Title: Medial prefrontal cortex compresses concept representations through learning

Brief title: Medial PFC compression during learning

Authors: Michael L. Mack[a], Alison R. Preston[b,c,d*], Bradley C. Love[e,f*]

Affiliations:
[a]Department of Psychology, University of Toronto, Toronto, ON, CA
[b]Department of Psychology, The University of Texas at Austin, Austin, TX, USA
[c]Center for Learning and Memory, The University of Texas at Austin, Austin, TX, USA
[d]Department of Neuroscience, The University of Texas at Austin, Austin, TX, USA
[e]Experimental Psychology, University College London, London, UK
[f]Alan Turing Institute, London, UK
*Authors contributed equally

Corresponding Author:
Michael L. Mack
Department of Psychology
University of Toronto
100 St. George Street, 4th Floor
Toronto, Ontario M5S 3G3
mack.michael@gmail.com

Keywords: prefrontal cortex; fMRI; attention; category learning; computational modeling

Abstract

Prefrontal cortex (PFC) is thought to support the ability to focus on goal-relevant information by filtering out irrelevant information, a process akin to dimensionality reduction. Here, we test this dimensionality reduction hypothesis by combining a data-driven approach to characterizing the complexity of neural representation with a theoretically-supported computational model of learning. We find direct evidence of goal-directed dimensionality reduction within human medial PFC during learning. Importantly, by using model predictions of each participant's attentional strategies during learning, we find that that the degree of neural compression predicts an individual's ability to selectively attend to concept-specific information. These findings suggest a domain-general mechanism of learning through compression in mPFC.

Introduction

Prefrontal cortex (PFC) is sensitive to the complexity of incoming information (Badre, Kayser, & D'Esposito, 2010) and theoretical perspectives suggest that a core function of PFC is to focus representation on goal-relevant features by filtering out irrelevant content (Mante, Sussillo, Shenoy, & Newsome, 2013; Wilson, Takahashi, Schoenbaum, & Niv, 2014). In particular, medial PFC (mPFC) is thought to represent the latent structures of experience (Schlichting, Mumford, & Preston, 2015; Zeithamova, Dominick, & Preston, 2012), coding for causal links (Chan, Niv, & Norman, 2016) and task-related cognitive maps (Schuck et al., 2016). At the heart of these accounts is the hypothesis that during learning, mPFC performs data reduction on incoming information, compressing task-irrelevant features and emphasizing goal-relevant information structures. This compression process is goal-directed and akin to how attention in category learning models dynamically selects features that have proven predictive across recent learning trials (Love & Gureckis, 2007; Love, Medin, & Gureckis, 2004). Although emerging evidence suggests structured representations occur in the rodent homologue of mPFC (Farovik et al., 2015), such coding in human PFC remains poorly understood. Here, we directly assess the data reduction hypothesis by leveraging an information-theoretic approach in human neuroimaging to measure how goal-driven learning is supported by attention updating processes in mPFC.

We focused on concept learning, given the recent findings that mPFC represents conceptual information in an organized fashion (Constantinescu, O'Reilly, & Behrens, 2016). Participants learned to classify the same insect images (Figure 1A), composed of three features that could take on two values (thick/thin legs, thick/thin antennae, pincer/shovel mandible), across three different learning problems (Shepard, Hovland, & Jenkins, 1961). These learning problems were defined by rules that required consideration of different numbers of features to successfully classify (see Table 1): the low category complexity problem was unidimensional (e.g., insects living in warm climates have thick legs, cold climate insects have thin legs), the medium category complexity problem depended on two features (e.g., insects from rural environments have thick antennae and shovel mandible or thin antennae and pincer mandible, urban insects have thick antennae and pincer mandible or thin antennae and shovel mandible), and the high category complexity problem required all three features (i.e., each insect's class was uniquely defined by a combination of features). By using the same stimuli for all three problems, the manipulation of conceptual complexity allowed us to target goal-specific learning processes.

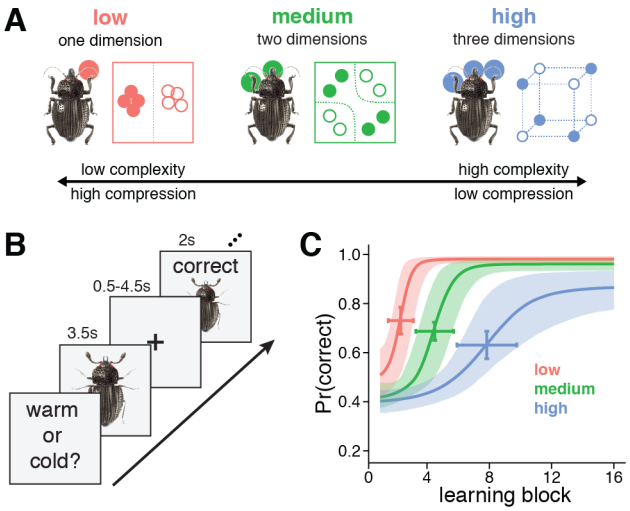

Figure 1: Experimental schematic and behavioral results. **A)** The learning problems differed in rule complexity (see Table S1). The low complexity problem was unidimensional (e.g., antennae size), medium complexity required a conjunction of two features (e.g., leg size and mandible shape), and high complexity required all three features. **B)** Learning blocks consisted of presentation of a stimulus for 3.5s, followed by a fixation cross for 0.5-4.5s, and then a feedback display for 2s that included the stimulus, accuracy of the response, and the correct category. Learning trials were separated by a delay of 2-6s of fixation. **C)** The probability of a correct response increased across learning blocks. The rate of learning differed according to the complexity of the problems.

This design allows us to ask the central question of whether compression in neural representation corresponds with the complexity of the problem-specific conceptual structure throughout learning. Complexity and compression have an inverse relationship; the lower the complexity of a conceptual space, the higher the degree of compression. For instance, in learning the unidimensional problem, variance along the two irrelevant feature dimensions can be compressed resulting in a lower complexity conceptual space. In contrast, learning the high complexity problem requires less compression because all three feature dimensions must be represented, resulting in a relatively more complex conceptual space. Differences in complexity across the three learning problems thus provide a means for testing how learning shapes the dimensionality of neural concept representations. Namely, brain regions involved in goal-directed data compression should *learn* to represent less complex problems with fewer dimensions.

Results

To test this prediction, we recorded functional magnetic resonance imaging (fMRI) data while participants learned the three problems and measured the degree that multivoxel activation patterns were compressed through learning using principal component

analysis (PCA; Figure 2A), a method for low-rank approximation of multidimensional data (Eckart & Young, 1936). Specifically, trial-level neural representations (Mumford, Turner, Ashby, & Poldrack, 2012) for each insect image were submitted to PCA, and the number of principal components (PC) that were necessary to explain 90% of the variance across trials within a learning block was used to calculate an index of neural compression (i.e., fewer PCs reflects more neural compression). This measure of neural compression was calculated across the whole brain with searchlight methods (Kriegeskorte, Goebel, & Bandettini, 2006) for each learning block in each problem. We then identified brain regions that reduce dimensionality with learning (i.e., learn to represent the less complex problems with fewer dimensions) by conducting a voxel-wise linear mixed effects regression on the searchlight compression maps. Specifically, at each voxel, we assessed how neural compression changed as a function of learning block and problem complexity and their interaction.

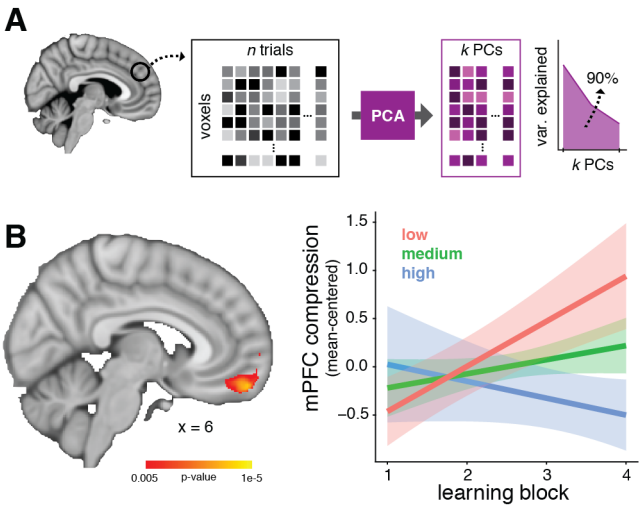

Figure 2: Neural compression analysis schematic and results. **A)** Principal component analysis (PCA) was performed on neural patterns evoked for each of *n* trial within a learning block. The number of principal components (PC) required to explain 90% of the variance (*k*) was used to calculate a neural compression score (1-*k*/*n*). We quantified neural compression as a function of problem complexity and learning block; the interaction of these factors reflects changes in the complexity of neural representations that emerge with learning. **B)** A whole brain voxel-wise linear mixed effects regression revealed an mPFC region that showed a significant interaction between learning block and problem complexity. The nature of the interaction in the mPFC region is depicted in the interaction plot on the left. Shaded regions representing 95% confidence bands.

Throughout the entire brain, only a region within mPFC showed an interaction of problem complexity and learning block (peak coordinates [4, 54, -18]; 653 voxels; voxel-wise threshold = 0.001, cluster extent threshold = 0.05; Figure 2B). The nature of the interaction within this cluster showed that mPFC compression corresponded with

problem complexity and emerged over learning blocks ($F_{1,253.8}$ = 19.02, p = 1.9×10^{-5}). Importantly, the interaction effect was independent of individual differences in learning performance (see Materials and Methods for details about the voxel-wise regression modeling). Because the stimuli were identical across the three problems, this finding demonstrates that learning-related compression is goal-specific, with mPFC requiring fewer dimensions for less complex goals.

To assess whether mPFC compression tracked changes in attentional allocation, we characterized the participant-specific attentional weights given to each stimulus feature across the three problems using a computational learning model (Love et al., 2004). Attention weight entropy indexed changes in attentional allocation; high entropy indicates equivalent weighting to all three features, whereas low entropy indicates attention directed to only one feature. We found that across the learning problems, attention weight entropy increased with conceptual complexity (β = 0.121, SE = 0.0176, t = 6.90, p = 1.43×10^{-8}; Figure 3B). Importantly, the increase in attention weight entropy mirrored the decrease in mPFC neural compression (β = -0.021, SE = 0.005, t = -4.30, p = 9.02×10^{-5}; Figure 3A), suggesting a link between the behavioral and neural signatures of dimensionality reduction.

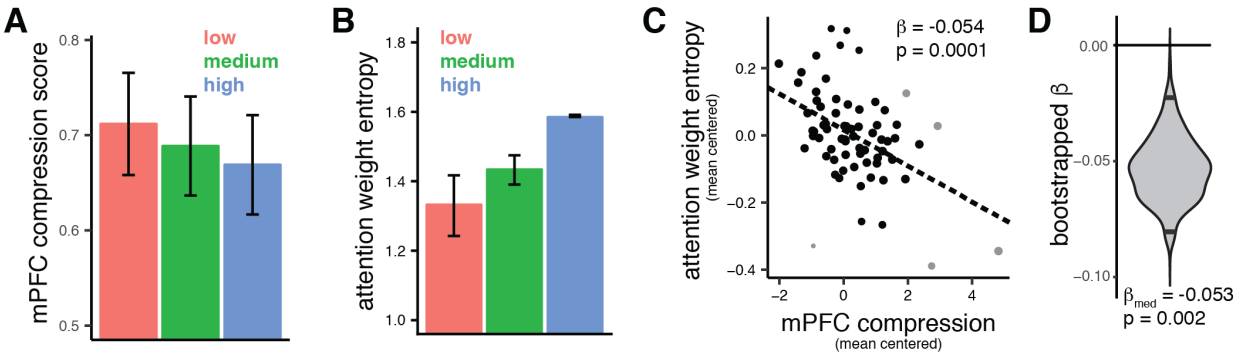

Figure 3: Relationship between mPFC compression and model-based attention weighting in the final learning block. **A)** mPFC neural compression decreased across problems, consistent with their complexity demands. **B)** Attention weight entropy (i.e., dispersion in attention weights) mirrored neural compression, showing attentional strategies consistent with the feature relevancy across the problems (error bars represent 95% CI of the means). **C)** mPFC compression predicted the degree of problem-specific attention weighting (indexed as attention weight entropy). The size of the scatterplot points depicts the weighting from a robust regression analysis; grey-colored points signify observations identified as outliers; the dashed line depicts the best-fitting regression line. **D)** The violin plot depicts a bootstrapped distribution of regression coefficients relating neural compression and attention entropy. Black lines within the distribution mark 95% confidence bounds.

To directly assess this relationship, we evaluated whether entropy of participants' attention weights was predicted by mPFC neural compression at the individual participant level. Specifically, if the ability to compress neural representations in a problem-appropriate fashion is related to participants' ability to attend to problem-relevant features, the prediction follows that participants with more neural compression for a given problem will also show more selective attention, thus lower entropy values. A regression analysis confirmed this hypothesis (β=-0.0536, *SE*=0.0132, t = -4.065, p = 0.0001; see Figure 3C).

To assess the reliability of this finding and evaluate the influence of potential outliers, we performed three additional analyses. First, we analyzed the relationship between neural compression and attention entropy with robust regression using a logistic weighting function. Robust regression accounts for potential outlier observations by down weighting observations that individually influence the estimation of a linear regression model between two variables. Consistent with the correlation results, the robust regression results showed evidence of a linear relationship between neural compression and attention weight entropy (β = -0.0532, *SE* = 0.0122, t = -4.358, p < 0.0001). The weighting of each observation estimated in the robust regression analysis is depicted in Figure 3C as the relative size of the data points. Second, we identified and removed potential outliers by evaluating the standardized difference in fit statistic (DFFITS) for each observation. Using the standard DFFITS threshold (Aguinis, Gottfredson, & Joo, 2013), five observations were identified as outliers (noted as grey data points in Figure 3C). Even with these potential outlier observations removed a strong relationship remained (β = -0.056, *SE* = 0.0122, t = -3.691, p = 0.0005). Third, a nonparametric bootstrap analysis of the linear relationship between neural compression and attention entropy showed a robust effect (see Figure 3D; median β = -0.053, p = *0.002, 95% CI* [-0.0785, -0.0201]). Collectively, these findings suggest that the degree of problem-specific neural compression in mPFC predicted participants' attentional strategies.

Discussion

By focusing on a mechanism by which mPFC forms and represents concepts through goal-sensitive dimensionality reduction, we show that neural representations in mPFC are shaped by experience. And, this shaping is adaptive, promoting efficient representation of information that focuses on encoding features that are most predictive of positive outcomes for a given goal. Importantly, by evaluating behavior through the lens of a theoretically-oriented computational model, we demonstrate that the process of learning to compress in mPFC is consistent with the mechanisms of SUSTAIN (Love

& Gureckis, 2007; Love et al., 2004). These findings provide a quantitative account of mPFC's role in the coding of schematic models or cognitive maps (Constantinescu et al., 2016; Schuck et al., 2016; Wikenheiser & Schoenbaum, 2016), specifically in the conceptual domain.

Successfully learning new concepts requires attending to goal-diagnostic features and ignoring irrelevant information to build abstracted representations that capture the structure defining a concept (Love et al., 2004). Viewed in these terms, concept learning has many parallels to schema formation, a mPFC-related function born out of lesion studies in the memory literature (Gilboa & Marlatte, 2017). Schemas are defined as structured memory networks that represent associative relationships among prior experiences and provide predictions for new experiences (Schlichting et al., 2015; Tse et al., 2011; van Kesteren, Fernández, Norris, & Hermans, 2010; Zeithamova et al., 2012). Schema-related memory behaviors are significantly impacted by mPFC lesions. For example, mPFC lesion patients exhibit a reduced influence of prior knowledge during recognition of items presented in schematically congruent contexts compared with healthy controls (Spalding, Jones, Duff, Tranel, & Warren, 2015). Moreover, mPFC lesions have been associated with a marked inability to differentiate schema-related concepts from concepts inappropriate for a given schema (Ghosh, Moscovitch, Melo Colella, & Gilboa, 2014). From this work, it is clear that mPFC is necessary for retrieving generalized representations built from prior events that are relevant to current experience. Such guided retrieval of relevant learned representations is key to building new concepts.

A key proposal of the SUSTAIN computational model we leveraged is that concept learning is decidedly goal-based, with concept representations adaptively formed to reflect the task at hand (Love et al., 2004). Recent rodent and human work support this proposal with findings that latent mPFC representations are goal-specific in nature, at least at the end of learning. Specifically, neural ensembles in the rodent homologue of mPFC have been demonstrated to represent higher order goal states that relate stimuli to behaviorally-relevant value (Farovik et al., 2015; Lopatina et al., 2017). Similarly, one human neuroimaging study recently localized latent representations of a complex task space relating 16 different goal states to mPFC activation patterns (Schuck et al., 2016). Importantly, these mPFC representations of goal states predicted participants' behavioral performance, supporting the notion that mPFC organizes knowledge based on goals to promote flexible behaviors.

Our findings provide important evidence for the role of mPFC during the *formation* of conceptual maps of experience. Although theoretical perspectives highlight the

important of mPFC in cognitive map formation (Wikenheiser & Schoenbaum, 2016; Wilson et al., 2014), empirical work has failed to directly examine the computations of mPFC contributions during encoding. Instead, evidence is limited to representations that are established after long periods of training (Constantinescu et al., 2016; Schuck et al., 2016). Relatedly, most current models of mPFC function in memory focus on its role in biasing reactivation of relevant prior experiences via the hippocampus (e.g., Miller & Cohen, 2001). Few directly address mPFC's impact at encoding, despite the fact that there is neuroimaging evidence for interactions between mPFC and memory centers during encoding (Mack, Love, & Preston, 2016; Schlichting & Preston, 2016; van Kesteren et al., 2010; Zeithamova et al., 2012). Our findings provide novel evidence for mPFC's important role in encoding processes that build goal-specific mental models. By linking mPFC coding to the learning mechanisms defined in SUSTAIN, our results suggest that mPFC influences encoding through dimensionality reduction wherein selective attention highlights goal-specific information and discards irrelevant dimensions. That mPFC was the only region identified in our analysis suggests that this influence is direct: inputs to mPFC are directly weighted to select goal-related information and discard irrelevant features. These weightings may then be fed back to memory centers (i.e., hippocampus) to impact neural coding of learning experiences (Mack et al., 2016).

Our hypothesized view of mPFC function is based on SUSTAIN's formalism of highly interactive mechanisms of selective attention and learning (Love et al., 2004), functions theoretically mapped onto interactions between PFC and the hippocampus (Love & Gureckis, 2007; Mack et al., 2016). Support for this view is found in recent patient work that has demonstrated a causal link between attentional processes and mPFC function in decision making (Noonan, Chau, Rushworth, & Fellows, 2017; Vaidya & Fellows, 2015, 2016). These studies have shown that lesions to ventral mPFC disrupt attentional guidance based on prior experience with cue-reward associations (Vaidya & Fellows, 2015), learning the value of task-diagnostic features during probabilistic learning (Vaidya & Fellows, 2016), and value comparison during reinforcement learning (Noonan et al., 2017). Relatedly, recent rodent work demonstrates the bidirectional flow of information between mPFC and hippocampus during context-guided memory encoding and retrieval (Place, Farovik, Brockmann, & Eichenbaum, 2016). Coupled with the recent demonstration of hippocampal-mPFC functional coupling during concept learning (Mack et al., 2016), the current findings align well with the view that mPFC is critical for evaluating and representing information in learning and decision making.

In summary, we show that learning can be viewed as a process of goal-directed dimensionality reduction and that such a mechanism is apparent in mPFC neural

representations throughout learning. Thus, mPFC plays a critical role not only in representing conceptual content, but in the process of *learning* concepts. Notably, dimensionality reduction through selective attention offers a reconciling account of many processes associated with mPFC including schema representation (Van Kesteren, Ruiter, Fernández, & Henson, 2012), latent casual models (Schuck et al., 2016), grid-like conceptual maps (Constantinescu et al., 2016), and value coding (Clithero & Rangel, 2013; Grueschow, Polania, Hare, & Ruff, 2015).

Materials and Methods

Participants

Twenty-three volunteers (11 females, mean age 22.3 years old, ranging from 18 to 31 years) participated in the experiment. All subjects were right handed, had normal or corrected-to-normal vision, and were compensated $75 for participating. One participant did not perform above chance in one of the learning problems, thus was excluded from analysis.

Stimuli

Eight color images of insects were used in the experiment (Figure 1A). The insect images consisted of one body with different combinations of three features: legs, mouth, and antennae. There were two versions of each feature (pointy or rounded tail, thick or thin legs, and shovel or pincer mandible). The eight insect images included all possible combinations of the three features. The stimuli were sized to 300 x 300 pixels.

Procedures for the learning problems

After an initial screening and consent in accordance with the University of Texas Institutional Review Board, participants were instructed on the classification learning problems. Participants then performed the problems in the MRI scanner by viewing visual stimuli back-projected onto a screen through a mirror attached onto the head coil. Foam pads were used to minimize head motion. Stimulus presentation and timing was performed using custom scripts written in Matlab (Mathworks) and Psychtoolbox (www.psychtoolbox.org) on an Apple Mac Pro computer running OS X 10.7.

Participants were instructed to learn to classify the insects based on the combination of the insects' features using the feedback displayed on each trial. As part of the initial

instructions, participants were made aware of the three features and the two different values of each feature. Before beginning each classification problem, additional instructions that described the cover story for the current problem and which buttons to press for the two insect classes were presented to the participants. One example of this instruction text is as follows: “Each insect prefers either Warm or Cold temperatures. The temperature that each insect prefers depends on one or more of its features. On each trial, you will be shown an insect and you will make a response as to that insect’s preferred temperature. Press the 1 button under your index finger for Warm temperatures or the 2 button under your middle finger for Cold temperatures.” The other two cover stories involved classifying insects into those that live in the Eastern vs. Western hemisphere and those that live in an Urban vs. Rural environment. The cover stories were randomly paired with the three learning problems for each participant. After the instruction screen, the four fMRI scanning runs (described below) for that problem commenced, with no further problem instructions. After the four scanning runs for a problem finished, the next problem began with the corresponding cover story description. Importantly, the rules that defined the classification problems were not included in any of the instructions; rather, participants had to learn these rules through trial and error.

The three problems the participants learned were structured such that perfect performance required attending to a distinct set of feature attributes (Figure 1A). For the low complexity problem, class associations were defined by a rule depending on the value of one feature attribute. For the medium complexity problem, class associations were defined by an XOR logical rule that depended on the value of the two feature attributes that were not relevant in the low complexity problem. For the high complexity problem, class associations were defined such that all feature attributes had to be attended to respond correctly. As such, different features were relevant for the three problems and successful learning required a shift in attending to and representing those feature attributes most relevant for the current problem. Critically, by varying the number of diagnostic feature attributes across the three problems, the representational space for each problem had a distinct informational complexity.

The binary values of the eight insect stimuli along with the class association for the three learning problems are depicted in Table 1. The stimulus features were randomly mapped onto the attributes for each participant. These feature-to-attribute mappings were fixed across the different classification learning problems within a participant. After the high complexity problem, participants learned the low and medium problems in sequential order. The learning order of the low and medium problems was

counterbalanced across participants. This problem order was used for purposes described in a prior analysis of this data (Mack et al., 2016).

	feature attribute			problem complexity		
stimulus	1	2	3	low	medium	High
1	0	0	0	A	A	B
2	0	0	1	A	B	A
3	0	1	0	A	B	A
4	0	1	1	A	A	B
5	1	0	0	B	A	A
6	1	0	1	B	B	B
7	1	1	0	B	B	B
8	1	1	1	B	A	A

Table 1: Stimulus features and class associations for the three learning problems. Each of the eight stimuli are represented by the binary values of the three feature attributes. The stimuli are assigned to different classes (A or B) across the low, medium, and high complexity learning problems according to rules that depend on one, two, or three of the feature attributes, respectively.

The classification problems consisted of learning trials (Figure 1a) during which an insect image was presented for 3.5s. During stimulus presentation, participants were instructed to respond to the insect's class by pressing one of two buttons on an fMRI-compatible button box. Insect images subtended 7.3° × 7.3° of visual space. The stimulus presentation period was followed by a 0.5-4.5s fixation. A feedback screen consisting of the insect image, text of whether the response was correct or incorrect, and the correct class was shown for 2s followed by a 4-8s fixation. The timing of the stimulus and feedback phases of the learning trials was jittered to optimize general linear modeling estimation of the fMRI data. Within one functional run, each of the eight insect images was presented in four learning trials. The order of the learning trials was pseudo randomized in blocks of sixteen trials such that the eight stimuli were each presented twice. One functional run was 194s in duration. Each of the learning problems included four functional runs for a total of sixteen repetitions for each insect stimulus. The entire experiment lasted approximately 65 minutes.

Behavioral analysis

Participant-specific learning curves were extracted for each problem by calculating the average accuracy across blocks of sixteen learning trials. These learning curves were used for the computational learning model analysis.

401
402 *Computational learning modeling*
403
404 Participant behavior was modeled with an established mathematical learning model,
405 SUSTAIN (Love et al., 2004). SUSTAIN is a network-based learning model that
406 classifies incoming stimuli by comparing them to memory-based knowledge
407 representations of previously experienced stimuli. Sensory stimuli are encoded by
408 SUSTAIN into perceptual representations based on the value of the stimulus features.
409 The values of these features are biased according to attention weights operationalized
410 as receptive fields on each feature attribute. During learning, these attention weight
411 receptive fields are tuned to give more weight to diagnostic features. SUSTAIN
412 represents knowledge as clusters of stimulus features and class associations that are
413 built and tuned over the course of learning. New clusters are recruited and existing
414 clusters updated according to the current learning goals. A full mathematical
415 formulization of SUSTAIN is provided in its introductory publication (Love et al., 2004).
416
417 To characterize the attention weights participants formed during learning, we fit
418 SUSTAIN to each participant's learning performance. First, SUSTAIN was initialized
419 with no clusters and equivalent attention weights across the stimulus feature attributes.
420 Then, stimuli were presented to SUSTAIN in the same order as a participant's
421 experience, and model parameters were optimized to predict each participant's learning
422 performance (mean accuracy averaged over blocks of 16 trials) in the three learning
423 problems through a maximum likelihood genetic algorithm optimization method (Storn &
424 Price, 1997). In the fitting procedure, the model state at the end of the first learning
425 problem was used as the initial state for the second learning problem. In doing so,
426 parameters were optimized to account for learning with the assumption that attention
427 weights, and knowledge clusters learned from the first problem carried over to influence
428 learning in the second problem. Similarly, model state from the second problem carried
429 over and influenced early learning in the third problem. Thus, problem order effects are
430 considered a natural consequence of our model fitting approach. The optimized
431 parameters were then used to extract measures of feature attribute attention weights
432 during the second half of learning in the three problems. Specifically, for each
433 participant, the model parameters were fixed to the optimized values and the model was
434 presented with the trial order experienced by the participant. After the model was
435 presented with the first 96 of trials, the values of the feature attribute attention weights
436 were extracted for each participant. This was repeated for each of the three learning
437 problems. The average value and 95% confidence intervals of the SUSTAIN's five free
438 parameters were: $\gamma = 3.286 \pm 2.064$, $\beta = 4.626 \pm 0.220$, $\eta = 0.308 \pm 0.145$, $d = 20.293 \pm$
439 5.724, $\tau_h = 0.112 \pm 0.039$.

MRI data acquisition

Whole-brain imaging data were acquired on a 3.0T Siemens Skyra system at the University of Texas at Austin Imaging Research Center. A high-resolution T1-weighted MPRAGE structural volume (TR = 1.9s, TE = 2.43ms, flip angle = 9°, FOV = 256mm, matrix = 256x256, voxel dimensions = 1mm isotropic) was acquired for coregistration and parcellation. Two oblique coronal T2-weighted structural images were acquired perpendicular to the main axis of the hippocampus (TR = 13,150ms, TE = 82ms, matrix = 384x384, 0.4x0.4mm in-plane resolution, 1.5mm thru-plane resolution, 60 slices, no gap). High-resolution functional images were acquired using a T2*-weighted multiband accelerated EPI pulse sequence (TR = 2s, TE = 31ms, flip angle = 73°, FOV = 220mm, matrix = 128x128, slice thickness = 1.7mm, number of slices = 72, multiband factor = 3) allowing for whole brain coverage with 1.7mm isotropic voxels.

MRI data preprocessing and statistical analysis

MRI data were preprocessed and analyzed using FSL 6.0 (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012) and custom Python routines. Functional images were realigned to the first volume of the seventh functional run to correct for motion, spatially smoothed using a 3mm full-width-half-maximum Gaussian kernel, high-pass filtered (128s), and detrended to remove linear trends within each run. Functional images were registered to the MPRAGE structural volume using Advanced Normalization Tools, version 1.9 (Avants et al., 2011).

Neural compression analysis

The goal of the neural compression analysis was to assess the informational complexity of the neural representations formed during the different learning problems. To index representational complexity, we measured the extent that neural activation patterns could be compressed into a smaller dimensional space according to principal component analyses (PCA). The compression analyses were implemented using PyMVPA (Hanke et al., 2009) and custom Python routines and were conducted on preprocessed and spatially smoothed functional data. First, whole brain activation patterns for each stimulus within each run were estimated using an event-specific univariate general linear model (GLM) approach (Mumford et al., 2012). This approach allowed us to model stable estimates of neural patterns for the eight insect stimuli across the trials in each learning problem. For each classification problem run, a GLM with separate regressors for stimulus presentation on each trial, modeled as 3.5s

boxcar convolved with a canonical hemodynamic response function (HRF), was conducted to extract voxel-wise parameter estimates for each trial. Additionally, trial-specific regressors for the feedback period of the learning trials (2s boxcar) and responses (impulse function at the time of response), as well as six motion parameters were included in the GLM. This procedure resulted in, for each participant, whole brain activation patterns for each trial in the three learning problems.

We assessed the representational complexity of the neural measures of stimulus representation during learning with a searchlight method (Kriegeskorte et al., 2006). Using a searchlight sphere with a radius of 4 voxels (voxels per sphere: 242 mean, 257 mode, 76 minimum, 257 maximum), we extracted a vector of activation values across all voxels within a searchlight sphere for all 32 trials within a problem run. These activation vectors were then submitted to PCA to assess the degree of correlation in voxel activation across the different trials. PCA was performed using the singular value decomposition method as implemented in the decomposition.PCA function of the scikit-learn (version 0.17.1) Python library. To characterize the amount of dimensional reduction possible in the neural representation, we calculated the number of principal components that were necessary to explain 90% of the variance (*k*) in the activation vectors. We scaled this number into a compression score that ranged from 0 to 1,

$$compression = 1 - \frac{k}{n},$$

where *n* is equal to 32, the total number of activation patterns submitted to PCA. By definition, 32 PCs will account for 100% of the variance, but no compression. With this definition of neural compression, larger compression scores indicated fewer principal components were needed to explain the variance across trials in the neural data (i.e., neural representations with lower dimensional complexity). In contrast, smaller compression scores indicated more principal components were required to explain the variance (i.e., neural representations with higher dimensional complexity). This neural compression searchlight was performed across the whole brain separately for each participant and each run of the three learning problems in native space.

Group-level analyses were performed on the neural compression maps calculated with the searchlight procedure. Each participant's compression maps were normalized to MNI space using ANTs (Avants et al., 2011) and combined into a group dataset. To identify brain regions that demonstrated neural compression that was consistent with the representational complexity of the learning problems, we performed a voxel-wise linear mixed effects regression analysis. The mixed effects model included factors of

problem complexity and learning block as fixed effects as well as participants as a random effect to predict neural compression. The interaction of problem complexity and learning block was the central effect of interest. We also included each participant's accuracy for the three problems within each learning block as a covariate. This regression model was evaluated at each voxel. A statistical map was constructed by saving the *t*-statistic of the interaction between complexity and learning block. The resulting statistical map was voxel-wise corrected at $p = 0.001$ and cluster corrected at $p = 0.05$ which corresponded to a cluster extent threshold of greater than 259 voxels. The cluster extent threshold was determined with AFNI (Cox, 1996) 3dClustStim (version 16.3.12) using the *acf* option, second-nearest neighbor clustering, and 2-sided thresholding. The 3dClustSim software used was downloaded and compiled on November 21, 2016 and included fixes for the recently discovered errors of improperly accounting for edge effects in simulations of small regions and spatial autocorrelation in smoothness estimates (Eklund, Nichols, & Knutsson, 2016).

We assessed the nature of the interaction in the mPFC cluster by extracting each participant's average neural compression score within the cluster for each problem across the four learning runs. The same linear mixed effects model described above was run on the extracted compression values. It is important to note that this analysis was conducted to characterize the interaction underlying the mPFC cluster and, therefore, does not represent a set of independent findings. The results of this model are shown in Table 2.

	estimate	SE	df	*t*	*p*
intercept	-2.260	0.632	218.4	-3.58	4.3×10^{-4}
block	0.789	0.164	219.1	4.82	2.7×10^{-6}
complexity	0.566	0.252	140.6	2.24	0.0266
accuracy	0.945	0.461	221.2	2.05	0.0415
block:complexity	-0.321	0.074	253.8	-4.36	1.9×10^{-5}

Table 2: Results of the linear mixed effects regression model predicting average neural compression within the mPFC region depicted in Figure 2B. The estimated values, standard errors (SE), Satterthwaite approximations of degrees of freedom (df), *t*-statistics, and *p* values are reported for each fixed effect.

Relating neural compression to behavioral signatures of selective attention

To evaluate the relationship between neural compression and model-based estimates of attention weighting, we first extracted individual participant-based measures of each. Because we were interested in the outcome of learning, we focused on the final learning

block. The participant-specific average neural compression within the mPFC cluster was extracted for each learning problem. We used the SUSTAIN parameter estimates of stimulus dimension attention weights to calculate a signature of selective attention. Specifically, the attention weight estimates for the three stimulus dimensions in each problem were transformed to sum to 1, thus creating a probability distribution representing the likelihood of attention to the three features. For example, given the attention weights [0.1, 0.1, 0.8], there is a probability of 0.8 that attention will be directed to the third stimulus dimension on any one trial. We then calculated entropy (Davis, Love, & Preston, 2012) across the attention weights for each problem separately:

$$entropy = -\sum_{i=1}^{3} a_i \log_2 a_i,$$

such that a_i is the attention weight for stimulus dimension *i*. This entropy measure indexed the dispersion of attention across the stimulus dimensions. For example, high attention weight entropy means that attention is unselective with all three stimulus dimensions equally weighted. On the other hand, low entropy means that attention is highly predictive with the majority of weight on a single dimension. As such, the attention weight entropy index offers a unique signature for optimal attentional strategy across the three learning problems: the lowest entropy should be seen in the low complexity problem, an intermediate entropy for the medium complexity problem, and the highest entropy for the high complexity problem. The effect of problem complexity on both mPFC neural compression and attention weight entropy was assessed with linear mixed effects regression (see Figure 3A and 3B).

We next evaluated the relationship between mPFC neural compression and attention weight entropy on an individual participant basis with linear regression and several follow up analysis. We first mean centered both measures within participant and entered the resulting measures into a regression model with neural compression as a predictor of attention weight entropy (Figure 3C). We performed three additional analyses to assess the reliability of the regression results and to evaluate the influence of potential outliers. First, we reran the regression analysis with robust regression using a logistic weighting function. Robust regression accounts for potential outlier observations by down weighting observations that individually influence the estimation of a linear regression model between two variables. The weighting of each observation estimated in the robust regression analysis is depicted in Figure 2C as the relative size of the data points. Second, we identified and removed potential outliers by evaluating the standardized difference in fit statistic (DFFITS) for each observation. The standard DFFITS threshold of $\pm 2\sqrt{(k+1)/n}$ (Aguinis et al., 2013) identified five observations as potential outliers (noted as a grey data point in Figure 3C). These observations were

589 excluded and the linear regression analysis was performed again. Third, we performed
590 a nonparametric bootstrap analysis to assess the robustness of the regression findings.
591 We randomly sampled participants' data with replacement from the compression and
592 entropy observations 5000 times, calculating and storing the regression coefficient on
593 each iteration. The 95% confidence interval of the resulting distribution of correlation
594 coefficients was then compared to 0 to determine the robustness of the mPFC
595 compression and attention weight entropy relationship (Figure 3D).
596
597
598 **Acknowledgments**
599
600 Thanks to Christiane Ahlheim for manuscript comments. M.L.M. was supported by an
601 NSERC Discovery Grant and NIMH grant F32-MH100904; A.R.P. by NIMH grant R01-
602 MH100121, and NSF CAREER Award 1056019; and B.C.L by Leverhulme Trust grant
603 RPG-2014-075, Wellcome Trust Senior Investigator Award WT106931MA, and NICHD
604 grant 1P01HD080679.
605
606 **Author Contributions**
607
608 All authors designed the experiment and wrote the paper. M.L.M. conducted the
609 research and data analysis.
610
611 **Competing Interests**
612
613 The authors declare no competing interests.
614
615

References

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods*, *16*(2), 270–301. https://doi.org/10.1177/1094428112470848

Avants, B. B., Tustison, N. J., Song, G., Cook, P. a., Klein, A., & Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, *54*(3), 2033–2044. https://doi.org/10.1016/j.neuroimage.2010.09.025

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron*, *66*(2), 315–326. https://doi.org/10.1016/j.neuron.2010.03.025

Chan, S. C. Y., Niv, Y., & Norman, K. A. (2016). A Probability Distribution over Latent Causes, in the Orbitofrontal Cortex. *The Journal of Neuroscience*, *36*(30), 7817–28. https://doi.org/10.1523/JNEUROSCI.0659-16.2016

Clithero, J. A., & Rangel, A. (2013). Informatic parcellation of the network involved in the computation of subjective value. *Social Cognitive and Affective Neuroscience*, *9*(9), 1289–1302. https://doi.org/10.1093/scan/nst106

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, *352*(6292), 1464–1468. https://doi.org/10.1126/science.aaf0941

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*(29), 162–173. https://doi.org/10.1006/cbmr.1996.0014

Davis, T., Love, B. C., & Preston, A. R. (2012). Striatal and hippocampal entropy and recognition signals in category learning: Simultaneous processes revealed by model-based fMRI. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 821–839. https://doi.org/10.1037/a0027865

Eckart, C., & Young, G. (1936). The Approximation of One Matrix by Another Low Rank. *Psychometrika*, *1*(3), 211–218.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, *113*(33), 201602413. https://doi.org/10.1073/pnas.1602413113

Farovik, A., Place, R. J., McKenzie, S., Porter, B., Munro, C. E., & Eichenbaum, H. (2015). Orbitofrontal Cortex Encodes Memories within Value-Based Schemas and Represents Contexts That Guide Memory Retrieval. *Journal of Neuroscience*, *35*(21), 8333–8344. https://doi.org/10.1523/JNEUROSCI.0134-15.2015

Ghosh, V. E., Moscovitch, M., Melo Colella, B., & Gilboa, A. (2014). Schema representation in patients with ventromedial PFC lesions. *The Journal of Neuroscience*, *34*(36), 12057–70. https://doi.org/10.1523/JNEUROSCI.0740-14.2014

Gilboa, A., & Marlatte, H. (2017). Neurobiology of Schemas and Schema-Mediated

Memory. *Trends in Cognitive Sciences*, *14*(0), 417–428. https://doi.org/10.1016/j.tics.2017.04.013

Grueschow, M., Polania, R., Hare, T. A., & Ruff, C. C. (2015). Automatic versus Choice-Dependent Value Representations in the Human Brain. *Neuron*, *85*(4), 874–885. https://doi.org/10.1016/j.neuron.2014.12.054

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V, & Pollmann, S. (2009). PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, *7*(1), 37–53. https://doi.org/10.1007/s12021-008-9041-y

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, *62*(2), 782–90. https://doi.org/10.1016/j.neuroimage.2011.09.015

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–3868.

Lopatina, N., Sadacca, B. F., McDannald, M. A., Styer, C. V., Peterson, J. F., Cheer, J. F., & Schoenbaum, G. (2017). Ensembles in medial and lateral orbitofrontal cortex construct cognitive maps emphasizing different features of the behavioral landscape. *Behavioral Neuroscience*, *131*(3), 201–212. https://doi.org/10.1037/bne0000195

Love, B. C., & Gureckis, T. M. (2007). Models in search of a brain. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(2), 90–108.

Love, B. C., Medin, D., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*(2), 309–332.

Mack, M. L., Love, B. C., & Preston, A. R. (2016). Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, *113*(46), 13203–13208. https://doi.org/10.1073/pnas.1614048113

Mante, V., Sussillo, D., Shenoy, K. V, & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84. https://doi.org/10.1038/nature12742

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*(1), 167–202. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11283309

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076

Noonan, M. P., Chau, B. K. H., Rushworth, M. F. S., & Fellows, L. K. (2017). Contrasting Effects of Medial and Lateral Orbitofrontal Cortex Lesions on Credit Assignment and Decision-Making in Humans, *37*(29), 7023–7035. https://doi.org/10.1523/JNEUROSCI.0692-17.2017

Place, R., Farovik, A., Brockmann, M., & Eichenbaum, H. (2016). Bidirectional prefrontal-hippocampal interactions support context-guided memory. *Nature*

Neuroscience, *19*(8). Retrieved from http://www.nature.com/doifinder/10.1038/nn.4327

Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications*, *6*, 8151. https://doi.org/10.1038/ncomms9151

Schlichting, M. L., & Preston, A. R. (2016). Hippocampal–medial prefrontal circuit supports memory updating during learning and post-encoding rest. *Neurobiology of Learning and Memory*, *134*, 91–106. https://doi.org/10.1016/j.nlm.2015.11.005

Schuck, N. W., Cai, M. B., Wilson, R. C., Niv, Y., Schuck, N. W., Cai, M. B., … Niv, Y. (2016). Human Orbitofrontal Cortex Represents a Cognitive Map of State Space Article Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron*, *91*(6), 1402–1412. https://doi.org/10.1016/j.neuron.2016.08.019

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classification. *Psychological Monographs*, *75*(13), 517.

Spalding, K. N., Jones, S. H., Duff, M. C., Tranel, D., & Warren, D. E. (2015). Investigating the Neural Correlates of Schemas: Ventromedial Prefrontal Cortex Is Necessary for Normal Schematic Influence on Memory. *J Neurosci*, *35*(47), 15746–15751. https://doi.org/10.1523/JNEUROSCI.2767-15.2015

Storn, R., & Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, *11*, 341–359. https://doi.org/10.1023/A:1008202821328

Tse, D., Takeuchi, T., Kakeyama, M., Kajii, Y., Okuno, H., Tohyama, C., … Morris, R. G. M. (2011). Schema-Dependent Gene Activation. *Science*, *333*(August), 891–895. https://doi.org/10.1126/science.1205274

Vaidya, A. R., & Fellows, L. K. (2015). Ventromedial Frontal Cortex Is Critical for Guiding Attention to Reward-Predictive Visual Features in Humans. *Journal of Neuroscience*, *35*(37). Retrieved from http://www.jneurosci.org/content/35/37/12813?ijkey=55dfc953354fc3a0477599ff2f60f293580642d5&keytype2=tf_ipsecsha

Vaidya, A. R., & Fellows, L. K. (2016). Necessary Contributions of Human Frontal Lobe Subregions to Reward Learning in a Dynamic, Multidimensional Environment. *The Journal of Neuroscience*, *36*(38), 9843–58. https://doi.org/10.1523/JNEUROSCI.1337-16.2016

van Kesteren, M. T. R., Fernández, G., Norris, D. G., & Hermans, E. J. (2010). Persistent schema-dependent hippocampal-neocortical connectivity during memory encoding and postencoding rest in humans. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(16), 7550–7555. https://doi.org/10.1073/pnas.0914892107

Van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211–219. https://doi.org/10.1016/j.tins.2012.02.001

Wikenheiser, A. M., & Schoenbaum, G. (2016). Over the river, through the woods:

cognitive maps in the hippocampus and orbitofrontal cortex. *Nat Rev Neurosci*, *17*(8), 513–523. https://doi.org/10.1038/nrn.2016.56

Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, *81*(2), 267–278. https://doi.org/10.1016/j.neuron.2013.11.005

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and Ventral Medial Prefrontal Activation during Retrieval-Mediated Learning Supports Novel Inference. *Neuron*, *75*, 168–179. https://doi.org/10.1016/j.neuron.2012.05.010