HHS Public Access

Author manuscript

J Cogn Neurosci. Author manuscript; available in PMC 2020 August 01.

Published in final edited form as:

J Cogn Neurosci. 2020 May ; 32(5): 989–1008. doi:10.1162/jocn_a_01541.

Neural mechanisms of strategic adaptation in attentional flexibility

Anthony W. Sali[1,2,3], **Jiefeng Jiang**[4], **Tobias Egner**[2,3]

[1]Department of Psychology, Wake Forest University, 27109

[2]Center for Cognitive Neuroscience, Duke University, 27708

[3]Department of Psychology and Neuroscience, Duke University, 27708

[4]Department of Psychology, University of Iowa, 52242

Abstract

Individuals are able to adjust their readiness to shift spatial attention, referred to as attentional flexibility, according to the changing demands of the environment, but the neural mechanisms underlying learned adjustments in flexibility are unknown. In the current study, we used functional magnetic resonance imaging (fMRI) to identify the brain structures responsible for learning shift-likelihood. Participants were cued to covertly hold or shift attention among continuous streams of alphanumeric characters and to indicate the parity of target stimuli. Unbeknownst to the participants, the stream locations were predictive of the likelihood of having to shift (or hold) attention. Participants adapted their attentional flexibility according to contextual demands, such that the response time cost associated with shifting attention was smallest when shift cues were most likely. Learning model-derived shift prediction error scaled positively with activity within dorsal and ventral fronto-parietal regions, documenting that these regions track, and update, shift likelihood. A complementary inverted encoding model analysis revealed that the pretrial difference in attentional selection strength between to-be-attended and to-be-ignored locations did not change with increasing shift likelihood. The behavioral improvement associated with learned flexibility may primarily arise from a speeding of the shift process rather than from preparatory broadening of attentional selection.

Keywords

Cognitive Flexibility; Inverted Encoding Model; Reinforcement Learning; Visual Attention

Attentional prioritization of a stimulus has traditionally been conceptualized as a product of low-level stimulus properties (Theeuwes, 1992, 1994; Yantis & Jonides, 1984), as well as top-down behavioral goals (Desimone & Duncan, 1995; Folk, Leber, & Egeth, 2002; Folk, Remington, & Johnston, 1992; Wolfe, Cave, & Franzel, 1989). However, recent work has suggested that our previous experiences also powerfully shape the likelihood that a stimulus will receive attentional selection (Anderson, Laurent, & Yantis, 2011; Awh, Belopolsky, &

Please direct correspondence to: Anthony W. Sali, Department of Psychology, Wake Forest University, P.O. Box 7778 Reynolda Station, Winston-Salem, NC 27109, saliaw@wfu.edu, Phone: 336-758-4813.

Theeuwes, 2012; Sali, Anderson, & Courtney, 2016; Sali, Anderson, & Yantis, 2014). Accordingly, in addition to spontaneous changes in individuals' readiness to shift spatial attention (Sali, Courtney, & Yantis, 2016), referred to here as *attentional flexibility*, individuals are able to adjust their shift-readiness according to the statistical regularities of their environment (Sali, Anderson, & Yantis, 2015). In particular, attention shift costs, the slowed performance in trials that require shifting attention versus holding attention, are smaller in task contexts that have been associated with a high likelihood of shifting attention in the past than in those that have been associated with holding attention. Moreover, neuropsychiatric disorders are associated with both elevated (e.g. attention deficit hyperactivity disorder and substance abuse; Barkley, 1997; Berridge, 2012; Keiflin & Janak, 2015; Vaurio, Simmonds, & Mostofsky, 2009) and deficient (e.g. schizophrenia and autism; Koster-Hale & Saxe, 2013; Moore, Dickinson, & Fletcher, 2011; Murray, Corlett, & Fletcher, 2010) levels of attentional flexibility. Thus, the question of how individuals learn to adjust attentional flexibility according to moment-by-moment changes in environmental demands is of clear public health relevance. In particular, the neural mechanisms underlying such learned states of attentional flexibility are presently unknown.

In the current study, we investigated these mechanisms by combining functional magnetic resonance imaging (fMRI) with a well-validated rapid serial visual presentation (RSVP) paradigm of spatial attentional selection in which task contexts differentially predicted the likelihood of an upcoming attention shift (e.g. Sali et al., 2015). Specifically, we addressed three key questions: First, we sought to characterize brain regions involved in the acquisition of shift-likelihood predictions. In line with computational models and empirical studies of how participants exploit statistical task regularities in guiding cognitive control processes (e.g., Botvinick et al., 2001; Blais, Robidoux, Risko, & Besner, 2007; Verguts & Notebeart, 2007; Abrahamse et al., 2016; Chiu, Jiang, & Egner, 2017; Jiang, Beck, Heller, & Egner, 2015; Jiang, Wagner, & Egner, 2018; Waskom, Frank, & Wagner, 2017), we assumed that the contextual adaptation of attentional flexibility is based on cumulative learning of the current probabilistic associations between context (temporal and/or spatial) and shift-likelihood. Accordingly, we used reinforcement learning (RL) models (Sutton & Barto, 1998) to fit behavioral response times and compute the degree to which an individual's shift predictions differed from the outcome (shift versus hold) of each trial, that is, the shift prediction error (*shift PE*), which serves to update shift-likelihood predictions. By employing trial-by-trial shift PE estimates in model-driven fMRI analyses (Chiu, Jiang, & Egner, 2017; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Jiang, Beck, Heller, & Egner, 2015; Jiang, Wagner, & Egner, 2018; O'Doherty, Cockburn, & Pauli, 2017; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003), we could identify brain regions that drive attentional flexibility learning.

Second, in addition to probing the neural mechanisms involved in the acquisition of attentional flexibility, we sought to determine the consequences of that learning on the brain's attentional prioritization of stimulus locations. To this end, we used an inverted encoding model (IEM) of spatial representation in visual cortex (Sprague, Ester, & Serences, 2014; Sprague, Saproo, & Serences, 2015; Sprague & Serences, 2013) to examine whether attentional orienting expectations, as defined by the RL model, modulated the strength of attentional selection at currently attended and unattended spatial locations, both preceding

and following a cued shift or hold of spatial attention. On the one hand, reduced shift costs for contexts with a high shift-likelihood may result from a preparatory increase in spatial selection at a to-be-attended location and/or decrease in selection at the currently attended location prior to the onset of a shift cue. Alternatively, learned states of attentional flexibility might not be associated with changes in preparatory attentional selection, but instead, with a speeding of the shift process itself. To investigate these questions, we examined whether trial-by-trial RL model-derived shift predictions covaried with the deployment of spatial attentional selection, as measured via the IEM.

Third, we used an IEM analysis of the post-cue brain activity to test whether participants shifted attention according to the orienting cues rather than zooming attention out to include both target and distractor locations (Castiello & Umiltà, 1990; Jefferies, Gmeindl, & Yantis, 2014). If participants shifted attention in response to cues, we would expect selection to be strongest at the location of the cue for hold trials, but strongest at the opposite location from the cue on shift trials. Conversely, if participants were able to attend to both stimulus locations simultaneously, we would expect to see equal selection at the cue and non-cue locations following the cue presentation. Furthermore, any difference in post-cue selection according to the probability context would shed light on the consequences of orienting expectations on the resulting strength of attentional selection in visual cortex.

Methods

Participants

Thirty-one individuals (14 male, 17 female) ranging in age from 21 to 38 years (M=28.6, SD=5.19) participated in exchange for monetary compensation. Of these participants, 2 were excluded for overall behavioral accuracies less than 75% on the RSVP task, 1 was excluded for a combination of excessive motion and falling asleep, and 5 were excluded for technical difficulties, resulting in a final sample size of 23 participants. One of the eligible participants completed a slightly different variation of the encoding model training task and is therefore excluded from the encoding model analyses alone. All participants completed a consent form that was approved by the Duke University Institutional Review Board. Compensation was $40 and the entire study lasted approximately 2 hours.

Design and Procedure

Attentional Flexibility Paradigm.—Participants completed a variant of a previously-established paradigm for tracking learning of attentional flexibility (Sali et al., 2015; Sali, Courtney, et al., 2016). On each trial, participants fixated a centrally-presented dot while covertly attending to one of two rapid serial visual presentation (RSVP) streams of alphanumeric characters (approximately 1.4° x 1.5°) that were positioned to the left and right of fixation (see Figure 1A; 3° distance to fixation measured center-to-center; diameter of fixation dot = 0.2°). At the beginning of the trial, a flashing asterisk appeared at the to-be-attended location for a total of 500 ms. Next, the two RSVP streams appeared and changed with a frame rate of 250 ms. During a distractor interval that was either 4 or 6 seconds (presented with equal frequency), each RSVP stream within each frame contained a single randomly-selected digit that ranged from 1 to 8. Following the distractor interval, a letter cue

appeared in the to-be-attended stream that signaled participants to either covertly shift attention to the opposite stream or to continue holding attention at the location of the cue. The letter “A” cued participants to hold attention at the location of the cue, while the letter “K” cued participants to covertly shift attention. The stimuli were presented within a circular grey aperture with a total diameter of approximately 16.7° of visual angle. Stimulus presentation in this and subsequent tasks was controlled with the Psychophysics Toolbox in Matlab (Brainard, 1997) and stimuli were viewed on a screen located in the bore of the scanner through a mirror that was attached to the head coil.

In order to measure the flexibility with which individuals shifted attention on a trial-by-trial basis, participants made a speeded parity judgment for the string of digits appearing at the cued location. Critically, as in earlier studies (Sali et al., 2015; Sali, Courtney, et al., 2016), all digits appearing at the cued location were either even or odd for a 2-second response window that began immediately following the presentation of the cue. Participants were asked to make the parity judgment based on the first digit that they detected. Specifically, they pressed a button with their left index finger if the stimuli were odd and a button with their right index finger if the stimuli were even. Participants only responded once during the response window and the digits at the unattended location continued to be randomly generated throughout the response window regardless of when the press occurred. Following the response window, there was a blank inter-trial interval of 1250 ms. Participants completed a total of four runs of the attentional flexibility paradigm and each run consisted of 60 trials in total.

In order to manipulate participants’ shift readiness, we varied the likelihood that participants would receive a cue to shift or hold attention across three location-defined contexts. Stimuli were always presented in one of the three pairs of spatial locations: along the horizontal meridian (x = −3° y = 0°; x = 3° y = 0°), diagonal with the left item above the horizontal meridian (x = −1.5° y = 2.598°; x = 1.5° y = −2.598°), and diagonal with the right item above the horizontal meridian (x =−1.5° y = −2.598°; x = 1.5° y = 2.598°; see Figure 1B). Importantly, for each context, the two stream locations were equidistant from the fixation point (approximately 3° of visual angle) such that the spatial representation of stimuli could later be realigned with a simple rotation procedure (see encoding model method below). Each of these location contexts (pair of two spatial locations) was associated with a different likelihood that the participant would receive a cue to shift attention. Specifically, one context was associated with 25% shift trials and 75% hold trials, while another was associated with 75% shift trials and 25% hold trials. The final context was associated with an equal number of shift and hold trials. Shift and hold trials were randomized within each context with the constraint that there could never be a repeat of the “rare” trial types within the 25% shift and 75% shift contexts. The probability associations of each of the three contexts were counterbalanced across participants and participants were not informed that stimulus location would predict the likelihood of shifting attention. Each context lasted for 20 trials such that all three contexts were presented once per experimental run in a blocked fashion. The order of contexts within each run was randomly selected.

After completing the experiment, participants answered a series of questions with increasing specificity to determine the degree to which they had explicit knowledge of the underlying

probabilistic structure of the task. Participants first read and answered the following question: “Was there any relationship between the three stimulus location contexts (–, \, or /) and the attention cues (K and A)?” They then read, “One location context was associated with *Mostly Hold Cues*, one was associated with *Mostly Shift Cues*, and one was associated with *Equal Shift and Hold Cues*,” and labeled diagrams of the three stimulus locations with labels of “Mostly Hold,” “Mostly Shift,” and “Equal.”

Model Training Task.—In addition to the attention task described above, participants completed 4 runs of a visual target detection task that was used to independently train an encoding model of spatial representation (Sprague, Ester, & Serences, 2016). Participants alternated between the attention and model training tasks across consecutive runs of the experiment, beginning with the attention task. During the model training task, participants viewed flashing checkerboard stimuli that appeared one at a time at locations throughout a central portion of the visual field while fixating a central point with a diameter of approximately 0.2 degrees. Each checkerboard subtended approximately 1.8° of visual angle and was presented within a circular grey aperture with a total diameter of approximately 16.7° of visual angle. During the task, each checkerboard was presented for a total of 3000 ms and flickered at a rate of 6 hz (see Figure 1C). Following each checkerboard presentation, the aperture and fixation point alone were presented for a variable inter-trial-interval (ITI), which ranged from 2000–6000 ms. These ITIs were randomly selected, without replacement, from a set of intervals spanning 2000–6000 ms in approximately 87 ms segments.

We varied the checkerboard locations across trials and across experimental runs. These locations were originally selected from a hexagonal grid of 37 potential locations that was centered in the grey aperture such that the distance between each location was 1.5°. Each trial’s checkerboard distribution was further jittered according to two different procedures. First the hexagonal grid was rotated 0°, 15°, 30°, or −15° for the four runs of the task. Next, each location was further jittered by randomly drawing coordinates from within a circle with a radius of 0.5° that was centered at the rotated grid location. At its widest point, the grid spanned 3 steps (4.5°). An aperture (0.8° in diameter) surrounded the fixation point on trials in which the checkerboard stimulus would overlap with the fixation dot. Given the spatial jittering and rotation described above, the maximum stimulated region of the visual field was 5.9° from fixation, making the total field of view 11.8° × 11.8°. Each of the 37 grid locations was used for one checkerboard stimulus per run.

In addition to the 37 trials described above, on an additional 10 trials, a checkerboard appeared at one of the locations (determined by the same procedure as above), but slightly dimmed during the 3000 ms presentation. Participants were instructed to press a button with their right index finger as soon as they detected the stimulus dimming.

Fixation Training.—Immediately prior to entering the fMRI scanner, participants completed a 30-minute training session. In addition to receiving instructions and completing practice trials of both tasks, participants also completed a separate task that was designed to train them to fixate their eyes on a central target (Guzman-Martinez, Leung, Franconeri, Grabowecky, & Suzuki, 2009). In particular, use of this visual display, which produced real-

time visual feedback if participants made an eye movement, has been associated with improvements in participants' fixation performance in subsequent tasks.

Behavioral Reinforcement Learning Model.—In order to determine the neural substrates of adapting shift-readiness to the statistical structure of the task, we considered a series of six learning models with different assumptions of how the learning of shift-readiness varied at the trial- and context-level. To compute a trial-by-trial measure of shift PE during the attentional flexibility task, we employed a reinforcement learning (RL) model (Sutton & Barto, 1998; see also Chiu et al., 2017; O'Doherty et al., 2017). In addition to modeling trial-level adjustments, we also considered block-by-block adaptations of shift-readiness, which were modeled using a separate free parameter. The performance of each model in explaining trial-level variance in response time was measured using cross-validation (see below). Model 1 used temporal integration that was independent of the location context, while model 2 added an extra free parameter to account for participants' tendencies to partially reset shift predictions when there was a change of context or run, and model 3 accounted for context-dependent learning. Model 4 included free parameters for both context-independent and context-dependent learning, while model 5 added the reset factor from model 2. Finally, model 6 was identical to model 5, with the exception of the reset parameter being restricted to either the absence of resetting or a complete reset of shift predictions with each change of context or start of a new run (see Table 1 for an overview of models. Each model is also explained in greater detail below).

Model 1 captured learning through temporal integration that was independent from location context. Model 1 thus accounted for trial-by-trial learning that spanned across the location contexts. Specifically, learning was modeled as: $P_{i+1} = P_i + \alpha_t(S_i - P_i)$, where for each trial i, P denoted shift-readiness on a linear scale from 0 (least ready to shift) to 1 (most ready to shift). S_i denoted the shift outcome on each trial (1=shift attention, 0=hold attention). Importantly, $(S_i - P_i)$ is a measure of the degree to which each shift outcome differs from the participant's expectation at that particular moment, i.e., the shift PE, which is employed to update predictions accordingly for the next trial. The absolute (unsigned) value of shift PE thus represents a generalized learning signal that indicates the degree to which expectations were violated. Finally, how much shift PE influenced the new prediction P_{i+1} was mediated by the learning rate, a free parameter denoted as α_t, which varied from 0.01 to 0.99. The model always began with a shift-readiness value of 0.5 for the first trial of the first run, reflecting a neutral belief of shift-readiness. The first trial of all subsequent runs began with the shift-readiness value dictated by the last trial of the previous run.

Model 2 was identical to Model 1, except that it allowed for the possibility that participants' shift predictions might reset approximately to neutral (0.5) whenever there was a change in the location context or a start of a new run. Just as in Model 1, Model 2 did not account for any context-specific learning. Since the degree of this resetting could vary across participants, we added another free parameter to the model, b, which varied from 0 to 1 to determine the degree to which predictions were reset to neutral at the start of each context and at the start of each run according to: $P_{i+1} = b * P_i + (1 - b) * 0.5$. Critically, this resetting step was applied only after the standard updating had been applied at the last trial of the

previous block. For all other trials, shift predictions updated according to the same equation as Model 1.

Both Models 1 and 2 update shift predictions according to trial history but do so in a manner that is context-independent, i.e., no memory of previous experience with a given context is used to adjust predictions when that context re-occurs. Conversely, Model 3 computed shift predictions separately for each location context. Unlike the earlier models, Model 3 did not account for any context-independent trial-by-trial learning of shift likelihood. The context-dependent prediction of Model 3 was defined as $P_{ci} + \alpha_c(S_{ci} - P_{ci})$, where all parameters were defined the same as in the earlier models. Importantly, shift predictions were updated for each context separately such that the prediction on the first trial of a particular context was equal to the value after updating during the last presentation of that specific context. The shift-readiness value for each context began at 0.5 and to simplify the model, we used the same context-based learning rate (α_c) for each context for a given individual.

Unlike Models 1–3, Model 4 accounted for the possibility that participants' shift readiness could vary based on a combination of both (a) the trial-by-trial outcomes captured by the temporal integration of Model 1 and (b) the context-specific learning of Model 3. Thus, Model 4 tested whether participants learned to associate particular contexts with the shift-likelihood probabilities above and beyond any effect of the ongoing trial history. Accordingly, shift-readiness was separately predicted by two learning processes: the context-independent prediction from Model 1 [$P_i + \alpha_t(S_i - P_i)$] and the context-dependent prediction from Model 3 [$P_{ci} + \alpha_c(S_{ci} - P_{ci})$]. Each prediction had a dedicated learning rate α_t and α_c for context-independent and context-dependent predictions, respectively, in order to account for the difference in which the predictions were updated.

Model 5 was identical to Model 4, but added the temporal integration resetting factor from Model 2. Consequently, Model 5 accounted for independent temporal integration and context-based sources of shift prediction but allowed for the possibility that temporal integration predictions reset, at least partially, to 0.5 whenever there was a change in context or run. All equations were identical to those listed above.

Finally, Model 6 also accounted for independent temporal integration and context-based sources of shift prediction but differed from Model 5 in how predictions were reset when there was a change in context or run. Unlike Model 5, we restricted the reset factor to values of only 1 or 0, effectively limiting the possibilities to no reset of predictions or to a complete reset of predictions as the context or run changed.

For each participant, we applied a grid search for the best-fitting learning rate parameter(s), and in the case of Models 2, 5, and 6, the reset parameter, to find the best set of free parameters that fit the participant's trial-level response time (RT) data. In each case, we employed a leave-one-run-out cross-validation procedure such that for each participant, we trained the models iteratively on all but one run of data and tested them on the left out run to prevent overfitting. Note that this procedure renders it unnecessary to penalize model complexity in the model comparison (see below). If an RL mechanism can account for shift-readiness learning, as in other domains of cognitive control (Chiu et al., 2017; Jiang et al.,

2015; Waskom, Frank, et al., 2017) participants' RTs should be longest on trials in which their shift expectations are violated, regardless of whether that violation is an unexpected shift or hold cue. Thus, for each participant, we exhaustively searched through a grid of possible a and b parameters, ranging from $a = 0.01$ to $a = 0.99$ and $b = 0.00$ to $b = 1.00$ in increments of 0.01, to select the one that would best minimize the sum of squared error when predicting RTs. For models 4, 5 and 6, this search was done for all possible combinations of a_t and a_c. We excluded all trials in which the participant failed to make an accurate response, trials immediately following an inaccurate response, and trials with an outlier RT according to a non-recursive procedure (see below; Van Selst & Jolicoeur, 1994) when fitting the model. For each combination of free parameters, we ran a general linear model to predict trial-by-trial response times with regressors that coded for the unsigned PE, block type (25% shift, 50% shift, 75% shift), and cue outcome (shift attention or hold attention). Importantly, unsigned PE can be conceptualized as the degree to which a participant's shift prediction differed from the trial outcome, regardless of whether participants were cued to shift or hold attention. For Models 4–6, two separate regressors coded context-independent and context-based unsigned prediction errors, due to the assumptions that shift-readiness was separately predicted by the two learning processes.

For each participant and for each model, we computed the mean squared error as a measure of model performance. Finally, we conducted a Bayesian model comparison using SPM12 (revision number 7219) (Stephan, Penny, Daunizeau, Moran, & Friston, 2009) to determine the model that best accounted for the behavioral data. Model evidence was computed for each model and each participant according to: $evidence_{ms} = -n_s * \ln(MSE_{ms})$, for each model, m, and each subject, s, where n denotes the total number of trials included for the given subject and MSE is the mean of the squared errors across the folds of the cross-validation procedure. The winning model was then run without cross validation to determine the best-fit parameters across the entire dataset for each participant.

fMRI Data Acquisition and Analysis.

All images were acquired with a GE MR750 3T scanner using an 8-channel head coil. We first acquired an anatomical T1-weighted Spoiled Gradient Echo (SPGR) acquisition in 116 1-mm-thick axial slices (TR = 8 ms, TE = 3 ms, FoV = 256 × 256 mm, voxel size = 1 × 1 × 1 mm). All functional images were acquired with a T2*-weighted gradient-echo EPI sequence in 40 interleaved slices (TR = 2000 ms, TE = 25 ms, flip angle = 90°, FoV = 192 × 192 mm, voxel size = 3 × 3 × 3 mm). We acquired a total of 278 volumes for each run of the attention task and a total of 176 volumes for each run of the mapping task. NIFTI images were created from dicoms using dcm2niix (version 1.0.20190410).

Preprocessing and analysis of the fMRI data was carried out in FSL (version 6.0.1) with the exception of the encoding model analysis, which used custom code written in Matlab. First, the skull was removed from the structural scans using fsl_anat. The first 4 functional volumes acquired from each run were excluded to allow the magnetization to reach steady state. The functional images were then corrected for slice acquisition timing, corrected for motion, coregistered to each participant's structural scan, smoothed with a Gaussian kernel of 5 mm FWHM (for all analyses other than the encoding model), intensity scaled, and after

running each GLM, resampled to an isotropic resolution of $2mm^3$ (again, only for analyses other than the encoding model). All data used for the encoding model analysis were preprocessed without spatial smoothing (other than that introduced during normalization into MNI space) to maximize our potential to reconstruct spatial representations. Each structural scan was brought into MNI space using nonlinear warping in FSL's FNIRT with a warp resolution of 10 mm. All functional data were subjected to a high pass filter with a cutoff of 100 s to remove low frequency changes in signal. Finally, all data for the encoding model were brought into MNI space with spline interpolation using the warp field generated by the structural scan, resampled to an isotropic resolution of $3mm^3$ to preserve the dimensionality of the original dataset, and were subsequently converted to z-scores according to the mean and standard deviation of each run.

In an initial general linear model (GLM), we modeled each trial as a function of cue type (shift vs. hold) and probability context (25% shift, 50% shift, 75% shift) using FSL's FEAT. The first level of analysis for this and all subsequent GLMs modeled each run for each participant independently by convolving a period of 2.25 seconds (the length of the cue plus response period) beginning at the time of each cue onset with a double-gamma hemodynamic response function (hrf). Specifically, we included regressors for accurate, non-outlier, response trials in each of the following conditions: shift trials in 25% shift contexts, shift trials in 50% contexts, shift trials in 75% contexts, hold trials in 25% shift context, hold trials in 50% contexts, and hold trials in 75% shift contexts. Finally, we modeled trials in which participants either failed to respond, made an incorrect response, or produced an outlier RT (see below; Van Selst & Jolicoeur, 1994) in a single regressor, as well as the 6 motion parameters from preprocessing, as regressors of non-interest.

In order to detect regions of the brain that code moment-by-moment PE signals, we ran two additional GLMs. One GLM was designed to identify regions that covaried with unsigned shift PE. Unsigned PE reflects the magnitude of the discrepancy between a participant's shift prediction at any moment, and the outcome that they experienced and thus does not differentiate between trials in which participants shifted or held attention. Two regressors modeled shift and hold trials for which the participant made an accurate, non-outlier, response. A third regressor included all trials in which an accurate, non-outlier, response was made and was parametrically modulated according to the magnitude of trial-by-trial demeaned unsigned shift PE (see RL model description above). As in the previous GLM, we also included a regressor for trials in which participants failed to respond, made an incorrect response, or produced an outlier RT, as well as regressors for the 6 subject motion parameters. A final GLM, which tested signed PE, was identical to the one above, except instead of including a single parametrically-modulated regressor, we included two signed PE regressors in order to examine the difference in PE signals between shift and hold trials. The first coded demeaned signed PE for shift attention trials and the second coded demeaned signed PE for hold attention trials. Unlike the analysis of unsigned PE, which does not specify the outcome (shift vs. hold) of the trial, by also testing signed PE, we were able to determine whether violations of shift predictions for shift trials and for hold trials produced differing patterns of brain activity.

All contrasts were initially carried out at the first level of analysis. For the GLM that did not include parametric PE regressors, we first sought to define brain regions that were associated with shifting covert spatial attention. Next, we contrasted activity in the 75% shift context with that in the 25% shift contrast to determine whether there were context-wide modulations of activity according to shift probabilities. Finally, we compared the shift cost in the 75% shift context with that in the 25% shift context (i.e., the interaction of context by cue type) to determine whether the difference in activity between shift and hold attention trials differed as a function of context-defined shift likelihood. To better illustrate the results of the interaction, we independently defined a medial superior parietal lobule (mSPL) / posterior parietal cortex (PPC) ROI for each participant using a leave-one-subject-out (LOSO) approach so that we could extract percent signal change from each condition. The LOSO approach required that we run the group analysis a total of 23 times with each subject's ROI defined as the resulting significant cluster according to FSL's FLAME stage 1 when that subject was excluded from the model. To convert the parameter estimates to units of percent signal change, we used a scale factor of 45.44, which was computed based on the estimated hrf baseline-to-maximum range of an isolated 2.25 second event in a dummy model (see http://mumford.fmripower.org/perchange_guide.pdf).

In the second and third GLMs, we sought to define which brain areas were associated with ongoing changes in unsigned and signed PE, respectively. In order to compute PE, we subtracted the subject-specific model-derived shift prediction for each trial from the model outcome (1= shift trial, 0 = hold trial). As a measure of general violation of expectations, we first defined unsigned PE as the absolute value of the PE scores. To identify the brain regions that were associated with unsigned PE, we contrasted the parameter estimate from the unsigned PE regressor vs. the implicit baseline of the model. Finally, to identify regions associated with PE for shift and hold trials separately, we contrasted shift trial signed PE and hold trial signed PE vs. the implicit baseline independently. Unlike unsigned PE, which represents a general error in prediction, signed PE carries information regarding whether the participant's PE is due to falsely expecting to shift or to hold attention. Given our method of computing PEs, shift trial PE was always positive while hold trial PE was always negative. Thus our contrasts for the third GLM were hold trial PE < baseline and shift trial PE > baseline. Finally, we contrasted shift and hold trial signed PE to determine whether there was a significant difference in the regions recruited for each type of attentional flexibility updating.

In all cases, first-level parameter estimates were brought into MNI space by applying the warp parameters from the structural registration and resampling to an isotropic resolution of $2mm^3$. A fixed effects second-level analysis was then run to compute parameter estimates for each participant. Finally, the parameter estimates from the second-level analyses were subjected to a third-level model that spanned all participants using FSL's FLAME stages 1 and 2.

Finally, we tested whether our RL model-based analysis accounted for the fMRI data better than a simpler model that only encoded whether or not a cue type was presented in a statistically improbable context. In this model, we replaced the unsigned PE regressor from GLM 2 with a regressor that was parametrically modulated in a binary fashion according to

the likelihood of a cue in a particular context. Cues that were presented in statistically unlikely contexts (e.g. a shift cue in a 25% shift context) were assigned a weighting of 1, while all other cues received a weighting of 0. This meant that all cues within the 50% shift context received a parametric weighting of 0. The residual time series from this simplified model, as well as those from the primary unsigned PE model, were brought into MNI space with spline interpolation. We averaged the residuals time-point by time-point within the mSPL / PPC LOSO ROIs described above. We then squared and summed these residual errors across time and across runs for each participant to compute the residual sum of squared error for each participant and model. We used the Bayesian Information Criterion to evaluate whether the RL model better accounted for the data than the simplified model, summing BICs across participants to get a group value per model.

Spatial Encoding Model.—In order to reconstruct spatial maps of attentional priority in visual cortex both in the time window preceding cue onset and in the window following cue onset, we trained an encoding model of visual spatial attention. Given the evidence in recent studies that spatial representations may be reconstructed from brain activity within visual cortex (e.g. Sprague, Ester, & Serences, 2014; Sprague, Saproo, & Serences, 2015; Sprague & Serences, 2013), we restricted the fMRI data to an ROI that encompassed bilateral V1, V2, V3, and human V4 (hV4) as defined according to a recent probabilistic atlas of visual cortical regions (Wang, Mruczek, Arcaro, & Kastner, 2015). Previous research has identified robust modulations of activity throughout extrastriate visual cortex in response to covert shifts of visual attention, making this an ideal site to probe for changes in the strength of attentional selection (e.g. Chiu & Yantis, 2009). To determine the final ROI, we computed the overlap of these atlas-defined regions, resampled to a 3mm isotropic voxel size, and a group mask, thereby removing any voxels for which there were missing data. As in earlier studies, (see Sprague et al., 2016), we defined a set of 37 identical spatial filters that were arranged in a hexagonal grid that was centered in the visual display. The spacing and FWHM of the spatial filters was set to equal a recent study on visual attention that used a similar model training task (Sprague, Itthipuripat, Vo, & Serences, 2018). The center of each filter was spaced 1.59 degrees apart such that they subtended an area both outside as well as inside the stimulus locations in the attention task. Each filter was a Gaussian-like function that was defined with a FWHM of 1.75° according to the equation:

$f(r) = \left(0.5 + 0.5\cos\left(\frac{\pi r}{s}\right)\right)^7 for\ r < s$; 0 *otherwise*, where r denotes the distance from each filter's center and s reflects the distance from each filter's center at which the amplitude of the filter reaches 0. Given the FWHM, s was set to 4.404° in order to make each filter a single round increase in amplitude at one location of the hexagonal grid.

For each of the non-target trials of the mapping task, we generated a binary stimulus mask by entering a 1 at each pixel location where the checkerboard stimulus was present and a 0 at all other pixels. This mask, and consequently the resulting reconstructions, spanned the total stimulated region of the visual field (11.8° by 11.8°) at a resolution of 119 by 119 pixels. Next we computed the overlap of each stimulus mask onto each channel and normalized these channel scores so that the maximum response was set to 1. To train the encoding model, we identified the volume recorded closest in time following the onset of the checkerboard stimulus and then averaged the volumes 6 seconds and 8 seconds after this

initial volume (capturing the peak BOLD response) for each trial. Each voxel's activity was then modelled as a weighted sum of the 37 spatial filters using ordinary least-squares linear regression according to $B = CW$, where B was the BOLD activity average from the model training task (n trials × m voxels), C was the normalized (ranging from 0 to 1) modeled response of each spatial filter for each trial (n trials × k channels), and W was the weight matrix that mapped changes in brain activity to changes in the channel responses (k channels × m voxels).

In order to reconstruct spatial representations, we used the Moore-Penrose pseudoinverse of the estimated weight matrix from the training data. Unlike the univariate process of determining the weight matrix, the pseudoinverse is multivariate and is based on all encoding models across all voxels within the ROI. As in the model training, we selected two volumes to average for each trial of the attentional flexibility task in which participants made an accurate response. We again excluded all trials that were flagged as RT outliers (Van Selst & Jolicoeur, 1994). For analyses looking at pretrial spatial selection, these volumes were acquired 2 seconds prior and at the time of cue onset. For early post-cue selection, we used volumes acquired 2 and 4 seconds after the onset of the attention cue. Finally, for analyses looking at late post-cue selection, we used volumes acquired 6 and 8 seconds after the onset of the attention cue. Importantly, the inverted weight matrix allows for data in voxel space to be mapped back into channel space. For any volume entered into the analysis from the attentional flexibility task, a spatial reconstruction was generated by first computing the level of activity in each spatial channel according to the recorded fMRI signal and inverted weights. Finally, by multiplying the channel activations by the basis set of functions and summing the output, we generated trial-by-trial spatial reconstructions (see Sprague et al., 2016). As a check of the accuracy of the IEM at the group level, we conducted a leave-one-run out procedure for each participant in which we iteratively trained the model on trials from all but one run, and reconstructed target locations for the remaining run, accounting for the trial-specific jittering of each presentation by shifting the basis functions (see task description).

Since stimuli could appear in 6 different locations falling along a circle in the attention task, we spatially rotated the basis set of functions for each trial such that the attended location (with respect to the period just before cue onset) always corresponded to the left location along the horizontal meridian and the unattended location always corresponded to the right location. We then averaged across trials within each context at the individual subject level to compute the mean reconstruction response.

In order to quantify the degree to which learned changes in attentional flexibility modulated spatial representations in the brain, we defined a 1.4° × 1.4° square that was centered at each target location along the horizontal meridian. After averaging the individual reconstructions of each participant for each condition, we averaged all values within each box to compute a single amplitude response that reflected the strength of attentional selection at each stimulus location for each participant. These averages were then entered into repeated measures ANOVAs to probe the influence of (a) attention (attended vs. unattended), (b) cue type (shift attention vs. hold attention), and (c) context manipulation (25% shift, 50% shift, 75% shift) on the magnitude of spatial selection. Critically, classification as attended or unattended

depended on both the temporal epoch as well as the trial type. For pretrial analyses, attended refers to the left square since the reconstructions were rotated such that the cue always appeared in the left location along the horizontal meridian. However, for early and late post-cue analyses, attended refers to the right square on shift trials and the left square on hold trials, accounting for the shift of attention (or lack thereof) that occurred over the course of the trial.

Experimental Design and Statistical Analysis

The experiment conformed to a within-subjects 3 × 2 factorial design. Accordingly, we analyzed the behavioral response time and accuracy data with separate two-way repeated measures analysis of variance (ANOVA) with factors of probability context (low shift probability, equal shift and hold probabilities, and high shift probability) and cue type (shift attention and hold attention). Only trials in which participants made an accurate response were included in the RT analysis. All behavioral and IEM analyses were corrected for violations of the sphericity assumption with the Geisser-Greenhouse correction when Mauchly's test was statistically significant.

Given our cue type probability manipulation, some cells of our design had far more data points than others. In order to prevent the possibility that any outlier removal procedure would disproportionately remove RTs from some conditions, we employed a non-recursive trimming method with a moving standard deviation cutoff criterion that is designed to match the level of data that would be trimmed with a sample of 100 RT measures per cell at a level of 2.5 SD above and below the mean of each condition (Van Selst & Jolicoeur, 1994). This procedure resulted in the loss of less than 3 percent of all trials with an accurate response. Preprocessing and GLM analyses of the fMRI data were carried out in FSL (version 6.0.1). The results of all contrasts at the group level of analysis were thresholded at a voxelwise level of Z=3.1, $p < .001$ and then cluster corrected to maintain a family-wise error rate of .05, thus robustly guarding against false-positive findings (Eklund, Nichols, & Knutsson, 2016). The IEM analyses used custom code running in Matlab (R2018a).

All behavioral and IEM data, as well as the code for running all behavioral and IEM analyses and for generating the corresponding figures, is available at https://osf.io/8zk6n/. The contrast images from the GLM analyses are available for download from https://neurovault.org/collections/3872/. All raw data are available upon request.

Results

Behavioral Results

We first tested whether attentional flexibility (or shift-readiness) varied across the three different shift-probability contexts. There was a significant main effect of cue type, $F(1,22) = 9.60$, $p = .005$, $\eta_p^2 = .304$, as shift trials were associated with slower responses than hold trials. However, the main effect of context was not statistically significant, $F(2,44) = 1.04$, $p = .360$, $\eta_p^2 = .045$. Critically, there was a significant interaction of context and cue, $F(2,44) = 27.18$, $p < .001$, $\eta_p^2 = .553$: as predicted, the behavioral cost in RT associated with shifting (versus holding) attention decreased as shift likelihoods increased (see Figure 2A).

We next analyzed behavioral accuracies. While the main effects of cue type, $F(1, 22) = 0.36$, $p = .553$, $\eta_p^2 = .016$, and context, $F(2, 44) = 1.54$, $p = .226$, $\eta_p^2 = .065$ failed to reach significance, the interaction approached significance, $F(2, 44) = 3.07$, $p = .057$, $\eta_p^2 = .122$. As in the RT data, the behavioral cost in accuracy associated with shifting as opposed to holding attention decreased as shift likelihood increased, thus ruling out the possibility of a speed-accuracy tradeoff accounting for the RT results (see Figure 2B).

One participant correctly stated the probability contingencies between location and shift likelihood during the open-ended question of the debriefing questionnaire. An additional participant wrote that they believed there was a relationship between location and shift likelihood, but was unable to specify the nature of that relationship. The remaining participants reported that they did not notice a relationship between location and shift likelihood. One participant failed to follow instructions on the subsequent forced choice judgment. Of the remaining 22 participants, 6 correctly matched the probability with the correct location for all three contexts. Thus, there was very little explicit knowledge of the task contingencies.

Model Training Task Performance

Participants indicated the detection of the occurrence of infrequent stimulus dimming by pressing a button. Overall, accuracies for target detection ranged from 75% to 100% ($M = 90.68$, $SD = 8.10$). Furthermore, occurrences of false alarms were low, ranging from 0 to 14 over the course of the experiment ($M = 2.00$, $SD = 3.02$). Importantly, the high accuracy on this demanding target detection task suggests that the participants were attending to each training task stimulus in accordance with the task instructions.

Reinforcement Learning Model

In order to probe the neural mechanisms associated with attentional flexibility learning, we fit the behavioral RT data with several RL models (see Method). In particular, the model fitting was performing using cross-validation to control for overfitting with extra free parameters. Furthermore, to select the RL model variant that best explained the behavioral data, we conducted a Bayesian model comparison of our 6 models. Model 6, which accounted for both context-independent and context-dependent learning as well as complete or absent resetting of context-independent learning with context and run changes, had the highest protected exceedance probability (0.617; see Figure 3A). That is, when considering the omnibus risk that all models had equal exceedance probability, Model 6, out of all 6 candidate models, had a probability of 0.617 to be the model best explaining the behavioral data. Overall, the omnibus risk, which was the probability that the exceedance probability was equal across all 6 candidate models, was .023, indicating a low chance of equal performance among candidate models. Moreover, the median R-squared of Model 6 was 0.12, when testing without cross-validation (see below), with a range of 0.05–0.29 across participants. Thus, our results suggest that, at the group level, attentional flexibility learning reflects the combination of both trial-by-trial adjustments in shift readiness that are independent of contexts and the tracking of shift expectations tied to particular learned contexts.

After selecting the winning model, we fit the three free parameters (the learning rate for context-independent trial-by-trial learning: α_t, the learning rate for context-dependent learning: α_c, and the reset factor: b, which could be 1 or 0 to reflect an absent or complete resetting of shift predictions, respectively). This fitting procedure determined optimal context-independent and context-dependent learning rates for each participant, which both indicate the degree to which that participant weighted the most recent (vs. more remote) previous trials when updating their shift predictions. As illustrated in Figures 3B–C, participants' learning rates ranged from 0.01 to 0.99 for both context-independent (M = 0.32, SD = 0.32) and context-dependent (M = 0.45, SD = 0.35) learning. Best-fit reset parameters for each participant are plotted in Figure 3D.

Given that the winning model binarized the prediction reset factor across the transition of contexts and runs, we tested whether participants with a reset parameter of 1 (n=11) demonstrated differential patterns in RT or accuracy compared to those with a reset parameter of 0 (n=12). Importantly, when testing RTs with an added between-subjects factor according to the reset factor, there remained a significant interaction of context by cue, $F(2,42) = 27.52$, Mauchly's $W = 0.71$, Geisser-Greenhouse corrected $p < .001$, $\eta_p^2 = .567$. The main effect of reset factor, as well as all interactions involving the reset factor, failed to reach statistical significance, ps > .309. An identical analysis of behavioral accuracies again yielded an interaction between context and cue that approached statistical significance, $F(2,42) = 3.03$, $p = .059$, $\eta_p^2 = .126$. As in the RT analysis, the main effect of reset factor as well as all interactions involving reset factor were nonsignificant, ps > .134. Furthermore, there were no significant differences in context-independent, t(21) = 0.17, p = .864, or context-dependent, t(21) = 0.90, p = .378, learning rates according to the reset parameter (see Figure 3E–F). Finally, there was a trend of a positive relationship between context-independent and context-dependent learning rates, $r(21) = .41$, p = .051 (see Figure 3G).

We computed trial-by-trial unsigned and signed prediction errors (PEs), defined as the difference between the model-based shift prediction and the actual outcome of that trial, according to each participant's best-fit learning rates and reset factor. The winning model incorporated two learners/predictions, which have to be integrated to guide behavior. In our modeling of the fMRI data, we therefore searched for brain regions that coded for the integrated context-independent and context-dependent predictions (which also mitigates against the possibility of high collinearity of PEs resulting from the two forms of prediction). To this end, we integrated the context-independent and context-dependent predictions with a weighted sum according to the relative contribution of each factor in accounting for each participant's behavioral RTs. Integrated shift predictions were thus computed according to: $P_{int} = \frac{B_t * P_t + B_c * P_c}{B_t + B_c}$, where B_t is the regression coefficient for context-independent predictions, B_c is the regression coefficient for context-dependent predictions, P_t is the trial-by-trial context-independent prediction and P_c is the trial-by-trial context-dependent prediction. If either B_t or B_c was negative for a given participant (n=19), we set that coefficient to 0. No participants had negative coefficients for both B_t and B_c. After computing trial-by-trial integrated shift predictions for each participant, we computed unsigned and signed PE. To illustrate the relationship between unsigned PEs derived from

Model 6 and trial-by-trial RTs, we divided the RT data for each participant into quintiles according to the corresponding strength of unsigned PE (see Figure 3H).

Imaging Data: Conventional Univariate Analyses

First, we assessed the main effects and interaction of the shift and context factors. To probe the main effect of shifting, we contrasted brain activity for trials in which participants shifted spatial attention compared to those in which they held attention at a single location. As in previous studies (e.g. Chiu & Yantis, 2009; Sali, Courtney, et al., 2016; Yantis et al., 2002), shifting attention was associated with an increase of brain activity, relative to when holding attention, within mSPL and surrounding bilateral PPC as well as the bilateral frontal eye fields (FEF; see Figure 4A; coordinates are displayed in Table 2). Conversely, there was greater activity within the ventromedial prefrontal cortex (vmPFC) for attention hold trials than for attention shift trials (see Figure 4B).

Next, we tested for a main effect of shift-probability context (collapsed across shift and hold trials) by contrasting 25% shift with 75% shift contexts, which did not yield any clusters that passed correction for multiple comparisons. We next ran the shift × context interaction contrast to detect brain regions where the effect of shifting varied as a function of shift-likelihood. We observed significantly greater activity within a cluster spanning mSPL / bilateral PPC / bilateral intraparietal sulcus (IPS) / left angular gyrus, as well as clusters in right angular gyrus, left superior / middle frontal gyrus, left lateral occipital cortex (LOC), left middle frontal/ inferior frontal/ precentral gyrus, left anterior frontal cortex, and right middle temporal gyrus for cues appearing in a context for which that cue was rare (e.g. a shift cue in the 25% shift context) than for cues appearing in a context for which that cue was common (e.g. a shift cue in the 75% shift context; see Figure 4C). Thus, the neural shift cost (shift > hold) in these regions was greatest in the condition where shifts were the least likely. Interestingly, there was overlap of the significant clusters identified in the shift > hold contrast and in the shift × context interaction. No clusters passed correction for multiple comparisons when testing the opposite relationship (i.e. contextually common cues > contextually uncommon cues).

To illustrate the source of this significant interaction, we extracted the average percent signal change across all voxels in the mSPL / PPC. We used a leave one subject out approach to define a mSPL / PPC region of interest for each participant using data from all but the left-out participant iteratively. By running FLAME 1 and selecting the cluster falling in the mSPL / PPC for each iteration, we independently defined a mSPL/ PPC ROI for the left-out participant. Averaging the parameter estimates from each participant's ROI yielded a main effect of cue type that approached significance, $F(1,22) = 4.22$, $p = .052$, $\eta_p^2 = .161$, as well as a significant interaction of cue by context, $F(2,44) = 4.95$, $p = .012$, $\eta_p^2 = .184$. The main effect of context failed to reach statistical significance, $F(2,44) = 0.31$, $p = .736$, $\eta_p^2 = .014$ (see Figure 4D).

Imaging Data: Model-based Analyses.

Next, we tested which brain regions may be directly responsible for learning shift-readiness by searching for regions whose activity scaled with the need to update shift-readiness

predictions, that is, with the trial-by-trial variation in the magnitude of unsigned shift PE. Critically, unsigned PE reflects the violation of shift expectations without regard to whether the participant had expected to shift or to hold attention. Unsigned shift PE scaled positively with activity in a set of regions spanning primarily the dorsal and ventral attention networks (see Figure 5A; coordinates are displayed in Table 3). In particular, we identified large clusters in PPC spanning the mSPL as well as bilateral IPS. Additionally, there were significant clusters in the left superior, middle, and inferior frontal gyri, right angular gyrus, the posterior cingulate cortex (PCC), and right LOC. No clusters passed the correction for multiple comparisons when testing for activity that scaled negatively with unsigned PE. We also tested whether shifting attention was still associated with an increase of activity within the mSPL/PPC when including unsigned PE as a regressor in the GLM. A contrast of shift trials > hold trials revealed significant clusters of activity within the mSPL, left FEF, and left superior LOC (see Figure 5B).

In addition to those regions where activity scaled with unsigned PE, it is possible that some brain areas only vary in response magnitude for violations of expectation for a particular cue outcome (hold or shift). Thus, we ran a final GLM with regressors for shift trial and hold trial signed PE. Unlike unsigned PE, signed PE accounts for whether a participant's PE is due to incorrectly expecting to shift or to hold attention. Given the design of our model (see Method), shift attention trials always had positive signed PE while hold attention trials always had negative signed PE. As the magnitude of signed PE for shift trials increased, we observed an increase in brain activity in many of the same regions identified in the unsigned PE analysis above. In particular, we observed a positive relationship between signed PE for shift trials and activity within bilateral IPS, mSPL, left superior, middle, and inferior frontal gyri / precentral gyrus, right SPL, bilateral superior LOC, and right temporoparietal junction (TPJ; see Figure 5C). Similarly, as the absolute magnitude of signed PE for hold trials increased (more negative values), there was a corresponding increase of activity within the mSPL (see Figure 5D; coordinates are displayed in Table 3). A direct comparison of activity associated with increasing shift trial signed PE and hold trial signed PE yielded no clusters that passed correction for multiple comparisons.

Finally, we tested whether our RL model-derived shift predictions accounted for the fMRI data better than a simpler model in which trials were modelled according to whether they were statistically improbable or likely given the current context. This simplified model was identical to the unsigned PE model described above, but instead of using unsigned PE as the parametric modulator, we assigned every trial in which the cue was statistically improbable (e.g. a shift cue in a 25% shift context or a hold cue in a 75% shift context) a 1 and every other trial a 0. We then computed the residual sum of squared error from each model in a mSPL / PPC ROI and compared the model fits using Bayesian Information Criterion (BIC). At the group level, the RL model resulted in lower BIC than the simplified model (difference in BICs = −7.331, $p = .025$), indicating that the former model better accounted for the variance in the fMRI data in the ROI.

In summary, the above results document that participants adjusted their attentional flexibility across the different shift probability contexts and that a RL model accounted for individual differences in learning. Moreover, a model-based fMRI analysis indicated that several

regions, largely clustered throughout the fronto-parietal network and occipital regions, coded moment-by-moment violations of shift predictions, serving to update expectations of contextual shift-likelihood. We next turned to an IEM analysis to adjudicate between different ways in which the behavioral effects could relate to attentional prioritization of spatial locations in visual cortex.

Spatial Encoding Model

From the above results alone, it is unclear whether varying shift-readiness carried any preparatory consequences on attentional selection. More specifically, it is possible that attentional selection broadens to include the to-be-attended location prior to the onset of a shift cue in cases when the participant expects to shift attention. Alternatively, predicted shifts may not change pre-cue attentional selection but instead, the behavioral effects reported here might stem from changes in the execution of the shift of attention itself (i.e., an accelerated shift in high shift-likelihood context). Furthermore, although we interpreted the above results as an indicator of a change in the readiness to shift attention according to contextual probabilities, it is possible that our task was not demanding enough to require an actual shift of attention. If participants were able to attend to both RSVP streams simultaneously, our results would not speak to the ways in which learning influences attentional control, but rather would more specifically apply to the sub-selection of stimuli within a broader focus of attention.

It is difficult (perhaps impossible) to adjudicate between these different mechanistic accounts of our findings based on behavioral data alone. Therefore, we used an inverted encoding model (IEM) analysis of data from regions V1, V2, V3, and hV4 (see Method for ROI definition; Figure 6A–B) to reconstruct the spatial deployment of attention across the three probability contexts. To start, we trained the encoding model on an independent set of data that was collected while participants viewed flickering checkerboard stimuli at locations throughout a central portion of the visual field that encompassed the stimulus locations from the attentional flexibility task. As a check of the accuracy of training, we iteratively trained the encoding model on the visual cortex data for each participant on all but one run of data and tested the model on the left out dataset. By spatially shifting the resulting reconstructions to undo the spatial jittering that was present during training (see Materials Methods), we were able to average across trials to assess the degree to which we could accurately reconstruct stimulus selection at each of the 37 grid locations. As illustrated in Figure 6C, we were able to accurately reconstruct the stimulus location based on brain activity. Moreover, when we applied this trained encoding model to the independent data from the attention shifting task, we were again able to successfully reconstruct the focus of attention. Figure 6D shows reconstructions of attention for hold attention trials only. Note that we have not spatially rotated these reconstructions, as we did in the main analysis, so that the original stimulus locations are apparent.

Next, we used the IEM to reconstruct the spatial deployment of attention throughout the visual field during the pretrial epoch (an average of volumes acquired 2 seconds prior to cue onset and at the time of cue onset) of the attentional flexibility task. If shift-readiness influences the preparatory allocation of attention prior to the onset of the cue, we would

expect to see differences in selection across the three probability contexts. Importantly, we spatially rotated all reconstructions so that the left location along the horizontal meridian was always the location where the cue would appear (and thus where attention should be directed prior to cue onset) and the right location along the horizontal meridian was the ignored RSVP stream at the time of cue onset (see Figure 6D). As illustrated in Figure 6E, pretrial attentional selection was confined to the to-be-attended location and did not differ according to the contextual probability manipulation. To quantify this result, we defined 1.4° × 1.4° squares centered at the locations of the two RSVP streams and averaged the amplitude of the response across all pixels in each square. A 2 × 3 repeated measures ANOVA with factors of attention (attended vs. unattended) and probability context (25% shift, 50% shift, 75% shift) yielded a significant main effect of attention, $F(1,21) = 50.40$, $p < .001$, $\eta_p^2 = .706$. However, the main effect of probability context, $F(2,42) = 2.15$, $p = .129$, $\eta_p^2 = .093$, failed to reach statistical significance. The assumption of sphericity was violated for the interaction (Mauchly's $W = .674$, $p = .019$) and we therefore applied the Geisser-Greenhouse correction. The interaction, $F(2,42) = 1.49$, $p = .240$ $\eta_p^2 = .066$, failed to reach statistical significance. These results suggest that participants primarily restricted attentional selection to the target location where they anticipated the cue onset and that the strength of this preparatory selection did not vary according to the likelihood of an upcoming shift.

It is possible that trial-by-trial variation of shift expectations might covary with the strength of attentional selection. The RL model-derived shift predictions allowed us to pursue a data-driven approach to infer moment-by-moment changes in behavioral flexibility. We therefore took the pretrial reconstructions and, for each participant, ran a linear regression to predict the strength of selection at the attended location minus the strength of selection at the unatteded location according to the trial-by-trial integrated shift predictions produced by the RL model. The regression coefficients ($M = -.02$, $SD = .10$) did not significantly differ from 0, $t(21) = -1.14$, $p = .266$. Thus, it appears that model-derived measures of trial-by-trial shift readiness were not associated with changes in the magnitude of attentional selection at the to-be-attended location during the pretrial window.

Finally, we examined evoked activity for shift and hold trials separately, by reconstructing the deployment of spatial attention for volumes acquired 2–8 seconds after the onset of the cue to determine whether participants shifted attention in accordance to the cues and whether post-cue selection varied according to the context probabilities. We divided this analysis according to whether data were taken in an early post cue temporal window (2 seconds and 4 seconds post cue) or in a late post cue temporal window (6 seconds and 8 seconds post cue). As in the above analysis, we quantified these reconstructions by averaging the amplitudes of pixels falling at the two target locations and running a 2 × 3 × 2 ANOVA with factors of attention (attended vs. unattended), probability context (25% shift, 50% shift, 75% shift), and cue type (shift trials vs. hold trials). Importantly, unlike the pretrial analysis, the classification of "attended" and "unattended" varied according to the cue type. Given the alignment of all reconstructions described above, the right target square in Figure 6 marked the attended location for shift attention trials, while the left target square continued to mark the attended location for hold attention trials. For the early postcue temporal window, there was a significant cue type by attention interaction, $F(1,21) = 82.21$, $p < .001$, $\eta_p^2 = .797$, such that the reconstruction strength was greatest at the unattended

location for shift trials, but greatest at the attended location for hold trials (see Figure 7). This pattern was likely due to sluggish nature of the hemodynamic response such that there was not yet a detectable shift of attention to the cued location. No other main effects or interactions reached statistical significance, *p*s > .056.

More interestingly, we next examined the volumes collected 6–8 seconds after the cue onset. These volumes indicated that participants shifted and held attention according to the cues, preferentially selecting one stimulus over the other instead of attending to both stimuli in the post-cue period (see Figure 8). There was a significant main effect of attention, $F(1,21) = 38.27$, $p < .001$, $\eta_p^2 = .646$, such that selection amplitudes were generally larger in the attended visual field than in the unattended field. Furthermore, there was a significant cue type by attention interaction, $F(1,21) = 24.71$, $p < .001$ $\eta_p^2 = .541$, as the difference in amplitude between attended and unattended locations was greater for hold trials than it was for shift trials. Finally, the three-way interaction reached significance, $F(2,42) = 3.47$, $p = .040$, $\eta_p^2 = .142$, such that the difference between attended and unattended locations was greatest for unexpected cue outcomes. The remaining main effects and interactions of the three-way ANOVA failed to reach statistical significance, *p*s > .118.

Discussion

In the current study, we examined the neural bases of attentional flexibility learning. As in earlier behavioral studies (Sali et al., 2015), participants adjusted their readiness to execute a covert shift of spatial attention according to context-based task regularities such that shift costs were the smallest when shifts were most likely and the largest when shifts were least likely. By fitting the behavioral RT data with a reinforcement learning model, we computed trial-by-trial measures of unsigned and signed shift PE, quantifying the degree to which a participant's shift expectations were violated on each trial. Entering unsigned shift PE as a regressor in a model of BOLD activity revealed that components of the dorsal and ventral attentional control networks, such as the mSPL, IPS, inferior frontal cortex, PCC, and right angular gyrus were implicated in coding moment-by-moment violations of shift expectations such that activity increased with the magnitude of unsigned shift PE. Interestingly, the PCC has previously been implicated in ongoing fluctuations in attentional flexibility such that activity is highest when participants are in an attentionally stable state (Sali et al., 2016). In the current study, the more unexpected a cue was, the larger the response we observed within the PCC. This PCC activity may therefore be instrumental in adjusting an individual's flexibility as they adapt to trial outcomes. Separate analysis of signed PE for shift trials revealed a similar relationship in many of the same brain regions, suggesting that the unsigned PE results may be driven by violations of expectations for shift trials. We also observed an increase of activity within the mSPL as signed PE for hold attention trials increased. Importantly, there were no clusters that passed the correction for multiple comparisons when directly contrasting regions associated with shift trial signed PE and hold trial signed PE, suggesting that similar neural mechanisms may compute violations of attentional control expectations in both cases. Taken together, our results show, for the first time, that the neural mechanisms that have been widely implicated in executing covert shifts of attention (Corbetta, Patel, & Shulman, 2008; Corbetta & Shulman, 2002; Yantis et al., 2002) as well as the PCC, which has recently been implicated in fluctuations of attentional

control states (e.g. Esterman, Noonan, Rosenberg & DeGutis, 2013; Kucyi, Esterman, Riley, & Valera, 2016; Sali et al., 2016) may also play a role in continually adjusting the flexibility of the attentional control system by computing the degree to which real outcomes violate shift expectations.

An important aspect of cognitive control is the ability to adjust flexibility according to learned expectations about the demands of the environment. In addition to learned changes in attentional flexibility with respect to shifts of spatial attention (Sali et al., 2015), individuals are able to adjust their readiness to switch tasks (Chiu & Egner, 2017; Crump & Logan, 2010; Waskom et al., 2017), and resolve stimulus conflict (King, Korb, & Egner, 2012) in response to the statistical regularities of their task environment. This ability to adjust one's level of flexibility in line with changing demands, known as meta-flexibility, is an important component of adaptive behavior that allows an individual to stably maintain a particular control set (such as attending to one's reading material in a noisy room) and yet moments later have the ability to rapidly switch task-sets and the deployment of attention as behavioral goals and demands change. The current study thus importantly adds to this body of work by investigating the neural mechanisms involved in these learned modulations for the domain of attentional control.

Our results are consistent with earlier work in the domain of stimulus-driven attentional orienting, showing that activity within the right TPJ, the IPS, inferior frontal gyrus, basal ganglia, and FEF was larger for unexpected shifts of attention than for expected shifts (Shulman et al., 2009). Critically, our experimental design differed from this earlier study in that participants executed volitional shifts of attention in response to endogenous cues rather than reoriented attention according to the abrupt onset of a salient target. To our knowledge, our study is the first to test the neural mechanisms responsible for updating experientially learned predictions of goal-oriented attentional orienting. Nevertheless, the finding that activity levels in components of the dorsal and ventral attentional control networks vary according to attention shift readiness is shared across both studies, suggesting that the computations involved in learned adjustments of attentional flexibility for both goal-directed and stimulus-driven attentional orienting may rely on at least partially overlapping neural mechanisms.

A growing literature has used model-based analysis of behavior and brain activity to study cognitive control learning (Chiu et al., 2017; Jiang et al., 2015; Waskom et al., 2017; Jiang, et al., 2018). As in the current study, violations of context-based expectations regarding a to-be-executed task scale with brain activity in frontoparietal control regions (Waskom et al., 2017). However, the results of the current study diverge from those regarding control learning in the domain of conflict resolution. By varying the likelihood that participants would receive conflict-inducing stimuli in a Stroop task, Jiang, Beck, Heller, and Egner (2015) found that the anterior insula tracks the volatility of control demands, while the caudate nucleus signals the prediction of upcoming demand. Relatedly, stimulus-control learning prediction errors in the domain of conflict resolution are also coded in the caudate nucleus (Chiu et al., 2017). In the present study, investigating learned shifts in spatial attention, we did not observe PE based activity within subcortical structures. An interesting topic of future study therefore remains the degree to which control learning for conflict

resolution and the deployment of attention, both among task sets held in memory and in visual cognition, rely on divergent neural mechanisms.

Individual differences in control learning strategies pose an additional important topic for future research. In the current study, the model that best accounted for the behavioral data included a binary prediction reset parameter such that for some participants, temporal integration shift predictions reset to a neutral state with each change of location context or start of a new run of trials. For the remaining participants, temporal integration spanned across each context and run. Participants likely began our study with differing beliefs regarding the task structure that may have influenced their learning strategy. For example, some participants may have expected that shift likelihood reset at the beginning of each block of trials. Although we found no significant differences in the behavioral learning effect or in the context-independent and context-dependent best-fit learning rate parameters according to the reset parameter fit, our goal in the current study was not to explore individual differences in learning. Rather, our analysis was meant to account for the group's performance as a whole. Critically, this means that the winning model from the Bayesian model comparison might not best capture any individual subject's learning strategies. Furthermore, while the winning RL model included context-independent and context-dependent predictions, for only 4 participants was there a positive association between both forms of prediction and trial-by-trial RT. An important topic for future inquiry is thus to explore these individual differences in factors such as the weighting of temporal integration and context-dependent predictions and the degree to which individuals reset shift predictions across changes in context.

An alternative approach to modeling attentional flexibility learning would be to use the model-derived shift predictions as a parametric modulator instead of prediction error. Here, we focused on prediction error, as these signals are most likely to reflect the learning process of interest (e.g. Sutton & Barto, 1998). Moreover, the current IEM analyses allowed us probe whether activity within visual regions of the brain varied according to contextual predictions. Nonetheless, the degree to which context-based shift predictions influence activity outside of visual cortex remains an interesting question for future inquiry.

The current study used an IEM model to reconstruct maps of spatial attentional selection (Sprague et al., 2014; Sprague et al., 2016; Sprague et al., 2015; Sprague & Serences, 2013), allowing us to track whether the deployment of spatial attention varied according to shift-readiness. Since all visual factors were held constant across contexts and spatial positions, any difference in reconstruction amplitude could be attributed to attention. States of high shift-readiness might be associated with a broadening of spatial attentional selection, including a weakening of attentional selection at a currently attended location and/or a strengthening of attentional selection at the location where attention will be deployed in the future (Jefferies et al., 2014). While we found robust evidence that participants maintained attention at the currently to-be-attended location, we found no evidence that suggested a relationship between attentional flexibility and pre-cue spatial selection. In particular, our data do not suggest that participants engaged in anticipatory shifts of attention prior to the onset of the cue. Relatedly, we did not find a significant relationship between model-derived

shift predictions and the strength of selection at the to-be-attended location in the pretrial epoch.

One possible interpretation of our results is that modulations of attentional flexibility are most strongly associated with a speeding of the shift process itself rather than in any change in the spatial selection of stimuli. In particular, our shift PE results suggest that the frontoparietal mechanisms involved in executing goal-oriented shifts of attention do respond differentially based on shift expectations. Interpretation of this analysis requires some caution given the lack of temporal precision available with fMRI data. In particular, although the difference did not reach statistical significance, selection strength was numerically greater following hold attention trials than following shift attention trials. Given the lag of the hemodynamic response it is possible that the preparatory differences we detected are at least in part a reflection of activity from the previous trial. Accordingly, due to the temporal structure of our task, we have restricted the window of analysis for the IEM to range from 2 seconds prior to cue onset to 8 seconds after the cue onset. Future research is thus needed to disambiguate the ways in which learned modulations of flexibility carry consequences for stimulus representation in visual cortex. Taken together, our results suggest that moment-by-moment changes in attentional flexibility may be associated with the execution of the shift process itself instead of large preparatory changes in attentional selection.

The study of attentional flexibility learning holds implications for understanding both healthy adaptive behavior as well as neuropsychiatric disorders. All individuals vary over time in their readiness to perform cognitive switches such as a shift of task set or of spatial attention. By understanding the neural mechanisms associated with both intrinsically-generated spontaneous fluctuations in flexibility, as well those involved in learned modulations of control states, we can account for variations in performance both within a single individual as well as understand individual differences in flexibility. Many neuropsychiatric disorders are associated with either abnormally elevated levels of flexibility (e.g. attention deficit hyperactivity disorder and substance abuse; Barkley, 1997; Berridge, 2012; Keiflin & Janak, 2015; Vaurio, Simmonds, & Mostofsky, 2009) or deficient levels of flexibility (e.g. schizophrenia and autism; Koster-Hale & Saxe, 2013; Moore, Dickinson, & Fletcher, 2011; Murray, Corlett, & Fletcher, 2010). However, the degree to which these impairments may be associated with deficient associative learning remain unknown. In one recent study, children with attention deficit hyperactivity disorder (ADHD) failed to form associations between stimuli and monetary rewards, as measured by the degree to which they later captured attention, to the same extent as their typically developing peers (Sali, Anderson, Yantis, Mostofsky, & Rosch, 2018). An interesting topic of future research thus remains the degree to which disorders such as ADHD are associated with impairments in meta-flexibility, or the ability to adjust cognitive flexibility according to changing environmental demands.

In sum, the current study employed a model-based analysis of behavioral and fMRI data to examine the neural mechanisms responsible for learned adjustments in attentional flexibility, or shift-readiness. Activity within components of the dorsal and ventral attentional control networks positively scaled with a trial-by-trial measure of updating shift predictions. Furthermore, an IEM failed to find statistically reliable pretrial context-based differences in

spatial selection, suggesting that attentional control learning might speed the shift process itself rather than influencing the breadth of attentional deployments. Together, our results suggest that a frontoparietal brain network is responsible for dynamically updating shift readiness in line with changing environmental demands.

Acknowledgements:

We would like to thank Thomas Sprague for advice and sharing code for the inverted encoding model analyses and Yu-Chin Chiu and John Pearson for assistance with the reinforcement learning model.

Funding: This research was supported by National Institute of Mental Health grant R01MH097965 to T.E. and National Research Service Award F32AG056080 to J.J.

References

Anderson BA, Laurent PA, & Yantis S (2011). Value-driven attentional capture. Proceedings of the National Academy of Sciences, 108(25), 10367–10371. 10.1073/pnas.1104047108

Awh E, Belopolsky AV, & Theeuwes J (2012). Top-down versus bottom-up attentional control: a failed theoretical dichotomy. Trends in Cognitive Sciences, 16(8), 437–443. 10.1016/j.tics.2012.06.010 [PubMed: 22795563]

Barkley RA (1997). Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. Psychological Bulletin, 121(1), 65–94. 10.1037/0033-2909.121.1.65 [PubMed: 9000892]

Berridge KC (2012). From prediction error to incentive salience: mesolimbic computation of reward motivation. European Journal of Neuroscience, 35(7), 1124–1143. 10.1111/j.1460-9568.2012.07990.x [PubMed: 22487042]

Blais C, Robidoux S, Risko EF, & Besner D (2007). Item-specific adaptation and the conflict-monitoring hypothesis: A computational model. Psychological Review, 114(4), 1076–1086. 10.1037/0033-295X.114.4.1076 [PubMed: 17907873]

Botvinick MM, Braver TS, Barch DM, Carter CS, & Cohen JD (2001). Conflict monitoring and cognitive control. Psychological Review, 108(3), 624–652. 10.1037/0033-295X.108.3.624 [PubMed: 11488380]

Brainard DH (1997). The psychophysics toolbox. Spatial Vision,10(4), 433–436. [PubMed: 9176952]

Castiello U, & Umiltà C (1990). Size of the attentional focus and efficiency of processing. Acta Psychologica, 73(3), 195–209. 10.1016/0001-6918(90)90022-8 [PubMed: 2353586]

Chiu YC, & Egner T (2017). Cueing Cognitive Flexibility: Item-Specific Learning of Switch Readiness. J Exp Psychol Hum Percept Perform. 10.1037/xhp0000420

Chiu YC, Jiang J, & Egner T (2017). The Caudate Nucleus Mediates Learning of Stimulus-Control State Associations. J Neurosci, 37(4), 1028–1038. 10.1523/jneurosci.0778-16.2016 [PubMed: 28123033]

Chiu YC, & Yantis S (2009). A domain-independent source of cognitive control for task sets: shifting spatial attention and switching categorization rules. J Neurosci, 29(12), 3930–3938. 10.1523/jneurosci.5737-08.2009 [PubMed: 19321789]

Corbetta M, Patel G, & Shulman GL (2008). The reorienting system of the human brain: from environment to theory of mind. Neuron, 58(3), 306–324. 10.1016/j.neuron.2008.04.017 [PubMed: 18466742]

Corbetta M, & Shulman GL (2002). Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci, 3(3), 201–215. 10.1038/nrn755 [PubMed: 11994752]

Crump MJC, & Logan GD (2010). Contextual control over task-set retrieval. Attention, Perception, & Psychophysics, 72(8), 2047–2053. 10.3758/bf03196681

Daw ND, Gershman SJ, Seymour B, Dayan P, & Dolan RJ (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron, 69(6), 1204–1215. 10.1016/j.neuron.2011.02.027 [PubMed: 21435563]

Desimone R, & Duncan J (1995). Neural Mechanisms of Selective Visual Attention. Annu. Rev. Neurosci, 18(1), 193–222. 10.1146/annurev.ne.18.030195.001205 [PubMed: 7605061]

Eklund A, Nichols TE, & Knutsson H (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences, 113(28), 7900–7905. 10.1073/pnas.1602413113

Esterman M, Noonan SK, Rosenberg M, & Degutis J (2013). In the zone or zoning out? Tracking behavioral and neural fluctuations during sustained attention. Cerebral Cortex, 23(11), 2712–2723. 10.1093/cercor/bhs261 [PubMed: 22941724]

Folk CL, Leber AB, & Egeth HE (2002). Made you blink! Contingent attentional capture produces a spatial blink. Perception & Psychophysics, 64(5), 741–753. 10.3758/bf03194741 [PubMed: 12201333]

Folk CL, Remington RW, & Johnston JC (1992). Involuntary covert orienting is contingent on attentional control settings. Journal of Experimental Psychology: Human Perception and Performance, 18(4), 1030–1044. 10.1037/0096-1523.18.4.1030 [PubMed: 1431742]

Guzman-Martinez E, Leung P, Franconeri S, Grabowecky M, & Suzuki S (2009). Rapid eye-fixation training without eyetracking. Psychonomic Bulletin & Review, 16(3), 491–496. 10.3758/PBR.16.3.491 [PubMed: 19451374]

Jefferies LN, Gmeindl L, & Yantis S (2014). Attending to illusory differences in object size. Atten Percept Psychophys, 76(5), 1393–1402. 10.3758/s13414-014-0666-7 [PubMed: 24696380]

Jiang J, Beck J, Heller K, & Egner T (2015). An insula-frontostriatal network mediates flexible cognitive control by adaptively predicting changing control demands. Nat Commun, 6, 8165 10.1038/ncomms9165 [PubMed: 26391305]

Jiang J, Wagner AD, & Egner T (2018). Integrated externally and internally generated task predictions jointly guide cognitive control in prefrontal cortex. Elife, 16(7), 10.7554/eLife.39597

Keiflin R, & Janak PH. (2015). Dopamine Prediction Errors in Reward Learning and Addiction: From Theory to Neural Circuitry. Neuron, 88(2), 247–263. 10.1016/j.neuron.2015.08.037 [PubMed: 26494275]

King JA, Korb FM, & Egner T (2012). Priming of Control: Implicit Contextual Cuing of Top-down Attentional Set. Journal of Neuroscience, 32(24), 8192–8200. 10.1523/jneurosci.0934-12.2012 [PubMed: 22699900]

Koster-Hale J, & Saxe R (2013). Theory of Mind: A Neural Prediction Problem. Neuron, 79(5), 836–848. 10.1016/j.neuron.2013.08.020 [PubMed: 24012000]

Kucyi A, Esterman M, Riley CS, & Valera EM (2016). Spontaneous default network activity reflects behavioral variability independent of mind-wandering. Proceedings of the National Academy of Sciences, 113(48), 13899–13904. 10.1073/pnas.1611743113

Moore JW, Dickinson A, & Fletcher PC (2011). Sense of agency, associative learning, and schizotypy. Consciousness and Cognition, 20(3), 792–800. 10.1016/j.concog.2011.01.002 [PubMed: 21295497]

Murray GK, Corlett PR, & Fletcher PC (2010). The Neural Underpinnings of Associative Learning in Health and Psychosis: How Can Performance Be Preserved When Brain Responses Are Abnormal? Schizophrenia Bulletin, 36(3), 465–471. 10.1093/schbul/sbq005 [PubMed: 20154201]

O'Doherty JP, Cockburn J, & Pauli WM (2017). Learning, Reward, and Decision Making. Annu Rev Psychol, 68, 73–100. 10.1146/annurev-psych-010416-044216 [PubMed: 27687119]

O'Doherty JP, Dayan P, Friston K, Critchley H, & Dolan RJ (2003). Temporal difference models and reward-related learning in the human brain. Neuron, 38(2), 329–337. [PubMed: 12718865]

Sali AW, Anderson BA, & Courtney SM (2016). Information processing biases in the brain: Implications for decision-making and self-governance. Neuroethics. 10.1007/s12152-016-9251-1

Sali AW, Anderson BA, & Yantis S (2014). The role of reward prediction in the control of attention. Journal of Experimental Psychology: Human Perception and Performance, 40(4), 1654–1664. 10.1037/a0037267 [PubMed: 24955700]

Sali AW, Anderson BA, & Yantis S (2015). Learned states of preparatory attentional control. Journal of Experimental Psychology: Learning, Memory, and Cognition, 41(6), 1790–1805. 10.1037/xlm0000146

Sali AW, Anderson BA, Yantis S, Mostofsky SH, & Rosch KS (2018). Reduced Value-Driven Attentional Capture Among Children with ADHD Compared to Typically Developing Controls. J Abnorm Child Psychol. 10.1007/s10802-017-0345-y

Sali AW, Courtney SM, & Yantis S (2016). Spontaneous Fluctuations in the Flexible Control of Covert Attention. Journal of Neuroscience, 36(2), 445–454. 10.1523/jneurosci.2323-15.2016 [PubMed: 26758836]

Shulman GL, Astafiev SV, Franke D, Pope DL, Snyder AZ, McAvoy MP, & Corbetta M (2009). Interaction of stimulus-driven reorienting and expectation in ventral and dorsal frontoparietal and basal ganglia-cortical networks. J Neurosci, 29(14), 4392–4407. 10.1523/jneurosci.5609-08.2009 [PubMed: 19357267]

Sprague TC, Ester Edward F., & Serences John T. (2014). Reconstructions of Information in Visual Spatial Working Memory Degrade with Memory Load. Current Biology, 24(18), 2174–2180. 10.1016/j.cub.2014.07.066 [PubMed: 25201683]

Sprague TC, Ester EF, & Serences JT (2016). Restoring Latent Visual Working Memory Representations in Human Cortex. Neuron, 91(3), 694–707. 10.1016/j.neuron.2016.07.006 [PubMed: 27497224]

Sprague TC, Itthipuripat S, Vo VA, & Serences JT (2018). Dissociable signatures of visual salience and behavioral relevance across attentional priority maps in human cortex. Journal of Neurophysiology, 119(6), 2153–2165. 10.1152/jn.00059.2018 [PubMed: 29488841]

Sprague TC, Saproo S, & Serences JT (2015). Visual attention mitigates information loss in small- and large-scale neural codes. Trends in Cognitive Sciences, 19(4), 215–226. 10.1016/j.tics.2015.02.005 [PubMed: 25769502]

Sprague TC, & Serences JT (2013). Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. Nature Neuroscience, 16(12), 1879–1887. 10.1038/nn.3574 [PubMed: 24212672]

Stephan KE, Penny WD, Daunizeau J, Moran RJ, & Friston KJ (2009). Bayesian model selection for group studies. Neuroimage, 460,(4), 1004–1017. 10.1016/j.neuroimage.2008.04.262

Sutton RS, & Barto AG (1998). Reinforcement learning: An introduction. MIT Press Cambridge, MA

Theeuwes J (1992). Perceptual selectivity for color and form. Percept Psychophys, 51(6), 599–606. [PubMed: 1620571]

Theeuwes J (1994). Stimulus-driven capture and attentional set: selective search for color and visual abrupt onsets. J Exp Psychol Hum Percept Perform, 20(4), 799–806. [PubMed: 8083635]

Van Selst M, Jolicoeur P (1994). A solution to the effect of sample size on outlier elimination. Quarterly Journal of Experimental Psychology, 47(3), 631–650. 10.1080/14640749408401131

Vaurio RG, Simmonds DJ, & Mostofsky SH (2009). Increased intra-individual reaction time variability in attention-deficit/hyperactivity disorder across response inhibition tasks with different cognitive demands. Neuropsychologia, 47(12), 2389–2396. 10.1016/j.neuropsychologia.2009.01.022 [PubMed: 19552927]

Wang L, Mruczek RE, Arcaro MJ, & Kastner S (2015). Probabilistic Maps of Visual Topography in Human Cortex. Cereb Cortex, 25(10), 3911–3931. 10.1093/cercor/bhu277 [PubMed: 25452571]

Waskom ML, Frank MC, & Wagner AD (2017). Adaptive Engagement of Cognitive Control in Context-Dependent Decision Making. Cereb Cortex, 27(2), 1270–1284. 10.1093/cercor/bhv333 [PubMed: 26733531]

Wolfe JM, Cave KR, & Franzel SL (1989). Guided search: An alternative to the feature integration model for visual search. Journal of Experimental Psychology: Human Perception and Performance, 15(3), 419–433. 10.1037/0096-1523.15.3.419 [PubMed: 2527952]

Yantis S, & Jonides J (1984). Abrupt visual onsets and selective attention: evidence from visual search. J Exp Psychol Hum Percept Perform, 10(5), 601–621. [PubMed: 6238122]

Yantis S, Schwarzbach J, Serences JT, Carlson RL, Steinmetz MA, Pekar JJ, & Courtney SM (2002). Transient neural activity in human parietal cortex during spatial attention shifts. Nature Neuroscience, 5(10), 995–1002. 10.1038/nn921 [PubMed: 12219097]

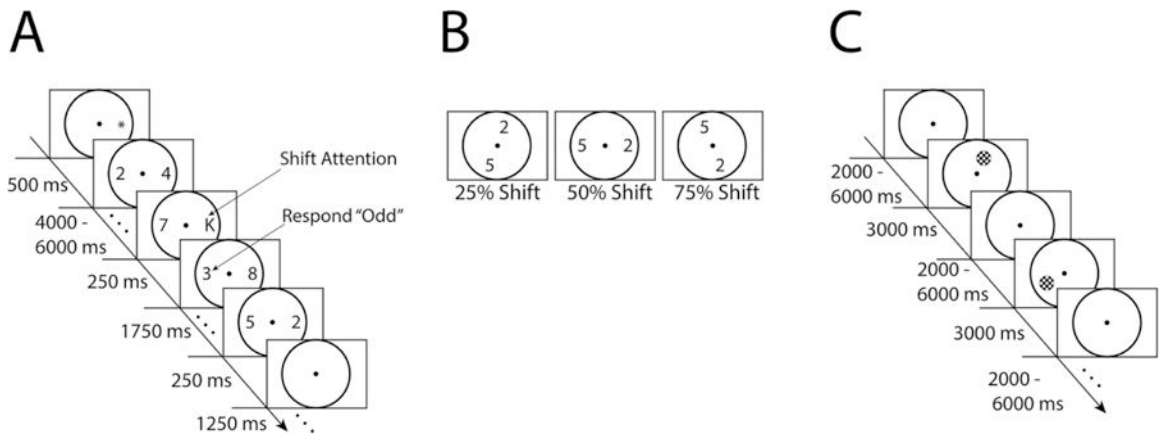

Figure 1.

Task Protocols. (A) Attentional flexibility learning paradigm. Participants covertly shifted spatial attention among continuous streams of alphanumeric characters in response to embedded visual cues and made behavioral responses regarding the parity of target stimuli (B) Three location contexts predicted the likelihood of shifting attention. The location-probability mappings were counterbalanced across participants. (c) Model training task. Participants viewed flickering checkerboard stimuli and made behavioral responses to indicate rare trials in which the checkerboard's luminance dimmed.

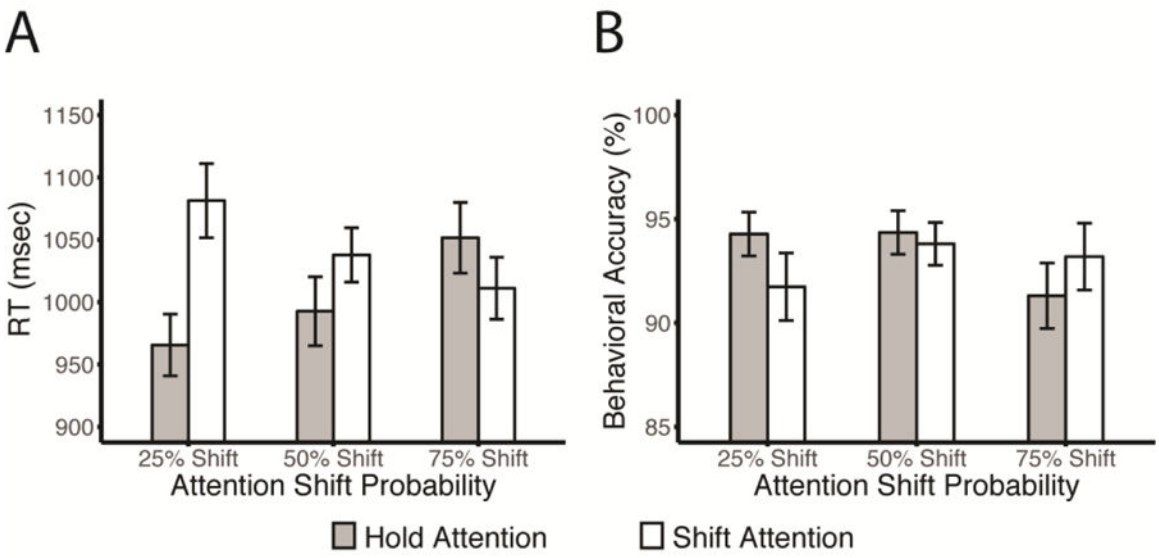

Figure 2.

Behavioral Results. (A) Response time and (B) performance accuracy as a function of attention shift probability and cue type for the attentional flexibility task. Error bars denote 1 between-subjects SEM.

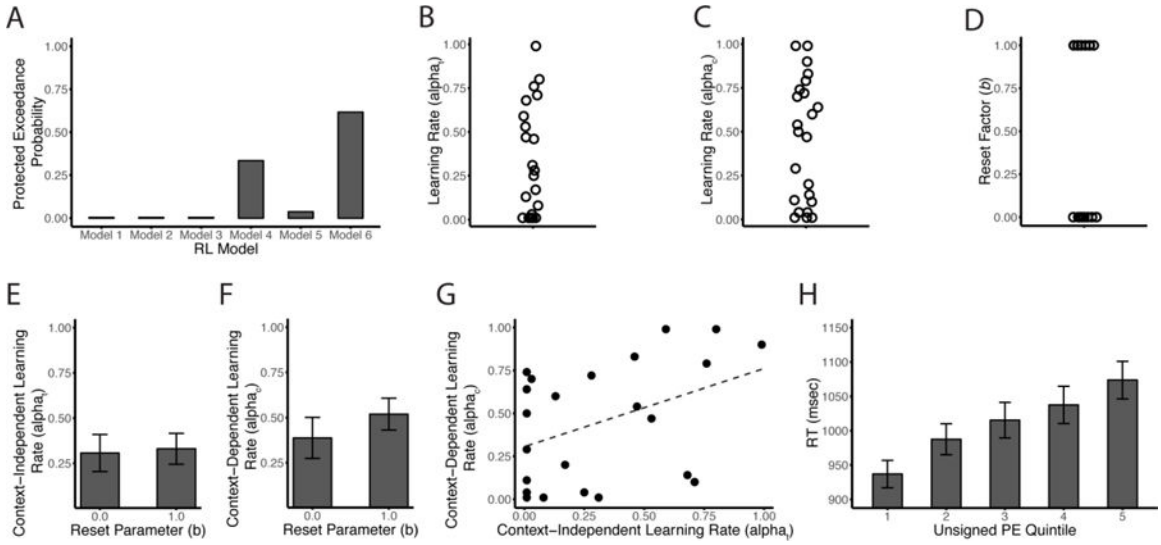

Figure 3.

(A) Protected exceedance probabilities for the 6 candidate models. (B) Model-fit context-independent learning rates. (C) Model-fit context-dependent learning rates. (D) Prediction reset factors. Each circle denotes one participant. There were no significant differences in (E) context-independent learning rates, or (F) context-dependent learning rates according to the best-fit reset parameter. (G) There was a trending positive relationship between context-independent and context-dependent learning rates. (H) Mean response time on the attentional flexibility learning paradigm as a function of the quintile of model-derived unsigned PE. Error bars denote 1 between-subjects SEM.

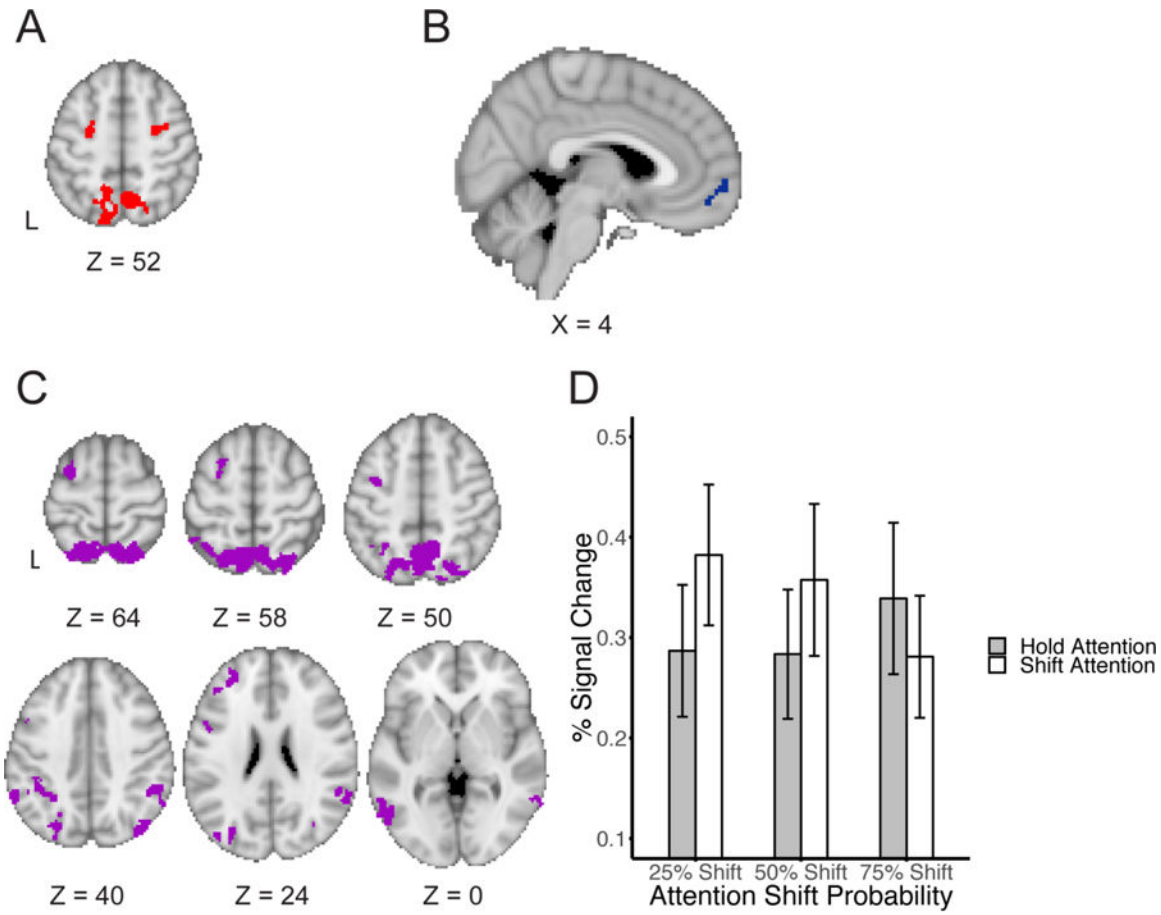

Figure 4.

(A) Shifting attention, relative to holding attention, was associated with greater activity within the PPC and FEFs. (B) Holding attention, relative to shifting attention, was associated with greater activity within the ventromedial prefrontal cortex. (C) The interaction of cue type (shift vs. hold) and context (25% shift vs. 75% shift) revealed that shift and hold cues appearing in statistically rare contexts (e.g. a shift cue in the 25% shift context) were associated with activity spanning bilateral PPC, left frontal cortex, left occipital cortex, right middle temporal gyrus, and right angular gyrus. (D) We independently defined a mSPL ROI for each participant using a leave one subject out approach. Extracted parameter estimates for each condition revealed that BOLD activity was greater for unexpected cues than for expected cues. Error bars denote 1 between-subjects SEM.

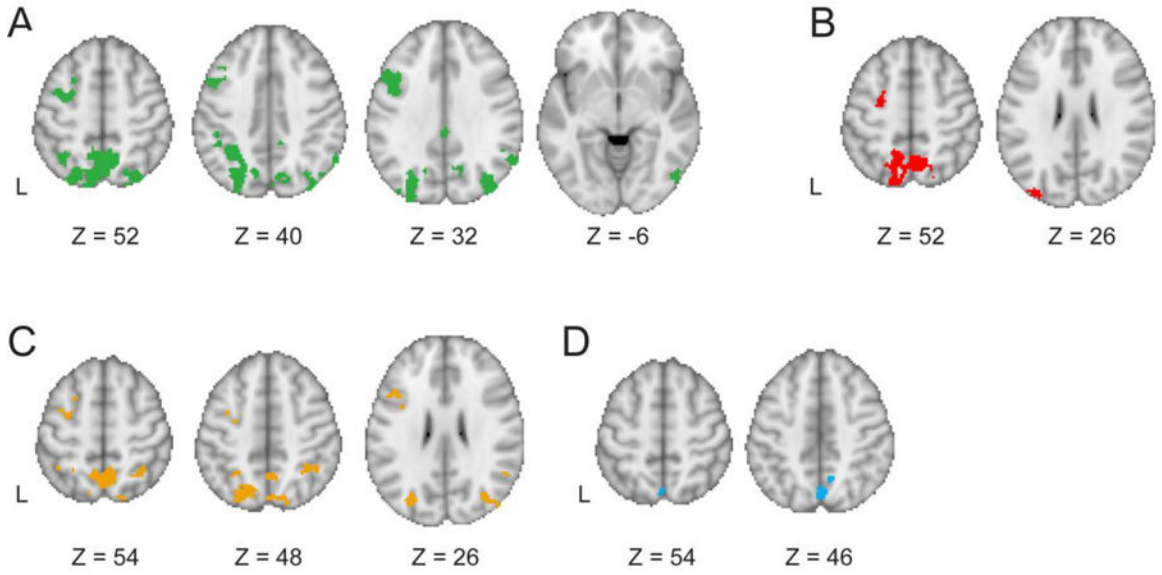

Figure 5.

Regions for which brain activity positively scaled with trial-by-trial (A) unsigned prediction errors. (B) When contrasting shift > hold trials with the PE regressor in the GLM, we found significant activity within mSPL, left FEF, and left superior lateral occipital cortex. Regions for which brain activity positively scaled with (C) signed shift trial prediction errors and (D) signed hold trial prediction errors.

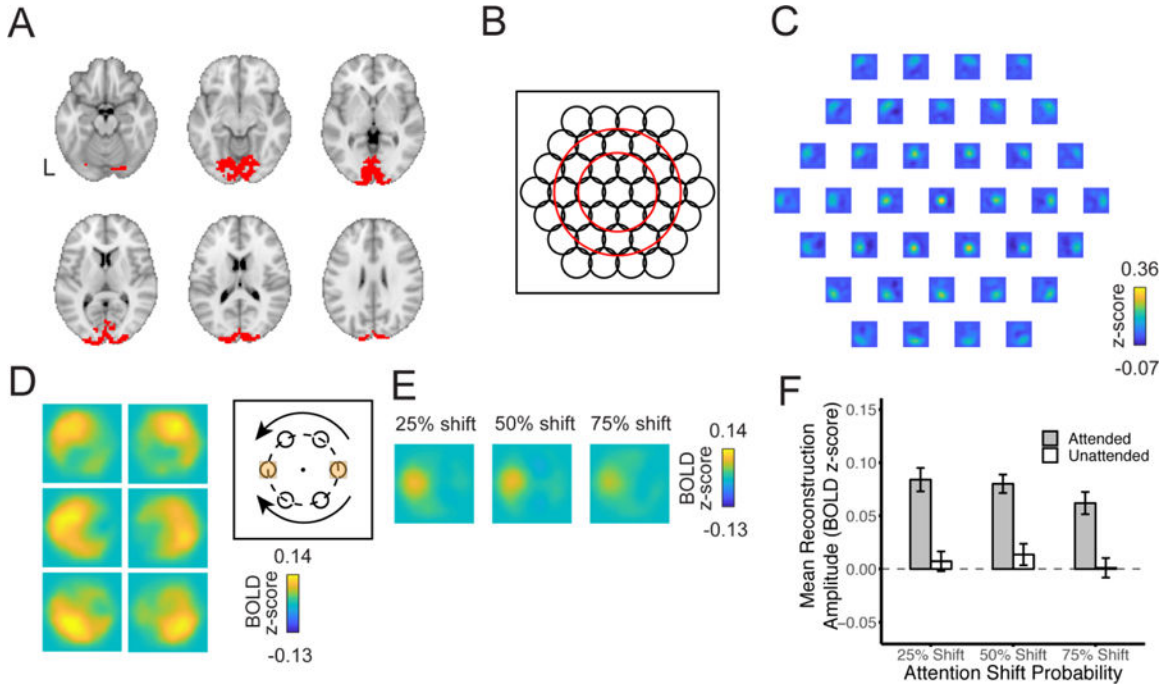

Figure 6.

(A) Visual cortex ROI spanning V1, V2, V3, and hV4. (B) The encoding model consisted of 37 basis functions that were organized in a hexagonal grid that was centered in the visual display. The diameter of each black circle is equal to the FWHM of one function. The inner red circle marks the 1.4° by 1.4° region of the RSVP stimuli that we averaged. (C) Leave-one-run-out cross validation of training data. Each reconstruction represents one of the 37 grid locations used in the model training task. (D) In order to average across participants, we spatially rotated the IEM basis functions such that the to-be-attended location prior to cue onset was located at the left target location falling along the horizontal meridian. Displayed are un-rotated reconstructions for hold attention trials based on BOLD volumes acquired 6 seconds and 8 seconds after the cue onset. After rotating the basis functions for each trial, we averaged reconstructions across trials, and then for each participant, computed an average of pixels falling within two target squares positioned at the left and right locations along the horizontal meridian (marked in yellow above). For the analysis of pretrial signal, the left square marks the attended location, while the right square marks the unattended location. For post-cue analyses, the designation of attended and unattended varied according to cue type such that the right square was attended on shift attention trials and the left square was attended on hold attention trials. (E) Rotated average pretrial spatial reconstructions as a function of probability context across all participants. (F) Average pretrial reconstruction amplitudes at attended and unattended stimulus locations. Spatial selection was greater at the attended location than at the unattended location, but did not vary based on probability context prior to the onset of the attention cue. Error bars denote 1 between-subjects SEM.

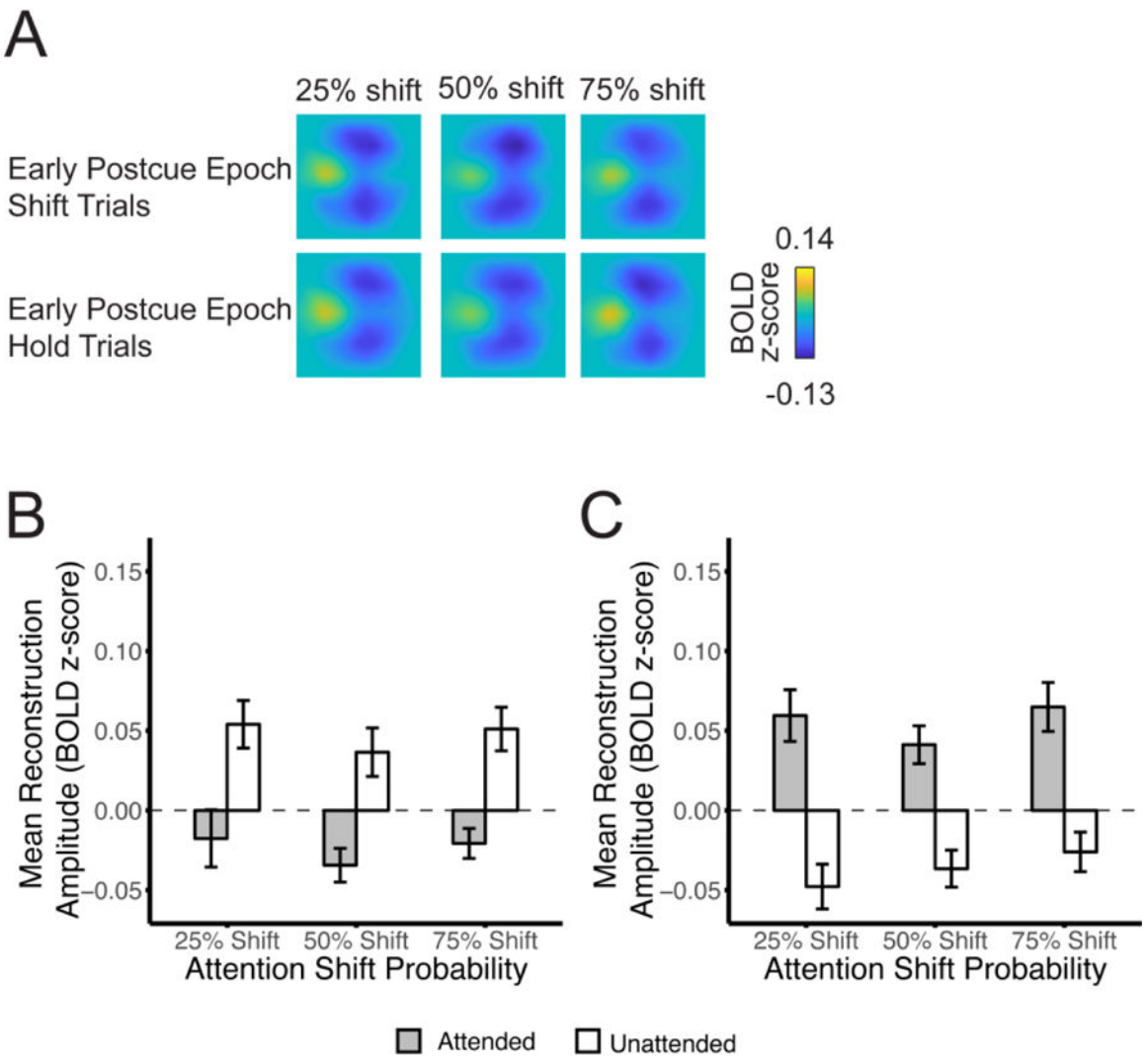

Figure 7.

(A) Spatial reconstructions from 2–4 seconds after the cue onset as a function of probability context and cue type. (B) Average amplitudes at the attended and unattended locations for shift trials, and (C) for hold trials. Attended refers to the right target square for shift trials and the left target square for hold trials. Error bars denote 1 between-subjects SEM.

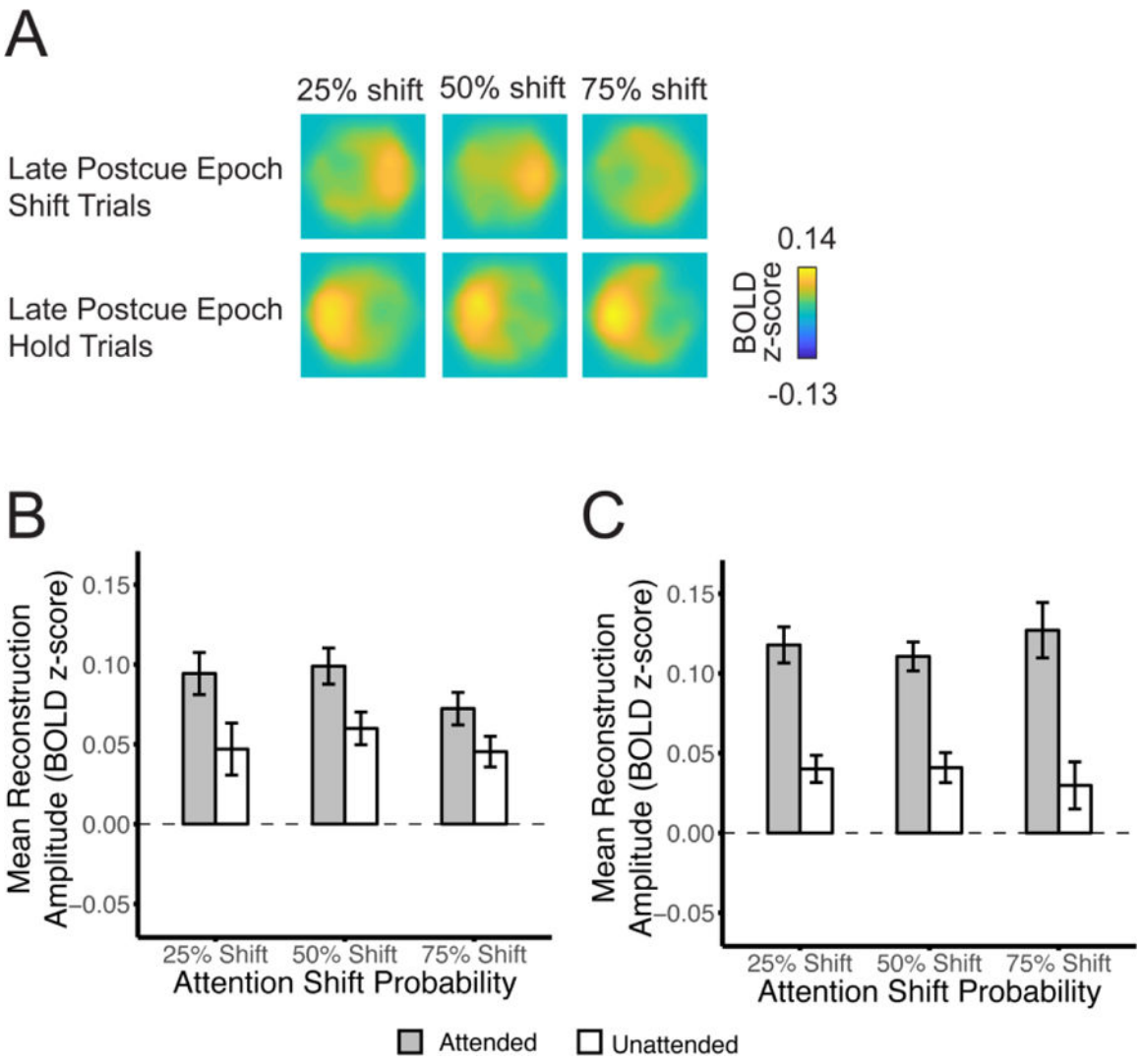

Figure 8.

(A) Spatial reconstructions from 6–8 seconds after the cue onset as a function of probability context and cue type. (B) Average amplitudes at the cued and non-cued locations for shift trials, and (C) for hold trials. Attended refers to the right target square for shift trials and the left target square for hold trials. Error bars denote 1 between-subjects SEM.

Table 1.

Candidate RL Models

	Context-Independent Learning	Context-Dependent Learning	Reset Factor
Model 1	yes	no	no
Model 2	yes	no	yes (partial reset allowed)
Model 3	no	yes	no
Model 4	yes	yes	no
Model 5	yes	yes	yes (partial reset allowed)
Model 6	yes	yes	yes (all-or-none reset required)

Table 2.

Clusters showing significant activity in the conventional univariate analyses.

Region	Cluster Size (voxels)	Z-statistic Maximum		
		x	y	z
Shift > Hold				
mSPL / PPC	646	4	−66	54
right FEF	126	30	−4	54
left FEF	112	−28	−8	62
Hold > Shift				
vmPFC	130	2	52	−12
Unexpected > Expected				
mSPL / bilateral PPC / bilateral IPS / left angular gyrus	3316	−26	−66	62
right angular gyrus	273	54	−44	36
left superior / middle frontal gyrus	221	−34	−4	50
left LOC	158	−54	−68	4
left middle frontal / inferior frontal / precentral gyrus	143	−46	2	30
right middle temporal gyrus	138	62	−50	4
left anterior frontal cortex	136	−28	46	22

Note. Cluster sizes and MNI coordinates are displayed for regions surviving cluster correction at a family-wise error rate of $p < .05$.

Table 3.

Clusters showing significant activity in the model-based analyses.

Region	Cluster Size (voxels)	Z-statistic Maximum		
		x	y	z
Unsigned PE				
mSPL / bilateral PPC / bilateral IPS	4081	42	−72	30
left SFG / left MFG	559	−26	2	62
left MFG / left IFG	378	−48	6	40
right angular gyrus	258	62	−58	28
PCC	186	0	−36	26
right LOC	109	52	−70	−4
Shift > Hold				
mSPL / PPC	760	4	−64	56
left FEF	133	−26	−10	52
left superior LOC	117	−36	−88	22
Signed PE Shift Trials				
mSPL / left IPS / left superior LOC	1188	−28	−78	26
right IPS	339	34	−74	26
right SPL / superior LOC	286	44	−50	48
left SFG / left MFG	248	−30	−10	64
left MFG / left IFG / left precentral gyrus	141	−42	4	32
right TPJ	125	54	−54	24
Signed PE Hold Trials				
mSPL	131	4	−68	50

Note. Cluster sizes and MNI coordinates are displayed for regions surviving cluster correction at a family-wise error rate of $p < .05$. SFG = superior frontal gyrus; MFG = middle frontal gyrus; IFG = inferior frontal gyrus.