, M. T. R., et al. (2011). Stress-related noradrenergic activity prompts large-scale neural network reconfiguration. *Science*, 334, 1151–1153. https://doi.org/10.1126/science.1209603, PubMed: 22116887

Huang, P., Carlin, J. D., Alink, A., Kriegeskorte, N., Henson, R. N., & Correia, M. M. (2018). Prospective motion correction improves the sensitivity of fMRI pattern decoding. *Human Brain Mapping*, 39, 4018–4031. https://doi.org/10.1002/hbm.24228, PubMed: 29885014

Hulbert, J. C., Henson, R. N., & Anderson, M. C. (2016). Inducing amnesia through systemic suppression. *Nature Communications*, 7, 11003. https://doi.org/10.1038/ncomms11003, PubMed: 26977589

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, PubMed: 12377357

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *Neuroimage*, 62, 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015, PubMed: 21979382

van Marle, H. J. F., Hermans, E. J., Ossewaarde, L., Henckens, M. J. A. G., Qin, S., van Kesteren, M. T. R., et al. (2011). Stress-related noradrenergic activity prompts large-scale neural network reconfiguration. *Science*, 334, 1151–1153. https://doi.org/10.1126/science.1209603, PubMed: 22116887

Huang, P., Carlin, J. D., Alink, A., Kriegeskorte, N., Henson, R. N., & Correia, M. M. (2018). Prospective motion correction improves the sensitivity of fMRI pattern decoding. *Human Brain Mapping*, 39, 4018–4031. https://doi.org/10.1002/hbm.24228, PubMed: 29885014

Hulbert, J. C., Henson, R. N., & Anderson, M. C. (2016). Inducing amnesia through systemic suppression. *Nature Communications*, 7, 11003. https://doi.org/10.1038/ncomms11003, PubMed: 26977589

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, PubMed: 12377357

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *Neuroimage*, 62, 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015, PubMed: 21979382
reconfiguration. *Journal of Neuroscience*, 36, 8551–8561. https://doi.org/10.1523/JNEUROSCI.0358-16.2016, PubMed: 27535904

Shine, J. M., Bissett, P. G., Bell, P. T., Koyejo, O., Balsters, J. H., Gorgolewski, K. J., et al. (2016). The dynamics of functional brain networks: Integrated network states during cognitive task performance. *Neuron*, 92, 544–554. https://doi.org/10.1016/j.neuron.2016.09.018, PubMed: 27693256

Shine, J. M., Hearne, L. J., Breakspear, M., Hwang, K., Müller, E. J., Sporns, O., et al. (2019). The low-dimensional neural architecture of cognitive complexity is related to activity in medial thalamic nuclei. *Neuron*, 104, 849–855. https://doi.org/10.1016/j.neuron.2019.09.002, PubMed: 31653463

Shine, J. M., & Poldrack, R. A. (2018). Principles of dynamic network reconfiguration across diverse brain states. *Neuroimage*, 180, 396–405. https://doi.org/10.1016/j.neuroimage.2017.08.010, PubMed: 28782684

Siegel, J. S., Mitra, A., Laumann, T. O., Seitzman, B. A., Raichle, M., Corbetta, M., et al. (2017). Data quality influences observed links between functional connectivity and behavior. *Cerebral Cortex*, 27, 4492–4502. https://doi.org/10.1093/cercor/bhw253, PubMed: 27550863

Smith, S. M. (2002). Fast robust automated brain extraction. *Human Brain Mapping*, 17, 143–155. https://doi.org/10.1002/hbm.10062, PubMed: 12391568

Spector, A., & Biederman, I. (1976). Mental set and mental shift revisited. *American Journal of Psychology*, 89, 669–679. https://doi.org/10.2307/1421465

Todd, M. T., Nystrom, L. E., & Cohen, J. D. (2013). Confounds in multivariate pattern analysis: Theory and rule representation case study. *Neuroimage*, 77, 157–165. https://doi.org/10.1016/j.neuroimage.2013.03.039, PubMed: 23558095

van der Bij, A. K., de Weerd, S., Cikot, R. J. L. M., Steegers, E. A. P., & Braspenning, J. C. C. (2003). Validation of the Dutch short form of the state scale of the Spielberger State-Trait Anxiety Inventory: Considerations for usage in screening outcomes. *Community Genetics*, 6, 84–87. https://doi.org/10.1159 /000073003, PubMed: 14560068

Waskom, M. L., Kumaran, D., Gordon, A. M., Rissman, J., & Wagner, A. D. (2014). Frontoparietal representations of task context support the flexible control of goal-directed cognition. *Journal of Neuroscience*, 34, 10743–10755. https://doi.org/10.1523/JNEUROSCI.5282-13.2014, PubMed: 25100605

Westphal, A. J., Wang, S., & Rissman, J. (2017). Episodic memory retrieval benefits from a less modular brain network organization. *Journal of Neuroscience*, 37, 3523–3531. https://doi.org/10.1523/JNEUROSCI.2509-16.2017, PubMed: 28242796

Wisniewski, D., Goschke, T., & Haynes, J.-D. (2016). Similar coding of freely chosen and externally cued intentions in a fronto-parietal network. *Neuroimage*, 134, 450–458. https://doi.org/10.1016/j.neuroimage.2016.04.044, PubMed: 27107470

Wisniewski, D., Reverberi, C., Tusche, A., & Haynes, J.-D. (2015). The neural representation of voluntary task-set selection in dynamic environments. *Cerebral Cortex*, 25, 4715–4726. https://doi.org/10.1093/cercor/bhu155, PubMed: 25057922

Woolgar, A., Afshar, S., Williams, M. A., & Rich, A. N. (2015). Flexible coding of task rules in frontoparietal cortex: An adaptive system for flexible cognitive control. *Journal of Cognitive Neuroscience*, 27, 1895–1911. https://doi.org/10.1162/jocn_a_00827, PubMed: 26058604

Woolgar, A., Thompson, R., Bor, D., & Duncan, J. (2011). Multivoxel coding of stimuli, rules, and responses in human frontoparietal cortex. *Neuroimage*, 56, 744–752. https://doi.org/10.1016/j.neuroimage.2010.04.035, PubMed: 20406990

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Spatial autocorrelation in univariate linear modeling of fMRI data. *Neuroimage*, 14, 1370–1386. https://doi.org/10.1006/nimg.2001.0931, PubMed: 11707093

Yang, W., Zhuang, K., Liu, P., Guo, Y., Chen, Q., Wei, D., et al. (2017). Choosing the rules: Distinct and overlapping frontoparietal representations of task rules for perceptual decisions. *Journal of Neuroscience*, 33, 11852–11862. https://doi.org/10.1523/JNEUROSCI.5193-12.2013, PubMed: 23864675