Analogous computations in working memory input, output and motor gating: Electrophysiological and computational modeling evidence

Rachel Rac-Lubashevsky and Michael J. Frank

Affiliations:

Rachel Rac-Lubashevsky, Brown Institute for Brain Sciences, Department of Cognitive Linguistic and Psychological Science, Brown University.

Michael J. Frank, Brown Institute for Brain Sciences, Department of Cognitive Linguistic and Psychological Science, Brown University.

Corresponding author:
Rachel Rac-Lubashevsky
Department of Cognitive, Linguistic & Psychological Sciences
Brown University
190 Thayer St
Providence RI 02912-1821
Email: rac.hunrachel@gmail.com

Abstract

Adaptive cognitive-control is achieved through a hierarchical cortico-striatal gating system that supports selective updating, maintenance, and retrieval of useful cognitive and motor information. Here, we developed a novel task that independently manipulated selective gating operations of working-memory (input), from working-memory (output), and in response (motor) and tested the neural dynamics and computational principles that support them. Increases in gating demands, captured by gate switches, were expressed by distinct EEG correlates at each gating level that evolved dynamically in partially overlapping time windows. EEG decoding analysis further showed that neural indexes of working-memory (category) and motor (action) representations were prioritized particularly when the corresponding gate was switching. Finally, the control mechanisms involved in gate switches were quantified by the drift diffusion model, showing elevated motor decision threshold in all gating levels. Together these results support the notion that cognitive gating operations scaffold on top of mechanisms involved in motor gating.

Introduction

Optimal flexible behavior requires an agent to not only respond to incoming sensory events but to adaptively adjust action selection based on context, including previous events in memory [1]. Moreover, while some events in memory need to be robustly maintained over time in the face of distracting interference, sometimes sensory events dictate that such memories should be disrupted and rapidly updated. This challenge is referred to as the stability vs flexibility tradeoff [2-4] and highlights the need for a context-dependent control mechanism that selectively gates information into and out of working-memory (WM) to guide actions [5-10]. Adaptive control is particularly crucial given a capacity-limited WM system in a complex environment, where only a subset of perceptual information is task-relevant and only a subset of currently maintained WM items may be useful for guiding ongoing behavior.

The PBWM (prefrontal cortex basal ganglia working memory) model is a computational model that leverages powerful mechanisms of dopaminergic reinforcement learning (RL) in basal ganglia (BG) so as to optimize gating policies that control access to and from prefrontal cortex, to be held in WM [9-11]. According to PBWM, gating of both motor and WM actions are implemented by a common canonical set of operations whereby phasic DA signals reinforce BG gating actions that yield successful task performance. A distinguishing characteristic of the PBWM model is that WM gating is an elaboration of the more established BG mechanism of motor gating [7,9,11]. Indeed, according to PBWM, WM gating can be further divided into *input gating*, whereby BG can control access to stored information in PFC, and *output gating*, whereby other BG circuits can control which among several maintained WM items should influence motor decisions (*response gating;* see Fig. 1*)*. Moreover, PBWM extensions further assume that frontostriatal gating processes operate across a spectrum of abstract actions ranging from action

plans in premotor cortex (PMC) to context-relevant cognitive information (like rules and task sets) in more anterior PFC [11-13]. Various imaging, lesion and pharmacological studies have provided support for PBWM gating mechanisms (see [14-16] for reviews). These cognitive and motor gating circuits are nested within a cortico-striatal hierarchical system that is arranged on a rostral-caudal axis in the frontal lobe (Fig .1) [17-20], such that the posterior response gate is dependent on, and constrained by, the more rostral cognitive gates [6,11,12]. Indeed, anatomical studies have revealed asymmetric topography with more fibers projecting information from anterior PFC to posterior BG [21]. Nevertheless, despite this converging evidence for hierarchical PFC-BG gating interactions, the core assumption that response, input and output gating share computational properties – or how they unfold in time – has not been rigorously tested.

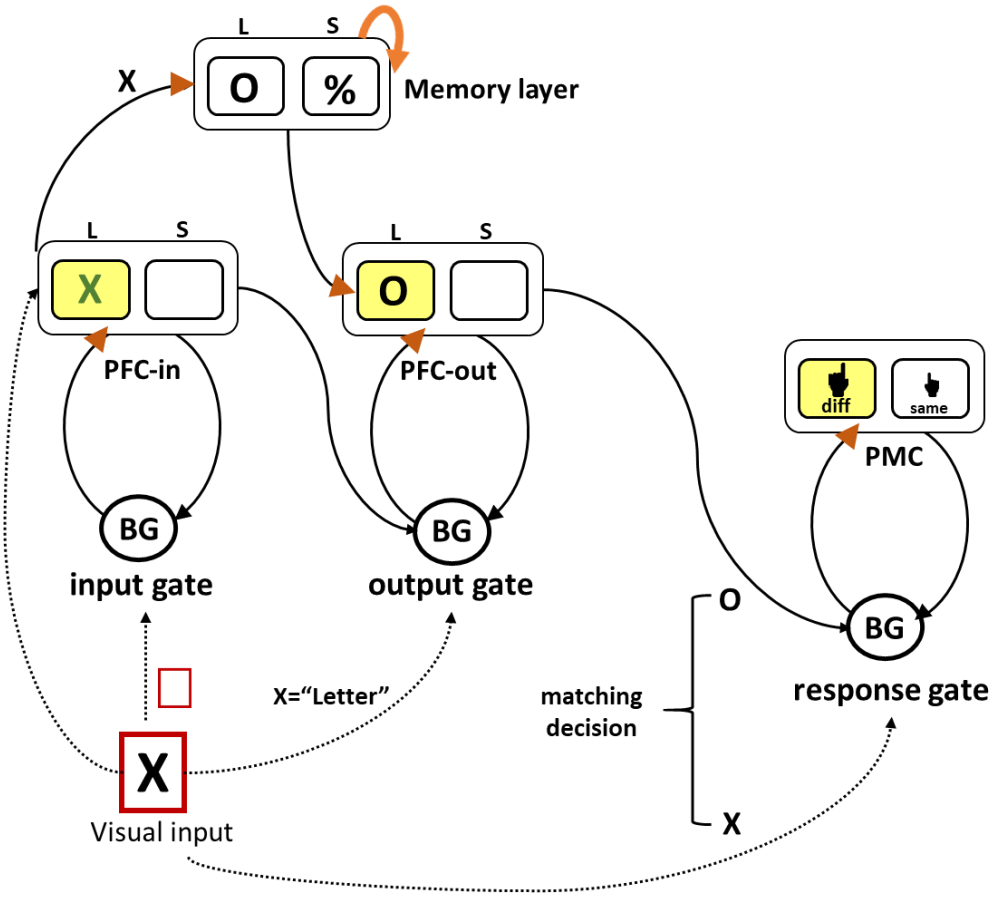

Figure 1. Schematic of the PBWM architecture for canonical gating operations. The gating system is composed of three core circuits that are hierarchically arranged from left to right. For illustration the operations needed to solve the task used in this paper are shown. Sensory information (e.g., the "X" in the visual input) is transiently represented in PFC-in. The input gate controls whether this information is updated to be maintained in a PFC memory layer, and if so, to which address or anatomical "stripe" within that layer (selective gating). The red frame represents a task cue signifying that memory should be updated, and the input gate selectively updates the left PFC stripe, replacing the "O" in memory with "X". The output gate controls which of the memory layer representations is represented in PFC-out (e.g., a deep layer of the PFC) to influence subsequent processing (e.g., here the relevant representation in memory is "O" and not "%", given that the input to be compared is in the letter category). The response gate controls which motor response to select (here, "same" or "different") by comparing the output-gated memory representation ("O") with the current visual input ("X"). At the beginning of a trial, the memory layer actively maintains the information from the previous trial. The orange arrow on the memory layer represents active maintenance. Dark orange arrow heads represent the gating signal by the BG that selects the relevant representation. The yellow squares represent the PFC stripe selected by the corresponding gating operation.

The N-back task is often used for examining variations in executive function [22-23]. While the N-back likely depends on both input and output gating (and can be modeled with PBWM [24]), experimentally, it provides a coarse measurement of WM control processes that are confounded with a variety of complex cognitive processes (e.g., encoding, inhibition, binding, matching, maintenance, updating and removal). To address this issue, and to focus on input gating specifically, we recently developed the reference-back task [25-29], a continuous WM updating task. Behavioral, EEG and fMRI data supports the notion that the reference-back task taxes BG input gating of PFC [27,29,30]. However, a core distinguishing feature of the PBWM framework from that of other gating models is that it affords *selective* gating, whereby items can be input gated to, and output gated from, distinct addresses in memory (represented as PFC "stripes" or ensembles in the PBWM model). Here we aimed to augment the reference-back task to more directly assess these separable content-addressable input and output gating functions, while also assessing how they relate to response gating.

To do so, we amended the reference-back task to mirror the SIR2 task, one of the key tasks developed to illustrate the need for learning selective gating policies in PBWM [10]. To manipulate the need for selective gating in a content-addressable manner, the SIR2 task includes two separate store cues (S1 or S2) in which the associated WM items had to be updated independently to separable WM stores and then accessed independently in response to corresponding recall probes (R1 or R2), in arbitrary order and over intervening distractors [10]. Through RL, PBWM learned a selective input gating policy that placed S1 and S2 items into separate PFC stripes, and an output gating policy that accesses the appropriate stripe given the corresponding R1 or R2 probe, thus implementing content-addressable memory. Given the unpredictable order in which items are stored and recalled, models without selective gating policies cannot solve the SIR2 task. This same selective gating ability was shown to useful to perform complex, hierarchically structured tasks and for supporting generalizability and task structure learning (11, 12, 31, 32]. However, the SIR2 task does not have psychometric properties useful for human experimentation.

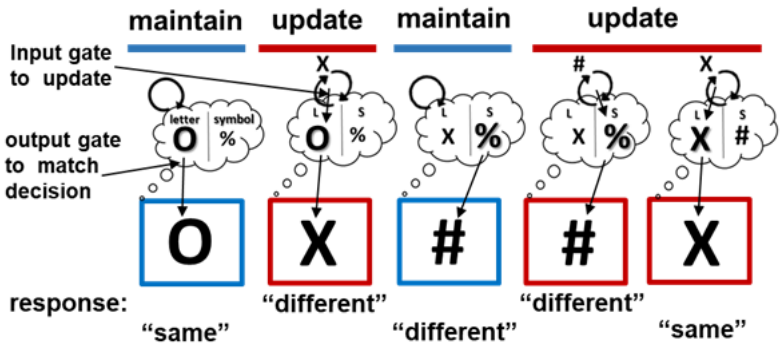

Figure 2. Trial events in the reference-back-2 task. Trials with red frame are updating trials, while blue trials are maintenance trials. On each trial, participants make a matching decision between the probe and last stimulus presented in red from the same category. Each side of the thought bubble represents an address (or PFC "stripe") in WM. During updating, input gates assign information to an address based on its category. During output gating, the relevant category is accessed from memory for the decision. The response gate makes the same/different selection. See the method section for more detail.

The reference-back-2 task (Fig. 2) retains all the key features of SIR2 and allows laboratory testing of content-addressable selective input and output gating. The reference-back-2 task, like the original reference-back, is a continuous WM updating task that is composed of two trial types. *Updating trials* require both a matching decision and WM updating, whereas *maintenance trials* are equivalent in their perceptual and decision making demands but do not require WM updating. To introduce the need for selective gating, the reference-back-2 task includes stimuli that belong to two different categories (e.g., symbol/letter), where only one of these representations is relevant on each trial. In updating trials (marked with a red frame), participants need to update only the stimulus in WM belonging to the same category as the presented stimulus, while continuing to maintain that of the other category; thus demanding selective input gating (see Fig. 1). The input gate, gates relevant information (e.g., the probed "X") into a specific WM memory address (stripe) given the appropriate context (e.g., the red frame indicates to update). Further, for making the matching decision, only the currently relevant category item should be considered; thus demanding selective output gating. <u>The output gate</u> selects amongst maintained WM representations (in the memory layer) and output gates relevant information for subsequent processing (e.g., given the letter category in the probe, the relevant representation in memory is "O" and not "%"). Finally, for motor responding, the participant needs to compare the probed stimulus ("X") to the WM representation selected by the output gate ("O"), in order to respond "same" or "different", thus demanding selective response gating.

Here, we aim to test three key predictions from PBWM. First, we assessed whether input, output, and response gating share analogous computations that evolve dynamically in time. In particular, PBWM predicts that when gating demands (at any level) switch from one trial to the next, the resulting conflict is associated with an independent response time cost.

Neurally, such switches give rise to a transient period of conflict at the corresponding region of PFC, analogous to the conflict experienced when needing to switch motor responses in PMC [33] (see also Fig. 4 below). Switches in higher level task rules give rise to anterior EEG signatures which precede those of more posterior motor EEG signatures (Collins et al., 2014). We thus assessed whether such neural markers of input, output and response gating are separately detectable, and whether they evolve serially or in parallel.

Second, we tested the core PBWM prediction that output gating is used to enhance the representation associated with the currently relevant category (bolded representation in Fig 2). Previous work has shown that attention to a relevant category, during either encoding or subsequent retrieval, acts to prioritize the attended representation and improve its memory (see reviews by [35-37]). WM prioritization is marked by sustained, elevated neural activity [38-40] and persistent alpha suppression [41] that increases decodability of the prioritized representations [42-44]. Such WM prioritization is thought to support "controlled attention" by enhancing top-down influence on subsequent processes [36, 45]. Such an account is consistent with output gating, but to date the gating WM literature and the WM prioritization literature have not been linked. To this end, we aimed to directly test the prediction that output gating (and switches thereof) places demands on prioritization to increase neural discriminability between the two candidate categories. Analogously, we also predicted that response switching will boost the neural decodability of the corresponding motor action representation, given that switches at the motor level place additional demands on response gating.

Finally, we tested whether the behavioral costs associated with gate switching can be assessed more quantitatively by computational models that summarize the impact of BG gating on response time distributions. The behavioral dynamics of BG gating neural networks at the

motor level can be approximated by the drift diffusion model (DDM) [46]. Moreover, in these models and related electrophysiological data, premotor conflict (e.g., following switches), lead downstream in the BG to an increase in the effective "decision threshold", buying more time to allow more cautious and deliberate response selection [33, 47-49]. Hierarchical PFC-BG models further suggest that this same computation is recruited when conflict occurs at higher level prefrontal task representations, preventing the lower level corticostriatal circuit from selecting actions until such conflict is resolved [12]. Thus according to PBWM, conflict at the WM levels (switches in WM gating) should give rise to an increase in the motor response decision threshold. We thus aimed to test whether neural markers of response switching and conflict are also impacted by switches of WM input and output gating, and whether they are accompanied by a concomitant change in decision threshold.

In sum, the current study was designed to test the hypotheses that selective input, output and response gating are managed by analogous computations and that they are distinguishable temporally. Very little is known on the order in which selective cognitive and action decisions are operating and how they unfold in time. This is important, in part, because input and output gating rely on close cortico-striatal circuits that may require high temporal resolution to disentangle. Furthermore, we investigate the functional role of WM gating and response gating. We employ a trial-by-trial decoding approach of EEG signals to quantify the implications of selective gating on the prioritization of WM and action representations. Finally, we will examine if the same control mechanisms that are engaged to response conflict during motor switching are also recruited in response to cognitive conflict arising from multiple items or task rules held in WM. To quantify these dynamics, we will employ an abstract mathematical model, the hierarchical drift-diffusion model (HDDM) [33,50].

Results

We first assessed the most basic prediction that switches at each gating level (input, output, response) will translate into separable costs on behavior. Three-way repeated-measures ANOVAs were run on mean RT and mean error rate as a function of switches in input (update, maintain) × output category (letters, symbols) × response (same, different).

For RT, significant main effects were observed for switches at all levels: input ($F_{1,29}$= 103.99, p<.001, η_p^2=.78), output ($F_{1,29}$= 54.06, p<.001, η_p^2=.65) and response ($F_{1,29}$= 5.40, p=.03, η_p^2=.16). The faster RT for repeats than for switches is consistent with the canonical gating model in which a switch at any level incurs conflict and delays motor responding. Significant under-additive interactions were also observed between input × output ($F_{1,29}$= 57.81, p<.001, η_p^2=.67), input × response ($F_{1,29}$= 81.51, p<.001, η_p^2=.74), and output × response ($F_{1,29}$= 36.51, p<.001, η_p^2=.56). The three-way interaction was also significant ($F_{1,29}$= 18.39, p<.001, η_p^2=.39) indicating that the under-additive interactions between switches at two levels is limited to the case where the third level was not switching.

For error rate, the only significant effect was for the 2-way input × response interaction ($F_{1,29}$= 43.00, p<.001, η_p^2=.60), such that switches in input gating demands were related to increased accuracy for response switching. This result indicates a speed-accuracy tradeoff given the RT results from that same interaction; this will be further investigated with the DDM model below. None of the main effects nor the other interactions were significant ($F_{1,29}$<3.7, p>.06). Note that an extended five-way ANOVA was also conducted with WM state (updating, maintenance), updating frequency (rare, frequent) and the three gating levels (switch, repeat) in

input, output and response gating. The results of this larger analysis are reported in Supplementary Results.

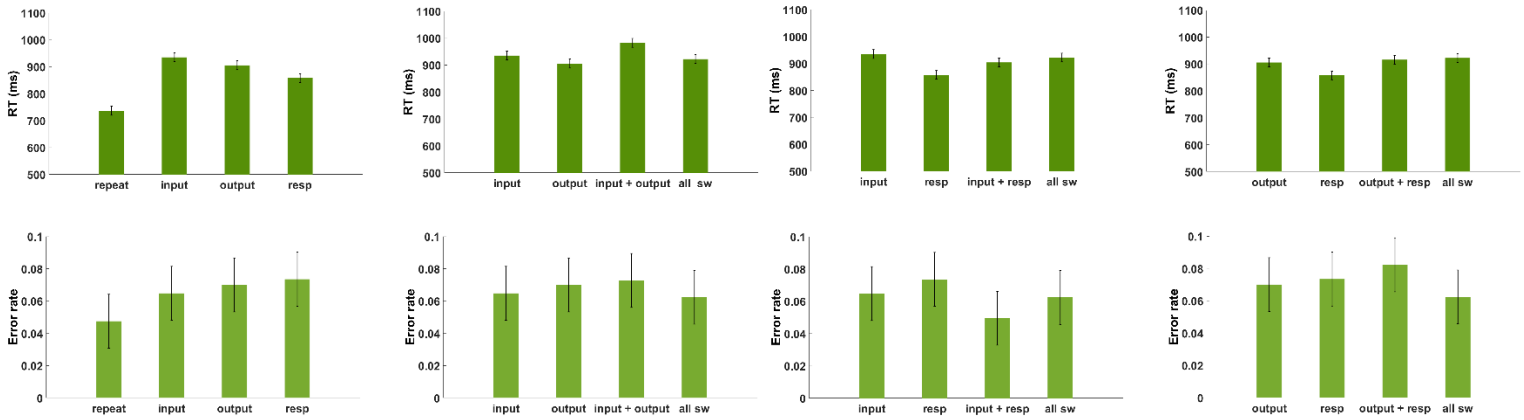

Figure 3. Behavioral effects of switching in gating demands. Mean RT (top row) and error rate (bottom row) demonstrate differences in performance across all levels of gate switching (left column), together with under-additive interactions at all levels (from left to right: input and output, input and response, output and response and when all three gates are switching).

Temporal evolution of neural gating dynamics across levels.

Given that switches in gating demands at all levels had observable effects on behavior, but with under-additive interactions, we also assessed whether such switches would be observable in neural activity. In PBWM, gating signals are used to displace prefrontal activity states with new information. Thus, the impact of gating is most observable during switch trials, which induce a transient period of conflict in cortex (the equivalent of an ERP in the model) when there is competition between the previous and currently relevant information (Fig 4a-b; [12,33, 34]. Notably, these models are hierarchical, such that neural signatures of gate switching evolve in time, with higher level gates inducing conflict prior to the response gate (Figure 4a; [34]). However, despite this seemingly serial process, the same model predicts that these processes evolve partly in parallel, leading to under-additive interactions. Indeed, previous

behavioral work has showed evidence for a combination of parallel and serial processing among cognitive and response selections, a result that was captured using the same hierarchical PFC-BG neural network model [51].

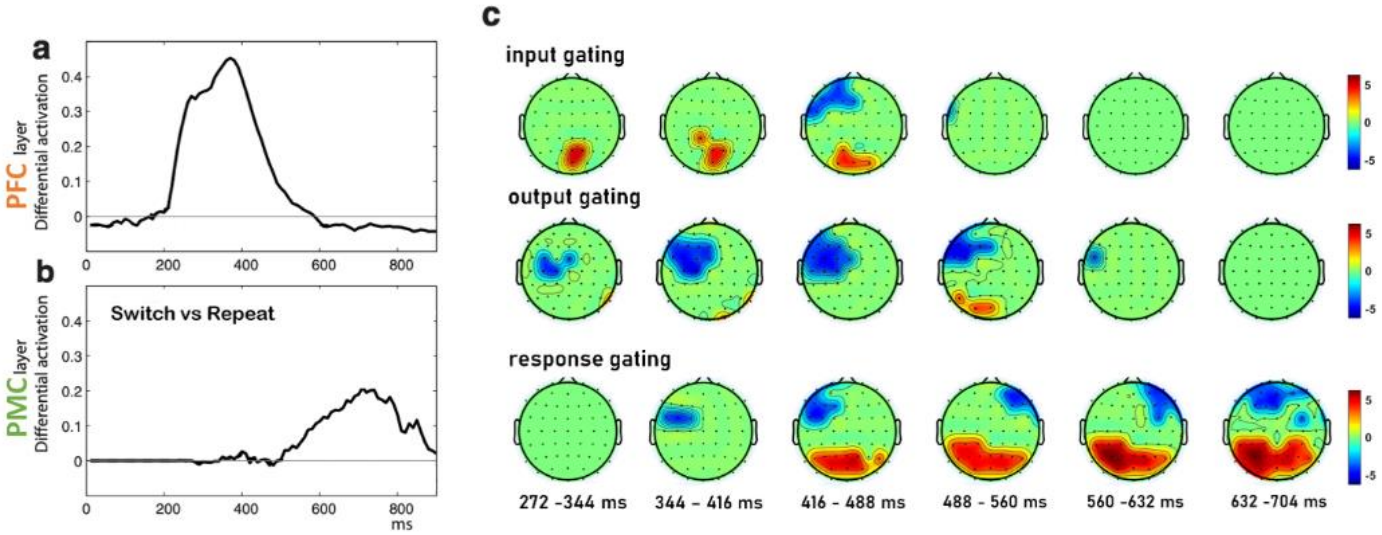

Figure 4. (a-b) Summed activation over units in PFC-BG neural network, showing impact of gate switch vs repeat (a) early in PFC and (b) later in motor cortex (adapted from [34]). (c) Current study. Scalp map topography for the predictors that survived permutation correction for the three gating types (input, output, response in three rows from top to bottom). The color in the scalp map represent the t-value of the average regression weights in each electrode at the time range indicated above each scalp map. Input, output and response gating signatures evolve sequentially in time, but with clear temporal overlap.

To evaluate whether gating processes are observable and whether they evolve in parallel or in serial, we focus on trial-to-trial switches that place demands on input, output, and response gating independently. Accordingly, we leverage three orthogonal switch versus repeat trial contrasts to index gating demands: (a) Input gating: The transition from maintenance trials to updating trials requires a shift from a closed to an open input gate, in order to selectively update the task-relevant category. These switch trials are compared to update repeat trials. (b) Output gating: The relevant category for guiding the match decision is selected from within memory using output gating. When the category switches across trials, a new address must be accessed, placing higher demands on the output gate relative to when the category repeats. (c) Response

gating: Switches of the same/different response should result in transient conflict at the level of the response decision relative to repeats.

We used a regression approach to extract spatiotemporal clusters that reflect the three gating effects on the trial-by-trial EEG signal, controlling for RT and multiple comparisons (see Methods; [52-53]). The results of this regression analysis yielded significant univariate patterns for WM input gating, output gating and response gating (Fig 4). Notably, the neural markers of all three gating levels (input, output and response) evolved dynamically in order (Fig 4), but with substantial overlap in time (410ms – 470ms), consistent with previous empirical studies and hierarchical computational models of cortico-striatal circuitry [12; 51]. The response index also persisted on its own after the input and output gate signatures (see Fig 4). The distinct univariate patterns of activity for each gating level, with largely overlapping time windows, provides neural and temporal support for the PBWM prediction that WM is managed by independent gate selection mechanisms operating sequentially but mostly in parallel [51].

Gating as a prioritization mechanism.

Having identified putative neural signatures of gating operations, we next sought to assess the functional impact of gate switching in terms of prioritization of task-relevant representations. The reference-back-2 task employed in this study is suited to study this question: in every trial, two representations are held in WM. Only one, however, is relevant to guide action selection (and to be selectively updated in WM when needed), while the other representation is prospectively relevant and therefore still requires maintenance. This design allowed us to test if selection within WM (output gating) shares the same neural manifestation as action selection (response gating). We leveraged a trial-by-trial decoding approach [52-53]. Thus in contrast to

the last section, in which we identified neural signatures of gate switching, here we used GLM to extract spatiotemporal clusters in the EEG signal that reflect the representations themselves (e.g., relevant category and response), controlling for RT, WM state (update, maintenance) and multiple comparisons ([52-53] see methods).

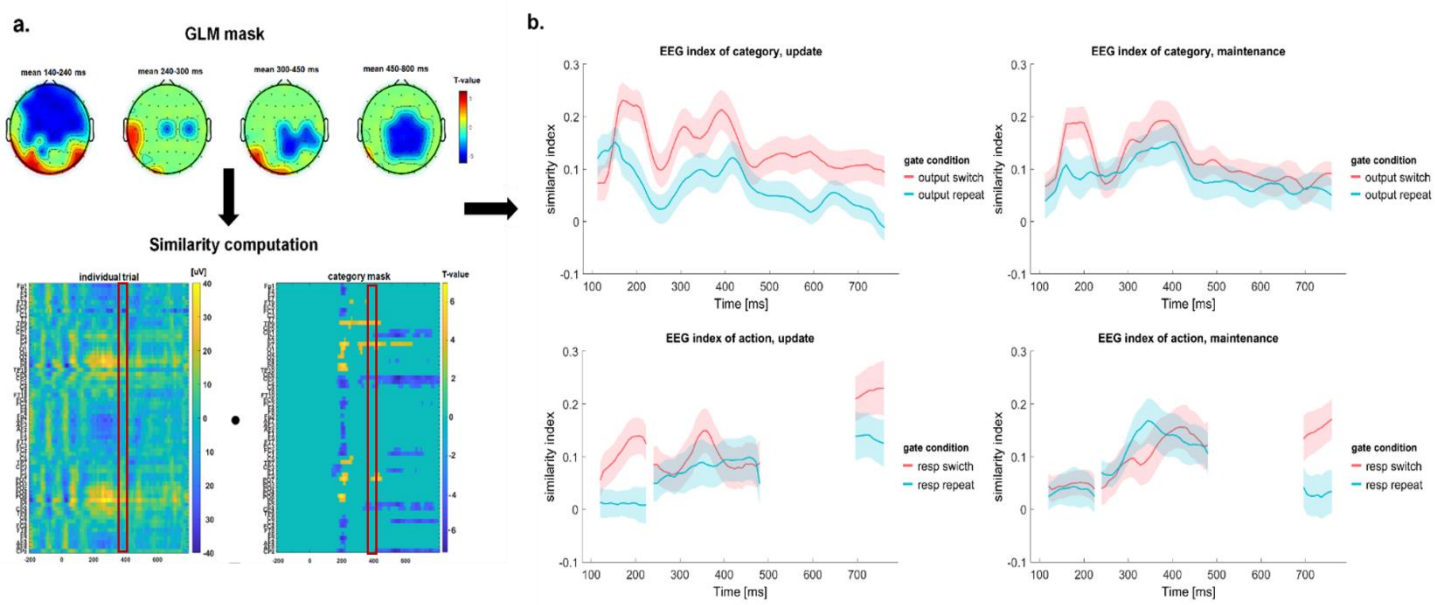

Figure 5. The temporal dynamics of neural prioritization by gating. (a) Trial-by-trial decoding of gated representations. Univariate patterns were identified for each representation of interest (category, response) using GLM. Trial by trial similarity indices obtained by computing the dot-product between the 2D voltage-time map on individual trials with that of the mask for the relevant representation. This similarity index is obtained over sliding window time bins (red rectangle). The result is a temporally evolving EEG index of similarity in time windows for which mask was significant (see methods for more detail). (b) The temporal dynamics of similarity indices to the relevant representations in updating (left) and maintenance (right) trials. While neural similarity to the relevant representation was evident in all cases, category level decoding was enhanced following switches in output gating, whereas response decoding was enhanced following switches in responses. The shaded error bars represent standard error of the mean (SEM) to reflect the time points where the difference between switch and repeat was significant.

Supporting the notion that switches at the relevant level enhance the need for gating, we found that the trial-wise neural similarity increased following switches of the relevant gate (see figure 5b). Specifically, output switching enhanced the neural difference between the two

categories whereas response switching enhanced the neural difference between the two motor action representations. These increases in decodability were very transient in maintenance trials (only around 200ms) while in updating trials, the benefit for output switching lasted throughout the trial (170-800ms; see Fig 5b). This finding is consistent with the PBWM model scheme in which selective updating of WM further enhances activity in the memory layer (Fig.2) which is compounded with impact of output and response switching operations. Together these results support the notion that selective gating at both cognitive and the motor level share the functional role of prioritization that increases decodability of the representation selected by the relevant gate.

Gate switching at any level increases decision threshold.

The behavioral results described above revealed that switches at any level of gating were related to increased RT cost. To further decompose separable cognitive processes that give rise to such changes in RT, we leveraged the drift diffusion model (DDM). As noted in the introduction, various studies have suggested that dorsomedial frontal cortex detects response conflict (which is elevated following switches) and it recruits BG mechanisms that increase the effective decision threshold [46-49], resulting in slower but more accurate decisions. Hierarchical PFC-BG models further suggest that conflict at the level of PFC representations recruits the same mechanism to prevent premature motor responding [12, 33]. We thus assessed whether gate switches at all levels of the WM hierarchy are quantitatively related to adjustments in decision threshold (thought to reflect cognitive control, thus related to slower but more accurate responding), or whether any such effects could be attributed to alterations in drift rate (thought to reflect increased difficulty and thus slower and less accurate) (Fig. 8).

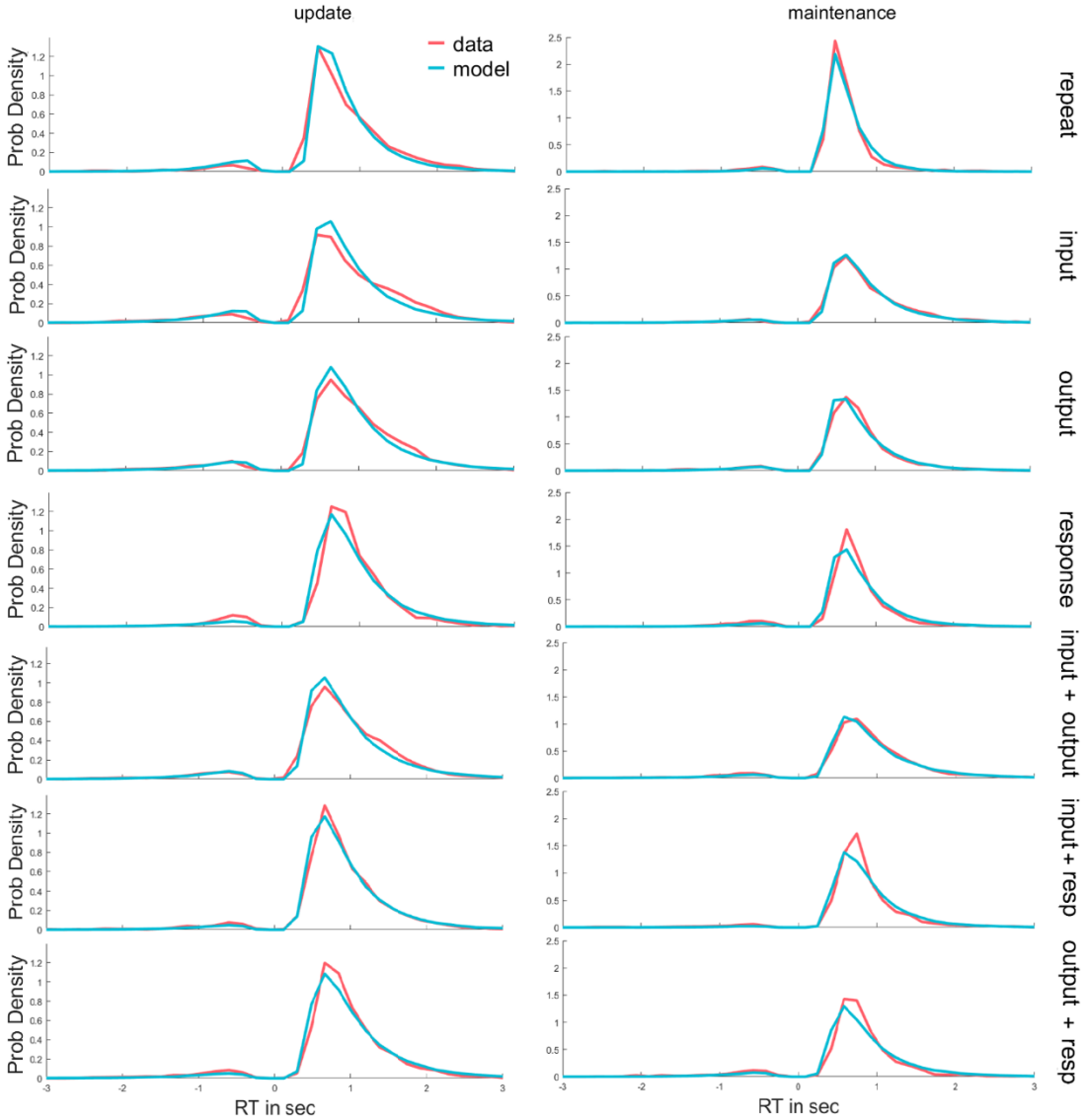

Figure 6. Model fit to behavior (a) Behavioral RT distributions across the group are shown for switching at each level (red line) together with posterior predictive simulation from the HDDM (light blue). Distributions of correct (the right positive tail) and incorrect (left negative tail) trials in updating trials (left) and maintenance trials (right) show good correspondence between data and model.

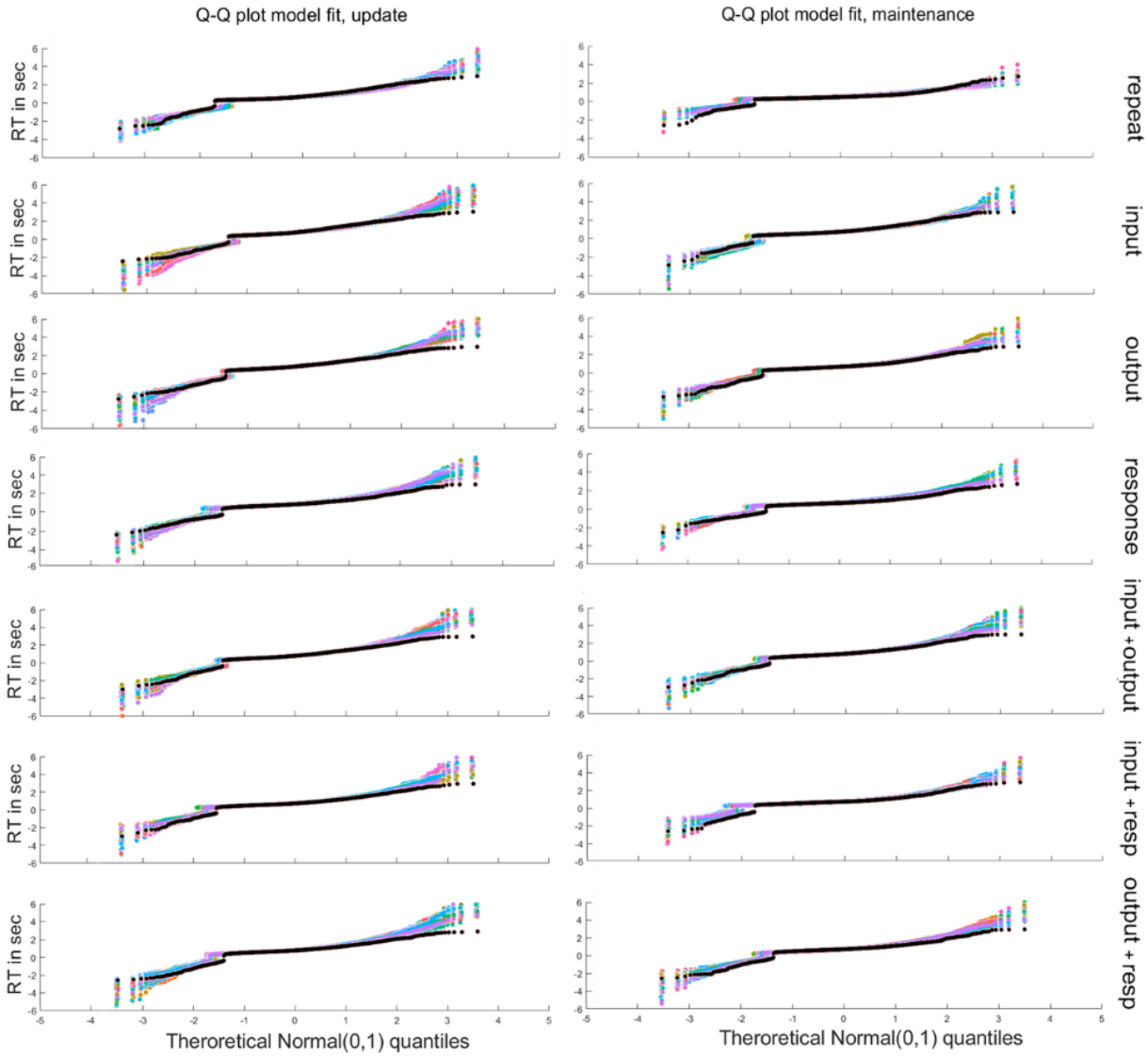

Figure 7. Model fit with Quantile-Quantile plots (a) Model fit to behavior can be more precisely viewed using quantile-quantile plots, showing quantiles of the empirical behavioral RT distributions (black) against the 50 simulation of RT distribution (colored lines, capturing model uncertainty) from the posterior predictive of the HDDM model, for correct response (positive RT) and incorrect responses (negative RT). Quantiles were computed at the group level.

Bayesian parameter estimation with HDDM revealed that the DDM provided an adequate fit to choice proportions and response time distributions (Fig 6). Moreover, switches at each independent level of gating (input, output, and response) were related to increased decision thresholds (see Fig 8 for statistics). Notably, in updating trials, such effects were incrementally larger when switches were closer to the response level. This result supports the notion that response conflict has preferential impact on motor decision threshold, but that switches in higher level gates can nevertheless recruit the same process. Relatedly, parameter estimates revealed under-additive interactions such that switches at higher gating levels had reduced impact if there were already switches at the lower level gates (Fig 8). These under-additive interactions mirrored those described in behavioral summary statistics above, and are also consistent with the EEG findings that gate switches overlapped in time. Maintenance trials showed relatively similar effects (Fig 8), with the exception that switches in input gating (i.e., closing the gate from updating to maintenance) were associated with increases in threshold that were comparable to those at the response level. We speculate about potential mechanisms for this effect in the discussion.

Notably, these threshold effects were obtained even though drift rate was also allowed to vary by condition. As noted, in the DDM, (lower) drift rates can also capture slowed RT when a condition is more difficult, as opposed to a control mechanism. Accordingly, we did observe that drift rates were slower in conditions that required more cognitive processing to succeed: updating vs. maintenance trials and switches at the cognitive level vs. switches at the response level. In both cases additional cognitive processing is required: updating trials require replacing WM content in addition to accessing previous content for responding, and switches at the cognitive levels are known to be more difficult than simple response switches (e.g., [54]). These

findings boost the interpretation that gate switches at any given level are specifically related to elevated decision thresholds, while accounting for impact of difficulty on drift rate.

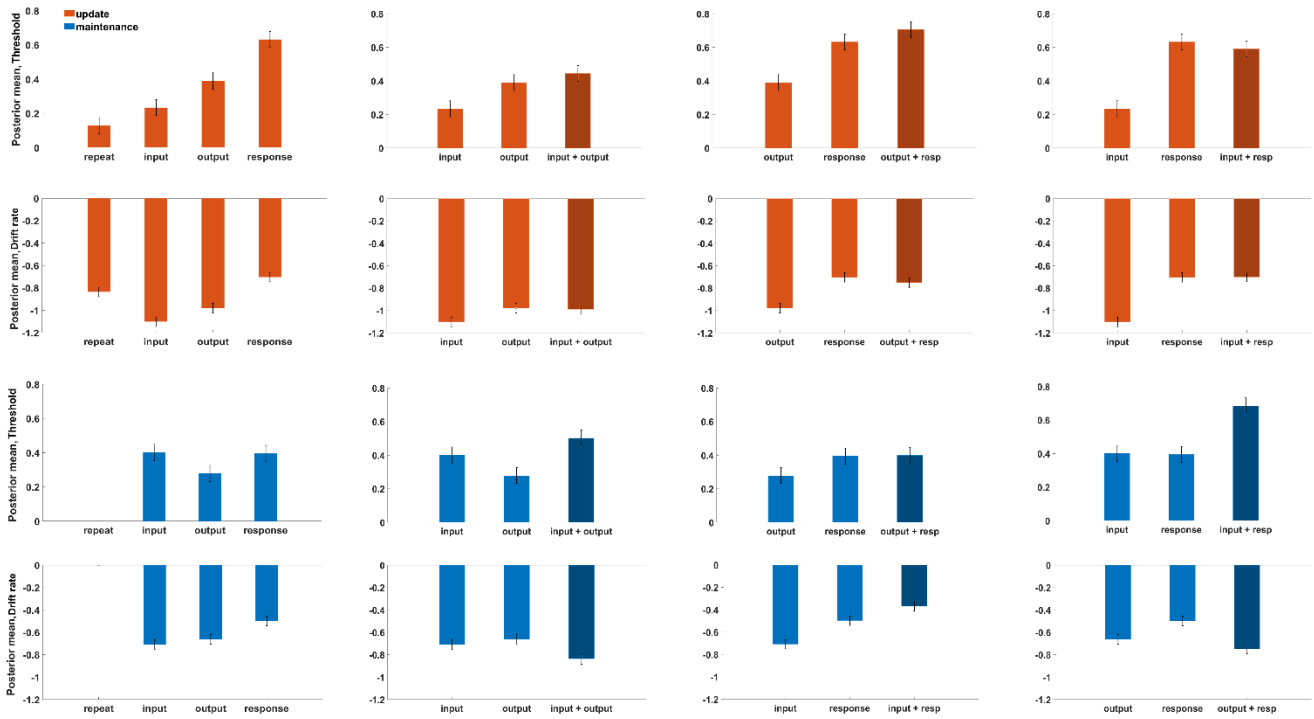

Figure 8. Under-additive interaction between gating selections. Decision threshold parameter estimates from HDDM in updating (first row, orange bars) and maintenance trials (third row, blue bars), exhibit main effects of gate switching at each independent level of gating (input, output, and response; left column), and under-additive interactions between gate switches in input-output, input-response, and output-response. Drift rate parameter estimations from the HDDM in updating (second row, orange bars) and maintenance trials (fourth row, blue bars) exhibit slowing in conditions that required more cognitive processing (left column) and exhibit facilitative interactions between gate switches. Dark color represents performance when the gates are switching together. Error bars reflect within-subject 95% highest density interval of the posterior distribution, in each case relative to maintenance repeat trials.

Discussion

These findings provide heretofore untested empirical support for the PBWM theoretical framework in which working memory involves a hierarchy of (content-addressable) selective input, output, and response gating operations. First, we developed a task that separately taxes the need for input and output gating processes, motivated directly by the computations in PBWM. The previous reference-back task focused on input gating [26, 27, 29] and its correlates in PFC and BG [30]. Here, we augmented that task by adding the need to track two independent categories over trials, thereby taxing selective input gating, while also orthogonally manipulating output gating. The advantage of the reference-back-2 task is that it is a continuous task that manipulates *selective* input and output gating demands to/from a particular address in memory, while preserving the need for continued maintenance of other information across trials.

Second, we confirmed the basic prediction that neural correlates of input, output, and response gating evolve dynamically and have analogous effects on behavior. The mass-univariate analysis of the EEG provided the temporal confirmation of this prediction by showing distinct spatiotemporal patterns for each gating level (input, output, response) that were independent and sequential, but partially overlapping, in time (see [51] for similar conclusions in PBWM simulations and behaviorally in a hierarchical rule task).

Third, we found that output gating (and particularly, switches therein) enhances the decodability of WM content selectively for the relevant category. The probabilistic and unpredictable switching between categories imposed a dynamic need to access distinct "addresses" within WM to update to, and read out from. This task property also had the important advantage of overcoming the challenge of measuring memory prioritization without being contaminated by the removal of irrelevant information [55] that often coincides with such

prioritization manipulations (e.g., [36]). Decoding analyses suggested that neural representations of relevant WM categories were enhanced following WM gate switches at the corresponding levels. Analogously, neural representations of motor responses were enhanced following response switches. We thus propose that gating operations, perhaps implemented by corticostriatal circuits, are partially responsible for previous observations of prioritization. Similar conclusions have been made in attentional shift experiments in which one or another stimulus category becomes relevant, and where striatal activity dictates whether PFC selectively enhances posterior representations of the relevant category [56]. Nevertheless, further investigations are needed to determine which neural mechanisms provide prioritization gain during gate switching. For example, gate switching may change the "active neural trace" (e.g., [43]) of representations or it may increase their accessibility for read-out (e.g., in other prefrontal layers as in PBWM). Other studies suggest that prioritization does not simply amplify relevant representations but that it also modifies the formatting of the non-selected representation so that they are less similar to the form that would be read out (e.g., [57-58]). Future research should also test whether gate switching facilitates coupling between cognitive rules and relevant task features like action representations [45, 59]

Finally, the above neural and behavioral findings were further supported by quantitative computational modeling using HDDM. We found that switches at any gating level were related to increases in the estimated decision threshold. This finding builds on previous studies showing that switches or response conflict gives rise to increased threshold via PFC-BG mechanisms [46, 48,49], but extends it to confirm the hierarchical PFC-BG model prediction that such effects hold for switches at more cognitive levels [12]. Nevertheless, such effects were still more dominant when gate switches were closer to the response level (they were strongest for response switch

then output switch then input switch). An exception to this conclusion was found whereby switches from updating to maintenance ("closing the gate") were accompanied by similarly large threshold adjustments as response switches. We speculate that closing the gate might also recruit the same prefrontal subthalamic pathway thought to be involved in transient response inhibition (e.g., [60-61]).

The cortico-striatal gating framework offers a theory of selective gating that includes not only plausible mechanisms for learning and generalization of gating policies [6, 7, 9, 18, 34] but also supports advanced cognitive control functions, like hierarchical control [12], and higher order learning and cognitive flexibility [11] in complex tasks like the reference-back-2 task used in this study. A key property of the cortico-striatal model that was further supported by current findings is that higher order cognitive "actions" (input and output gating operations) scaffold on top of the canonical computational motor gating operations. Indeed, striatal dysfunctions such as in Parkinson's disease causes not only progressive motor degeneration but also cognitive deficits that are both related to the dysfunction of the gating system [62]. More recent work also found that degraded ability to perform selective updating of WM is the key marker of the cognitive deficits in Parkinson's [63].

Furthermore, converging evidence from modeling [12, 64], and experimental work (e.g., [65-66]) demonstrated under-additive interactions when more than one order of switching was needed (e.g., switching task-sets and responses or switching context and items in WM). These results suggest that action selection is partially parallel, such that decisions about a response are processed to some degree even while the identity of the cognitive rule is uncertain. Such an interpretation is consistent with the parallel processing hypothesis across the gating system [51]. The HDDM analysis of the RT distribution of correct and error trials in the reference-back-2 task

further support the parallel processing account as under-additive interactions were observed between the gating loops. Specifically, the threshold parameter showed greater slowing when two gates were switching but with threshold predominated by the lower level gate. The drift-rate on the other hand, showed an under-additive facilitation where switching two gates was as fast as switching the lower level gate. The gain in processing speed for switching the higher level gate together with a lower level gate, recapitulates an effect often described in the task-switching literature (e.g., [64, 67]), whereby the task switch cost decreases or is abolished by response alterations. A possible interpretation for the facilitation finding is that WM updating decisions increase the mutually facilitative effect of switching across the gating system.

Finally, our study has various limitations. First and foremost, while the EEG method allowed us to assess the temporal dynamics of gating signals and representation decoding, it does not afford the ability to assess the involvement of corticostriatal circuits specifically. Functional imaging studies have implicated striatum in input gating in the reference back task [30], and output gating in other tasks [6], but testing the involvement of this circuit in selective content-addressable gating awaits further study.

Second, with the aim to increase our understanding of the latent cognitive processes that give rise to the full behavior in the reference-back-2 task and not just the mean RT and error rate, we leveraged the HDDM framework. However, it is likely that the decision process engaged during working memory gating diverges from that assumed by the standard DDM, and other models should be considered in future work (including those with time-varying drift rates and/or boundaries. Such models are difficult to estimate, but recent tools open the door for such investigations [68]. Moreover, future research could also employ a model-based approach where

neural patterns are linked to cognitive mechanisms through computational models (e.g., [47, 69]).

Methods

Participants

Thirty-two right-handed adults (aged 18–35; X female, X male) with normal or corrected-to-normal vision completed the experiment. All spoke English natively, were screened for neurological medications/conditions, and provided informed consent in accordance with the Research Protections Office at Brown University. Two participants were excluded from the analysis, due to technical problems with the experiment.

Stimuli and Procedure

Stimuli presentation and behavioral data collection were implemented using the Psychophysics Toolbox extensions in Matlab [70-71]. One out of four possible stimuli was presented on each trial. The stimuli were each of a distinct category: letters (“X”, “O”) and symbols (“%”, “#”) with two stimuli in each category. The stimuli appeared in a random order. Each trial started with a presentation of a stimulus inside a colored frame (red or blue), that indicated whether it was an updating trial or a maintenance trial. Participants had to make a matching decision between the presented stimulus and the last item category that was presented inside the updating color (e.g., red). Therefore, letters had to be compared with letters, and symbols with symbols. For all trials (regardless of blue or red) required participants to follow the same rule of comparing the stimulus with that of the corresponding category in the most recent updating trial. Participants were instructed about the meaning of the color (i.e., which one corresponded to reference to be updated and which one indicated the stimulus should only be compared to that in memory). The selected reference color (red or blue) was counterbalanced

between participants. The color of the frame in each trial was biased with 75% probability for one color in the first six blocks of the experiment and 75% probability for the other color for the last six blocks of the experiment. The order of color bias was chosen randomly. The biased color manipulation allowed us to have a more stable measure of selective output gating where the relevant category changed (output switch) but fewer trials in which there was also a color (input policy) switch. Participants were instructed that there will be a color bias that will flip once during the experiment. “Same” and “different” responses were indicated by using the right and left index fingers, respectively, to press ‘Z’ and ‘/’ on the keyboard. Response mappings were chosen randomly for each participant.

Stimulus presentation was limited to 3 sec. The response was followed by an inter-trial interval that was jittered between 800-1000 ms. Participants were instructed to keep their eyes fixated on the center of the screen throughout the experimental blocks. The first two trials in a block were always updating trials with stimuli from the two categories. Participants were instructed that accuracy will not be measured in the first two trials in the block. The experiment comprised of 12 blocks, including 90 trials each. Participants had to reach 80% accuracy on the practice block before they began the experiment. Participants were allowed to repeat the practice block up until 4 times.

Electroencephalogram (EEG) Recording and processing

Scalp voltage was measured using 62 Ag/AgCl electrodes referenced to a site immediately posterior to Cz using a Synamps2 system (bandpass filter 0.5–100 Hz, 500-Hz sampling rate). Preprocessing was conducted using the EEGLAB toolbox [72]. During pre-processing, data were low-pass filtered at 30 Hz and high-pass filtered at 0.1 Hz. Epochs were segmented from -200 to +800 ms surrounding stimulus onset and were baseline corrected from –

200 to 0 ms before the onset of the stimulus. The epoched data were visually inspected and those containing large artifacts due to facial electromyographic (EMG) activity or other artifacts (except for eye blinks) were manually removed. Independent components analysis (ICA) was next conducted using EEGLAB's runica algorithm. Components containing blink, oculomotor artifacts, or other artifacts that could be clearly distinguished from genuine neural activity signals, were subtracted from the data.

Data processing for univariate EEG analysis

To extract the neural correlates in the EEG signal of conditions of interest we employed a mass univariate approach. A multiple regression analysis was conducted for each participant, in which the EEG amplitude at each electrode site and time point was predicted by the conditions of interest while controlling for other factors such as RT (such an approach was recently used [52-53]). For the regression analysis, the EEG signal recorded with 500 Hz sampling rate was down-sampled by a factor of 4, resulting in 125 time points for the selected window of -200:800 ms around stimulus onset. The EEG signal was z-scored before it was entered to the robust multilinear regression analysis to account for remaining noise in the data [53].

Two separate multilinear regressions were run. The first regression assessed the neural correlates for gate switching. It included 7 regression factors: log of RT (to remove variability due to slower responses in some conditions; [53]), and 6 contrasts dummy coded as 1 and 0: WM state (updating vs maintenance), input gating (WM state switch vs repeat), output gating (category switch vs repeat), response gating (action switch vs repeat), bias (frequent updating vs rare updating), and finally one interaction between WM state × input gating (to search for a selective input gating pattern in updating trials).

The other regression assessed the neural prioritization of representations, i.e. to identify an EEG signal that differentiates between representations at each gating level. The second regression included 4 factors: log of RT and 3 contrasts dummy coded as 1 and 0: WM state (updating vs maintenance), category (symbols vs letters) and action (same vs different).

Statistical analysis of GLM weights

Statistics on the regression weights were performed across participants for all electrodes and time points by testing the significance of each point against 0. To correct for multiple comparisons, we performed cluster-mass correction by permutation testing [74] with custom written Matlab scripts. Cluster-based test statistics were calculated by taking the sum of the t-values within a cluster of significant points with threshold for a t test significance level of P = 0.001. This was repeated 1000 times, generating a distribution of maximum cluster-mass statistics under the null hypothesis. Only clusters with greater mass than the maximum cluster mass obtained with 95% chance permutations were considered significant [53]. The results of the second regression analysis yielded significant univariate patterns for the two representation types (category and action). Note that a different regression model was also run to assess the neural correlates of stimulus representations. This regression produced a significant univariate pattern only for the "O" representation but not for any of the other stimuli. Therefore, we did not continue with further analyses related to stimulus decodability.

Trial-by-trial decoding

Using the GLM masks we then computed the dot product between individual trials (voltage maps of electrode * time) and the identified masks (electrode * time maps of t-values of significant pixels; [53]). This computation produced a trial-level similarity measure (see Fig. 5a)

that presumably reflects how similar the EEG signal of the probed representation in memory on a given trial to the mask activity of this representation. To visualize the temporal dynamics of the EEG index (i.e. how the decoded representations evolves in time), the similarity measure procedure was calculated in sliding time bins of 40 ms across the epoch, and applied only to time bins in which the original mask was significant. The similarity index was calculated 121 times in each trial, with a 32ms overlap between time bins. The temporal similarity indexing was obtained by calculating the mean similarity between trials where one representation was probed and when the other one was probed.

Note that to test the effect of gate switching on prioritization, we excluded trials from the temporal dynamic analysis where the same stimulus was repeated from the previous trial. This exclusion allowed us to control for impact of perceptual switches that could have elicited an involuntary stimulus-driven enhancement of memory representations that match the probed category stimulus (e.g., [28,75]). Therefore, we restricted analysis to output repeat trials involving a perceptual change (e.g., from “X” to “O”), while output switch trials also included a category change (e.g., from “X” to “%”).

Drift diffusion modeling

The DDM is a common sequential sampling model of two-choice RT tasks. The advantage of this model is that it can translate response time distributions and error rates to the underlying generative parameters in each task condition, and it has previously been used to summarize decision dynamics that arise from BG gating [33, 47-49]. The core parameters in the model are threshold (or boundary separation) a, drift rate v, and non-decision time t. Threshold is the distance between the response boundaries, where higher threshold indicates that more evidence needs to be accumulated before committing to a choice, leading to slower but more

accurate responses. Conflict or switching in motor responses leads to elevated decision thresholds [33, 47-49]. Drift rate is the rate in which evidence is accumulated. Larger drift rates are usually interpreted to reflect higher quality of evidence that is expressed by faster and more accurate response times. Finally, non-decision time captures the processes that are not related to the decision making process such as stimulus encoding and motor execution.

We tested whether the DDM provides a good model of RT distributions of correct and incorrect choices during the reference-back-2 task in which the decision threshold can be adjusted as a function of conflict or switches at any gate level (while controlling for any differences in drift rate). The underlying decision-making process during the reference-back-2 task were estimated from the DDM likelihood functions [33]. We used hierarchical Bayesian estimation of DDM parameters, where individual's fit is constrained and informed by the group distribution leading to more accurate estimation of parameters at both the individual and group level [33].

Parameter estimation in the Hierarchical Bayesian framework used the Markov-chain Monte-Carlo (MCMC). HDDM is especially beneficial for estimating individual parameters while optimizing the tradeoff between random and fixed-effects [33]. The first two trials in each block, omission trials, and trials with very fast RT (<200ms) were excluded from the analysis. RT was limited to 3sec. There were three thousand samples generated from the posterior using four chains. The first thousand (burn-in) and every second (thinning) were discarded. Proper chain convergence was tested between the MCMC chains, using the $\hat{R}$ statistic [76], which measures the degree of variation between chains relative to the variation within chains. The maximum $\hat{R}$ value across all parameters in all eight models was 1.03, indicating that all chains converged successfully [77].

Statistical analysis was performed on the group mean posteriors. The Deviance Information Criterion (DIC) was used for model comparison [78] which balances model fit against complexity. We first tested the simplest model where the five conditions (WM state, input switching, output switching, response switching, and updating frequency) were used as regression weights without interactions and with group level estimates (DIC: 29339). We systematically added interactions, and the model with the best fit was one in which drift rate and threshold exhibited 4-way interactions (DIC 28806). This best fitting model was estimated again allowing for subject-level estimates in each condition (DIC 27232):

Threshold ~ *a_mr* + Update_frequency + (WM state × input × output × response) +…

Drift rate ~ *v_mr* + Update_frequency + (WM state × input × output × response).

A_mr is the intercept for threshold and *v_mr* is the intercept for drift-rate, Update_frequency reflects the high or low frequency of updating, and the 4-way interaction captured all combinations of switch or repeat during updating and maintenance trials (15 combinations in total). Within subject regressions were used such that all gate switch effects are evaluated relative to the maintenance repeat condition (the intercept), using patsy in HDDM. Significance was determined if the 95% confidence interval of the posterior mean did not overlap with 0.

Behavior characterization and model validation

We first plotted the overall RT distributions for each gating condition and interaction, separately for updating and for maintenance trials (Fig. 6) using Gramm plotting toolbox [79]. This plot showed that the HDDM model captures choice proportions (of correct and incorrect trials) and the different shape of the RT distributions within each gating condition,

simultaneously. A critical test of a the DDM model is that it can capture the full RT distributions in each condition. Although we used model fit statistics (DIC) to select the best-fitting model, it is also important to validate that the best fitting model can capture the observed response proportions and RT distributions. For this model validation we generated quantile- quantile plots describing the correspondence between behavior and DDM predictions (Fig. 7) [79]. We compared the shape of the empirical RT distribution to the shape of the simulated RT distribution by plotting each data against a theoretical normal distribution. The plot in Fig. 7 exhibits the RTs on the y-axis as a function of the estimated quantiles on the x-axis based on the inverse of the continuous cumulative distribution function (quantile function) that was estimated at the group level. RTs of the model were simulated 50 times from the posterior predictive of the HDDM model and plotted on top of each other to show the uncertainty in the model. As can be seen, the empirical RT was mostly within the range of the simulated RT with a small over-estimation at the right tail of the distribution.

References:

1. Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, **24**, 167-202.
2. Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in Cognitive Sciences, **16***, 106-113.
3. Dreisbach, G., & Fröber, K. (2019). On how to be flexible (or not): Modulation of the stability-flexibility balance. *Current Directions in Psychological Science,* **28**, 3-9.
4. Goschke, T., & Bolte, A. (2014). Emotional modulation of control dilemmas: The role of positive affect, reward, and dopamine in cognitive stability and flexibility. *Neuropsychologia,* **62**, 403-423.
5. Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. *Control of cognitive processes: Attention and performance XVIII*, 713-737.
6. Chatham, C. H., Frank, M. J., & Badre, D. (2014). Corticostriatal output gating during selection from working memory. *Neuron,* **81**, 930-942.
7. Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cognitive Affective Behavior Neuroscience,* **1**, 137-60.
8. Frank, M.J. and O'Reilly, R.C. (2006) A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. *Behavior Neuroscience,* **120** , 497-517.
9. Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2006). Banishing the homunculus: making working memory work. *Neuroscience,* **139**, 105-118.

10. O'Reilly, R.C. & Frank, M.J. (2006). Making working memory work: A computational model of learning in the frontal cortex and basal ganglia. *Neural Computation*, **18**, 283-328.
11. Frank, M.J. & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cerebral Cortex*, **22**, 509-526.
12. Collins, A.G.E. & Frank, M.J. (2013). Cognitive control over learning: Creating, clustering and generalizing task-set structure. *Psychological Review*, **120**, 190-229.
13. Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. Philosophical Transactions of the Royal Society B: *Biological Sciences*, **362**, 1585-1599.
14. Frank, M. J. (2011). Computational models of motivated action selection in corticostriatal circuits. *Current Opinion in Neurobiology*, **21**, 381-386.
15. Chatham, C. H., & Badre, D. (2015). Multiple gates on working memory. *Current Opinion in Behavioral Sciences*, **1**, 23-31.
16. Cools, R. (2019). Chemistry of the Adaptive Mind: Lessons from Dopamine. *Neuron*, **104**, 113-131.
17. Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in cognitive sciences*, **12**, 193-200.
18. Badre, D., & Frank, M. J. (2012). Mechanisms of hierarchical reinforcement learning in cortico–striatal circuits 2: Evidence from fMRI. *Cerebral cortex*, **22**, 527-536.
19. Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron*, **66**, 315-326.

20. Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, **302**, 1181-1185.

21. Verstynen, T. D., Badre, D., Jarbo, K., & Schneider, W. (2012). Microstructural organizational patterns in the human corticostriatal system. *Journal of Neurophysiology*, **107**, 2984-2995.

22. Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, **18**, 394-412.

23. Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the N-back task: a question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **33**, 615.

24. Chatham, C. H., Herd, S. A., Brant, A. M., Hazy, T. E., Miyake, A., O'Reilly, R., & Friedman, N. P. (2011). From an executive network to executive control: A computational model of the n-back task. *Journal of Cognitive Neuroscience*, **23**, 3598-3619.

25. Rac-Lubashevsky, R., & Kessler, Y. (2016). Dissociating working memory updating and automatic updating: The reference-back paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **42**, 951.

26. Rac-Lubashevsky, R., & Kessler, Y. (2016). Decomposing the n-back task: An individual differences study using the reference-back paradigm. *Neuropsychologia*, ***90***, 190-199.

27. Rac-Lubashevsky, R., & Kessler, Y. (2018). Oscillatory correlates of control over working memory gating and updating: An EEG study using the reference-back paradigm. *Journal of Cognitive Neuroscience*, **30**, 1870-1882.

28. Rac-Lubashevsky, R., & Kessler, Y. (2019). Revisiting the relationship between the P3b and working memory updating. *Biological Psychology*, **148**, 107769.

29. Rac-Lubashevsky, R., Slagter, H. A., & Kessler, Y. (2017). Tracking real-time changes in working memory updating and gating with the event-based eye-blink rate. *Scientific Reports*, **7**, 1-9.

30. Nir-Cohen, G., Kessler, Y., & Egner, T. (in press). Neural substrates of working memory updating. *Journal of Cognitive Neuroscience*.

31. Kriete, T., & Noelle, D.C. (2011) Generalization benefits of output gating in a model of prefrontal cortex. *Connection Science*, **23**, 119-129.

32. Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*, **110**, 16390-16395.

33. Wiecki, T.V. & Frank, M.J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review*, **120**, 329-55.

34. Collins, A. G., Cavanagh, J. F., & Frank, M. J. (2014). Human EEG uncovers latent generalizable rule structure during learning. *Journal of Neuroscience*, **34**, 4677-4685.

35. Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing information during working memory: beyond sustained internal attention. *Trends in Cognitive Sciences*, **21**, 449-461.

36. Oberauer, K. (2019). Working memory and attention–A conceptual analysis and review. *Journal of Cognition*, **2**, 1-23.

37. Olivers, C. N., Peters, J., Houtkamp, R., & Roelfsema, P. R. (2011). Different states in visual working memory: When it guides attention and when it does not. *Trends in Cognitive Sciences,* **15**, 327-334.

38. D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology*, **66**, 115-142.

39. LaRocque, J. J., Lewis-Peacock, J. A., & Postle, B. R. (2014). Multiple neural states of representation in short-term memory? It's a matter of attention. *Frontiers in Human Neuroscience,* **8**, 1-14.

40. Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends in Cognitive Sciences,* **19,** 394-405.

41. Günseli, E., Fahrenfort, J. J., van Moorselaar, D., Daoultzis, K. C., Meeter, M., & Olivers, C. N. (2019). EEG dynamics reveal a dissociation between storage and selective attention within working memory. *Scientific Reports,* **9**, 1-13.

42. LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2013). Decoding attended information in short-term memory: an EEG study. *Journal of Cognitive Neuroscience,* **25**, 127-142.

43. LaRocque, J. J., Riggall, A. C., Emrich, S. M., & Postle, B. R. (2017). Within-category decoding of information in different attentional states in short-term memory. *Cerebral Cortex,* **27**, 4881-4890.

44. Lewis-Peacock, J. A., & Postle, B. R. (2012). Decoding the internal focus of attention. *Neuropsychologia,* **50**, 470-478.

45. Olivers, C. N., & Roelfsema, P. R. (2020). Attention for action in visual working memory. *Cortex,***131**, 179-194.

46. Ratcliff, R., & Frank, M. J. (2012). Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. *Neural Computation*, **24**, 1186-1229.

47. Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience*, **14**, 1462-1467.

48. Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., & Badre, D. (2015). fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *Journal of Neuroscience*, **35**, 485-494.

49. Herz, D. M., Zavala, B. A., Bogacz, R., & Brown, P. (2016). Neural correlates of decision thresholds in the human subthalamic nucleus. *Current Biology*, **26**, 916-920.

50. Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, **20**, 873-922.

51. Ranti, C., Chatham, C. H., & Badre, D. (2015). Parallel temporal dynamics in hierarchical cognitive control. *Cognition*, **142**, 205-229.

52. Collins, A. G. E., & Frank, M. J. (2016). Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. *Cognition*, **152**, 160-169.

53. Collins, A. G., & Frank, M. J. (2018). Within-and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proceedings of the National Academy of Sciences*, **15**, 2502-2507.

54. Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. *Psychological Bulletin*, **136**, 849-874.

55. Lewis-Peacock, J. A., Kessler, Y., & Oberauer, K. (2018). The removal of information from working memory. *Annals of the New York Academy of Sciences*, **1424**, 33-44.

56. van Schouwenburg, M. R., den Ouden, H. E., & Cools, R. (2010). The human basal ganglia modulate frontal-posterior connectivity during attention shifting. *Journal of Neuroscience*, **30**, 9910-9918.

57. van Loon, A. M., Olmos-Solis, K., Fahrenfort, J. J., & Olivers, C. N. (2018). Current and future goals are represented in opposite patterns in object-selective cortex. *ELife*, **7**, e38677.

58. Yu, Q., Teng, C., & Postle, B. R. (2020). Different states of priority recruit different neural representations in visual working memory. *PLoS biology*, **18**, e3000769.

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

60. Aron, A. R., Durston, S., Eagle, D. M., Logan, G. D., Stinear, C. M., & Stuphorn, V. (2007). Converging evidence for a fronto-basal-ganglia network for inhibitory control of action and cognition. *Journal of Neuroscience*, **27**, 11860-11864.

61. Jahfari, S., Ridderinkhof, K. R., Collins, A. G., Knapen, T., Waldorp, L. J., & Frank, M. J. (2019). Cross-task contributions of frontobasal ganglia circuitry in response inhibition and conflict-induced slowing. *Cerebral Cortex*, **29**, 1969-1983.

62. Wiecki, T. V., & Frank, M. J. (2010). Neurocomputational models of motor and cognitive deficits in Parkinson's disease. In *Progress in Brain Research* (Vol. **183**, pp. 275-297). Elsevier.

63. Salmi, J., Ritakallio, L., Fellman, D., Ellfolk, U., Rinne, J. O., & Laine, M. (2020). Disentangling the role of working memory in Parkinson's disease. *Frontiers in Aging Neuroscience*, **12**, 1-17.

64. Brown, J. W., Reynolds, J. R., & Braver, T. S. (2007). A computational model of fractionated conflict-control mechanisms in task-switching. *Cognitive Psychology*, ***55***, 37-85.

65. Risse, S., & Oberauer, K. (2010). Selection of objects and tasks in working memory. *Quarterly Journal of Experimental Psychology*, **63**, 784-804.

66. Sali, A. W., Jiang, J., & Egner, T. (2020). Neural Mechanisms of Strategic Adaptation in Attentional Flexibility. *Journal of Cognitive Neuroscience*, **32**, 989-1008.

67. Korb, F. M., Jiang, J., King, J. A., & Egner, T. (2017). Hierarchically organized medial frontal cortex-basal ganglia loops selectively control task-and response-selection. *Journal of Neuroscience*, **37**, 7893-7905.

68. Fengler, A., Frank, M., Govindarajan, L., & Chen, T. (2020). Likelihood Approximation Networks (LANs) for Fast Inference of Simulation Models in Cognitive Neuroscience. bioRxiv.

69. Turner, B. M., Palestro, J. J., Miletić, S., & Forstmann, B. U. (2019). Advances in techniques for imposing reciprocity in brain-behavior relations. *Neuroscience & Biobehavioral Reviews*, **102**, 327-336.

70. Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, **10**, 433-436.

71. Pelli, D. G. (1997). The Video Toolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, **10**, 437-442.

72. Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, **134**, 9-21.

73. Cooper, P. S., Karayanidis, F., McKewen, M., McLellan-Hall, S., Wong, A. S., Skippen, P., & Cavanagh, J. F. (2019). Frontal theta predicts specific cognitive control-induced behavioural changes beyond general reaction time slowing. *Neuroimage*, **189**, 130-140.

74. Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, **15**, 1-25.

75. van Ede, F., Board, A. G., & Nobre, A. C. (2020). Goal-directed and stimulus-driven selection of internal representations. *Proceedings of the National Academy of Sciences*, **117**, 24590-24598.

76. Gelman, A. (1996). Inference and monitoring convergence. Markov chain Monte Carlo in practice, 131-143.

77. Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457-472.

78. Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series b (statistical methodology)*, **64**, 583-639.

79. Morel, (2018). Gramm: grammar of graphics plotting in Matlab. *Journal of Open Source Software*, **3**, 568, https://doi.org/10.21105/joss.00568.

Acknowledgements

This research was funded by the National Institutes of Health under award number R01 MH084840-08A1. Work by the first author was partially supported by THE ISAREL SCIENCE FOUNDATION (grant No. 96/19) and by Fulbright United states - Israel Educational Foundation.