Working memory load strengthens reward prediction errors.

Abbreviated title: Working memory load strengthens prediction errors

Authors: Anne G.E. Collins[1-3], Brittany Ciullo[3], Michael J Frank[3-4], David Badre[3-4]

Affiliations

1. Department of Psychology, University of California, Berkeley, Berkeley, CA 94720.
2. Helen Wills Neuroscience Institute, Berkeley, CA, 94720
3. Department of Cognitive, Linguistics and Psychological Sciences, Brown University, Providence RI 02912
4. Brown Institute for Brain Science, Providence, RI 02912.

Corresponding author:
Anne G.E. Collins
3210 Tolman Hall, Department of Psychology, UC Berkeley, Berkeley, CA 94720
annecollins@berkeley.edu,

Number of pages: 30
Number of figures: 6
Number of tables: 3
Abstract: 185 words
Introduction: 497 words
Discussion: 1487

The authors declare no competing financial interests.

Acknowledgements: We thank Christopher R Gagne for his role in data collection.

Abstract:

Reinforcement learning in simple instrumental tasks is usually modeled as a monolithic process in which reward prediction errors are used to update expected values of choice options. This modeling ignores the different contributions of different memory and decision-making systems thought to contribute even to simple learning. In an fMRI experiment, we asked how working memory and incremental reinforcement learning processes interact to guide human learning. Working memory load was manipulated by varying the number of stimuli to be learned across blocks. Behavioral results and computational modeling confirmed that learning was best explained as a mixture of two mechanisms: a fast, capacity-limited, and delay-sensitive working memory process together with slower reinforcement learning. Model-based analysis of fMRI data showed that striatum and lateral prefrontal cortex were sensitive to reward prediction error, as shown previously, but critically, these signals were reduced when the learning problem was within capacity of working memory. The degree of this neural interaction related to individual differences in the use of working memory to guide behavioral learning. These results indicate that the two systems do not process information independently, but rather interact during learning.

Significance Statement

Reinforcement learning theory has been remarkably productive at improving our understanding of instrumental learning as well as dopaminergic and striatal network function across many mammalian species. However, this neural network is only one contributor to human learning, and other mechanisms such as prefrontal cortex working memory, also play a key role. Our results show in addition that these other players interact with the dopaminergic RL system, interfering with its key computation of reward predictions errors.

Intro:

Reinforcement learning (RL) theory (Sutton & Barto 1998) proposes that we can learn the value associated with various choices by computing the discrepancy between the reward we obtain and our previously estimated value, and proportionally adjusting our estimate. This discrepancy, the reward prediction error (RPE), signals a valenced

surprise at the outcome being better or worse than expected and a direction to adapt behavior (Pessiglione et al. 2006; Schönberg et al. 2007; Daw & Doya 2006). In the brain, cortico-basal ganglia loops appear to implement a form of algorithmic RL: Dopamine-dependent plasticity in the striatum may reinforce selection of choices leading to positive RPEs and weaken those leading to negative RPEs (Frank et al. 2004; Collins & Frank n.d.). Dopaminergic neurons exhibit phasic changes in their spike rates that convey RPEs (Montague et al. 1996; Schultz 2002), and dopamine release in target regions provides a bidirectional RPE signal (Hart et al. 2014). Human imaging studies have indeed found that striatal BOLD correlates with RPE and is enhanced by DA manipulations (Pessiglione et al. 2006; Schönberg et al. 2007; Jocham et al. 2011).

However, other neurocognitive processes contribute to learning besides the integration of reward history by RL. Specifically, executive processes (such as those involved in representing sequential or hierarchical task structure) contribute substantially to human learning over and above incremental RL (Daw et al. 2011; Badre & Frank 2011; Botvinick et al. 2009; Collins & Koechlin 2012; Collins & Frank 2013). Even in basic stimulus-response learning tasks, working memory (WM) contributes substantially to instrumental learning beyond RL (Collins & Frank 2012; Collins et al. 2014), as evidenced by both behavioral analyses and quantitative computational model fits. Two effects of WM were evident in learning. As WM set size increased (working memory load), learning curves per stimulus were slowed. Second, accuracy per trial declined as a function of the number of intervening items (working memory delay). These WM effects decayed with further experience, as the more reliable but slower RL process gained control of behavior. A hybrid model of WM and RL provided a better fit to these data than either process itself (Collins & Frank 2012; Collins et al. 2014),.

This prior behavioral work implies that WM contributes to RL processes. Here, we investigate the neural markers of learning and RPEs to determine whether they are interact with WM. While many RL studies have revealed neural correlates of RPEs that relate to learning, these studies have not manipulated or estimated WM factors that could contribute to (and potentially confound) these signals. Identifying separate markers of systems that contribute jointly to behavior also provides an opportunity to explore whether they interact (e.g., competitively or cooperatively). Specifically, we tested whether frontoparietal networks associated with cognitive control and striatal systems

associated with RL would show parametric modulations of RPE signaling as a function of WM load during learning. We also tested whether such interactions would be predictive of the extent to which individuals relied on WM contributions to RL behaviorally.

Methods:

Participants:

We scanned 26 participants (ages 18-31, mean age 23, 15 males/11 females). All 26 participants are included in the behavioral analysis. 5 participants were excluded from fMRI analysis prior to analyzing their fMRI data due to head movement greater than our voxel size. 2-6 blocks were excluded from 3 other participants due to movement during data collection towards the end of the scan. All participants were right-handed with normal or corrected-to-normal vision and were screened for the presence of psychiatric or neurological conditions and contraindications for fMRI. All participants were compensated for their participation and gave informed, written consent as approved by the Human Research Protection Office of Brown University.

Experimental design:

The task (Fig1) was similar to that described previously (Collins & Frank 2012; Collins et al. 2014),, itself adapted from a classic Conditional Associative Learning paradigm (Petrides 1985). On each trial, subjects had to respond with one of three responses (button presses on a response pad) when presented with a centrally displayed single stimulus. Subjects had to learn over trials which response was correct for each stimulus, based on binary deterministic reinforcement feedback (Collins & Frank 2012; Collins et al. 2014),.

To manipulate working memory demands separately from RL components, we systematically varied the number of stimuli (denoted as set size n_s) to be learned within a block. Larger set sizes provide greater load on working memory, and also impose on average larger delays between repetitions of the same stimulus. Subjects experienced 3 blocks of each of the set-sizes one through six. In each block, subjects learned about a different category of visual stimulus (such as sports, fruits, places, etc.), with stimulus

category assignment to block set size counterbalanced across subjects. Block ordering was also counterbalanced within subjects to ensure an even distribution of high/low load blocks across each third of the experiment.

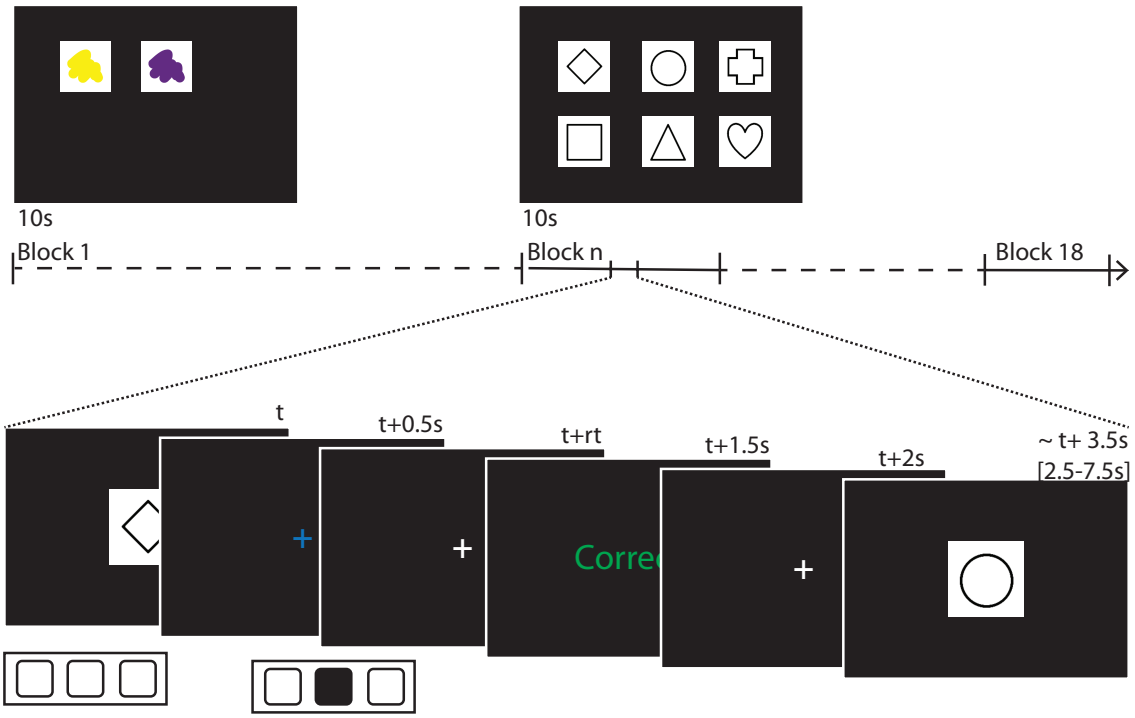

Figure 1: Experimental Protocol. At the beginning of each block, subjects were shown for 10 s the set of stimuli they would see in that block. In this example, Block 1 uses color patches for stimuli and has a set size n_s *=2*; Block n uses shapes and has n_s *=6.* Each trial included the presentation of a stimulus for 0.5s followed by a blue fixation cross until subject pressed one of three buttons, or up to 1.5s after trial onset. Button press caused the fixation cross to turn white. Feedback was presented for 1s, and came 1.5s after trial onset. Feedback consisted of the words correct or incorrect in green and red, respectively. The inter-trial interval consisted of a white fixation cross with jittered duration to allow trial by trial event-related analysis of fMRI signal. Blocks set sizes varied between 1 and 6, and the order was randomized across subjects.

At the beginning of each block, subjects were shown the entire set of stimuli for that block and were encouraged to familiarize themselves with them for a duration of 10 sec (figure 1 top). They were then asked to make their response as quickly and accurately as possible after each individual stimulus presentation. Within each block, stimuli were presented 12 times each in a pseudo-randomly intermixed order.

Stimuli were presented in the center of the screen for up to 0.5s seconds, followed by a blue fixation cross for up to 1s or subjects making a choice by pressing one of 3 buttons, at which time the fixation cross turned white (figure 1 bottom). Feedback was presented 1.5s after stimulus onset for 0.5s as either “Correct” in green, “Incorrect” in red, or “Too

slow" if the subject failed to answer within 1.5s. A white fixation cross followed with jittered duration of mean 1.5s, ranging [.5 6.5]s, before the next stimulus was presented.

Subjects were instructed that finding the correct action for one stimulus was not informative about the correct action for another stimulus. This was enforced in the choice of correct actions, such that, in a block with e.g., n_S=3, the correct actions for the three stimuli were not necessarily three distinct keys. This procedure was implemented to ensure independent learning of all stimuli (i.e., to prevent subjects from inferring the correct actions to stimuli based on knowing the actions for other stimuli). Prior to entering the scanner, subjects went through the instructions and practiced on a separate set-size 2 sets of images to ensure they were familiarized with the task.

Computational model:

RLWM model:

To better account for subjects' behavior and disentangle roles of working memory and reinforcement learning, we fitted subjects' choices with our hybrid RLWM computational model. Previous research showed that this model, allowing choice to be a mixture between a classic delta rule reinforcement learning process and a fast but capacity-limited and delay-sensitive working memory process, provided a better quantitative fit to learning data than models of either WM or RL alone (Collins & Frank 2012; Collins et al. 2014),. The model used here is a variant of the previously published models. We first summarize its key properties, following by the details:

- RLWM includes two modules which separately learn the value of stimulus-response mappings: a standard incremental procedural RL module with learning rate α, and a WM module that updates S-R-O associations in a single trial (learning rate 1) but is capacity-limited (with capacity K).
- The final action choice is determined as a weighted average over the two modules' policies. How much weight is given to WM relative to RL (the mixture parameter) is dynamic and reflects the probability that a subject would use WM vs. RL in guiding their choice. This weight depends on two factors. First, a *constraint* factor reflects the *a priori* probability that the item is stored in WM, which depends on set size n_S of the current block relative to capacity *K* (i.e., if n_S>*K*, the probability that an item is stored is *K*/ns), scaled by the subject's overall reliance of WM vs. RL (factor $0<\rho<1$), with higher values reflecting relative greater confidence in WM function. Thus, the

constraint factors indicates that the maximal use of WM policy relative to RL policy is $w_0 = \rho$ x min(1, K/n_S). Second, a *strategic* factor reflects the inferred reliability of the WM compared to RL modules over time: initially, the WM module is more successful at predicting outcomes than the RL module, but because it has higher capacity and less vulnerability to delay, the RL module becomes more reliable with experience.

- Both RL and WM modules are subject to forgetting (decay parameters ϕ_{RL} and ϕ_{WM}). We constrain $\phi_{RL} < \phi_{WM}$ consistent with WM's dependence on active memory).

Learning model details.

Reinforcement learning model: All models include a standard RL module with simple delta rule learning. For each stimulus *s*, and action *a*, the expected reward *Q(s,a)* is learned as a function of reinforcement history. Specifically, the *Q* value for the selected action given the stimulus is updated upon observing each trial's reward outcome r_t (1 for correct, 0 for incorrect) as a function of the prediction error between expected and observed reward at trial *t*:

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha \, x \, \delta_t,$$

where $\delta_t = r_t - Q_t(s,a)$ is the prediction error, and α is the learning rate. Choices are generated probabilistically with greater likelihood of selecting actions that have higher *Q* values, using the softmax choice rule:

$$p(a|s)=exp(\beta Q(s,a))/ \Sigma_i(exp(\beta Q(s,a_i)).$$

Here, β is an inverse temperature determining the degree with which differences in *Q*-values are translated into more deterministic choice, and the sum is over the three possible actions a_i.

Undirected noise. The softmax temperature allows for stochasticity in choice, but where stochasticity is more impactful when the value of actions are similar to each other. We also allow for "slips" of action ("irreducible noise", i.e., even when Q value differences are large). Given a model's policy π = p(a|s), adding undirected noise consists in defining the new mixture policy:

$$\pi' = (1- \varepsilon) \pi + \varepsilon U,$$

where U is the uniform random policy (U(a) = 1/n_A, n_A=3), and the parameter $0<\varepsilon<1$ controls the amount of noise (Collins & Koechlin 2012; Collins & Frank 2013; Guitart-

Masip et al. 2012). (Nassar & Frank 2016) showed that failing to take into account this irreducible noise can render fits to be unduly influenced by rare odd datapoints (e.g. that might arise from attentional lapses), and that this problem is remedied by using a hybrid softmax-ε–greedy choice function as used here.

Forgetting. We allow for potential decay or forgetting in Q-values on each trial, additionally updating all Q-values at each trial, according to:

$$Q \leftarrow Q + \phi\,(Q_0 - Q),$$

where $0<\phi<1$ is a decay parameter pulling at each trial the estimates of values towards initial value $Q_0 = 1/n_A$. This parameter allows us to capture delay-sensitive aspects of WM, where active maintenance is increasingly likely to fail with intervening time and other stimuli, but also allows us to separately estimate any decay in RL values (which is typically substantially lower than in WM).

Perseveration. To allow for potential neglect of negative, as opposed to positive feedback, we estimate a perseveration parameter *pers* such that for negative prediction errors (delta<0), the learning rate α is reduced by $\alpha = (1\text{-}pers) \times \alpha$. Thus, values of *pers* near 1 indicate perseveration with complete neglect of negative feedback, whereas values near 0 indicate equal learning from negative and positive feedback.

Working Memory. To implement an approximation of a rapid updating but capacity-limited WM, this module assumes a learning rate $\alpha = 1$ (representing the immediate accessibility of items in active memory), but includes capacity limitation such that only at most K stimuli can be remembered. At any trial, the probability of working memory contributing to the choice for a given stimulus is $w_{WM}(t) = P_t(WM)$. This value is dynamic as a function of experience (see next paragraph). As such, the overall policy is:

$$\pi = w_{WM}(t)\pi_{WM} + (1 - w_{WM}(t))\,\pi_{RL}$$

where π_{WM} is the WM softmax policy, and π_{RL} is the RL policy. Note that this implementation assumes that information stored for each stimulus in working memory pertains to action-outcome associations. Furthermore, this implementation is an approximation of a capacity/resource-limited notion of working memory. It captures key aspects of working memory such as 1) rapid and accurate encoding of information when low amount of information is to be stored; 2) decrease in the likelihood of storing or maintaining items when more information is presented or when distractors are presented

during the maintenance period; 3) decay due to forgetting. Because it is a probabilistic model of WM, it cannot capture specifically which items are stored, but it can provide the likelihood of any item being accessible during choice given the task structure and recent history (set size, delay, etc).

Inference: The weighting of whether to rely more on WM vs. RL is dynamically adjusted over trials within a block based on which module is more likely to predict correct outcomes. The initial probability of using WM $w_{WM}(0) = P_0(WM)$ is initialized by the *a priori* use of WM, as defined above, $w_{WM}(0) = \rho \times min(1, K/n_S)$, where ρ is a free parameter representing the participant's overall reliance on WM over RL.
On each correct trial, $w_{WM}(t)=P_t(WM)$ is updated based on the relative likelihood that each module would have predicted the observed outcome given the selected correct action a_c; specifically:

- for WM, $p(correct|stim, WM) = w_{WM}\, \pi_{WM}(a_c) + (1-w_{WM})1/n_A$
- for RL, $p(correct|stim, RL)$ this is simply $\pi_{RL}(a_c)$

The mixture weight is updated by computing the posterior using the previous trial's prior, and the above likelihoods, such that

$$P_{t+1}(WM) = \frac{P_t(WM) \times p(correct|stim, WM)}{P_t(WM) \times p(correct|stim, WM) + P_t(RL) \times p(correct|stim, RL)}$$

and $P_{t+1}(RL)=1-P_{t+1}(WM)$.

Models Considered. We combined the previously described features into different learning models and conducted extensive comparisons of multiple models to determine which fit the data best (penalizing for complexity) so as to validate the use of this model in interpreting subjects' data. For all models we considered, adding undirected noise, forgetting and perseveration features significantly improved the fit, accounting for added model complexity (see model comparisons).

This left three relevant classes of models to consider:

- RL: This model combines the basic delta rule RL with forgetting, perseveration and undirected noise features. It assumes a single system that is sensitive to delay and asymmetry in feedback processing. This is a 5-parameter model

(learning rate α, sofmax inverse temperature β, undirected noise ε, decay ϕ_{RL}, and *pers* parameter).

- RL6: This model is identical to the previous one, with the variant that learning rate can vary as a function of set-size. We have previously shown that while such a model can capture the basic differences in learning curves across set-sizes by fiting lower learning rates with higher set sizes, it provides no mechanism that would explain these effects, and still cannot capture other more nuanced effects (e.g. changes in the sensitivity to delay with experience). However it provides a benchmark to compare with RLWM. This is a 10-parameter model (6 learning rate α_{ns}, sofmax inverse temperature β, undirected noise ε, decay ϕ_{RL}, and *pers* parameter).
- RLWM: This is the main model, consisting of a hybrid between RL and WM. RL and WM modules have shared softmax β and *pers* parameters, but separate decay parameters, ϕ_{RL} and ϕ_{WM}, to capture their differential sensitivity to delay. Working memory capacity is $0<K<6$, with an additional parameter for overall reliance on working memory $0<\rho<1$. Undirected noise is added to the RLWM mixture policy. This is an 8-parameter model (capacity K, WM reliance ρ, WM decay ϕ_{WM}, RL learning rate α, RL decay ϕ_{RL},softmax inverse temperature β, undirected noise ε, and *pers* parameter).

In the RLWM model presented here, the RL and WM modules are independent, and only compete for choice at the policy level. Given our findings showing an interaction between the two processes, we also considered variants of RLWM including mechanisms for interactions between the two processes at the learning stage. These models provided similar fit (measured by AIC) to the simpler RLWM model. We chose to use the simpler RLWM model, because the more complex model is less identifiable within this experimental design, providing less reliable parameter estimates and regressors for model-based analysis.

RLWM fitting procedure: We used matlab optimization under constraint function `fmincon` to fit parameters. This was iterated with 50 randomly chosen starting points, to increase likelihood of finding a global rather than local optimum. For models including

the discrete capacity *K* parameter, this fitting was performed iteratively for capacities K = {1,2,3,4,5}, using the value gave the best fit in combination with other parameters.

Softmax β temperature was fit with constraints [0 100]. All other parameters were fit with constraints [0 1]. We considered sigmoid-transforming the parameters to avoid constraints in optimization and obtain normal distributions, but while fit results were similar, distributions obtained were actually not normal. Thus, all statistical tests on parameters were non-parametric. See table 4 for fit parameter statistics.

Other competing models:

In order to further test whether "single system" models, as opposed to hybrid models including an RL and a WM component, could account for behavior, we tested other algorithms embodying alternative assumptions in which behavior is governed by a single learning process (either RL or WM).

- The WMd model is similar to a WM module, with the following changes. A) there is no capacity limitation. B) Instead of being fixed, the decay parameter is fixed to an initial value which then decreases toward 0 with each stimulus encounter, modeling the possibility that forgetting in WM itself might decrease with practice. This model includes 5 parameters: β, ε and *pers* as defined above, the initial value of decay $decay_0$, and ξ the decay factor.
- The WMdi model adds an interference mechanism to WMd, such that the decay factor of a given stimulus additionally increases with every encounter of a different stimulus. This adds one parameter to the previous model.
- The RLi model is identical to the basic RL model, with an added interference mechanism: on each trial, the Q-value of non-observed stimuli with the chosen action is updated in the same way as the observed stimuli, but with a fraction of the learning rate αi. This captures the possibility that credit is assigned to the wrong stimulus, modeling the possibility that WM-like effects might reflect interference within a pure RL system. This model includes 6 parameters.

Model Comparison: We used the Akaike Information Criterion to penalize model complexity - AIC (Burnham & Anderson 2002). Indeed, we previously showed that in the case of the RLWM model and its variants, AIC was a better approximation than Bayesian Information Criterion (BIC; Schwarz 1978) at recovering the true model from

generative simulations (Collins & Frank 2012). Comparing RLWM, RL6 and RL-only showed that models RL6 and RL-only were strongly non-favored, with probability 0 over the whole group. Other single process models were also unable to capture behavior better than RLWM (Fig. 3 E).

Model Simulation: Model selection alone is insufficient to assess whether the best fitting model sufficiently captures the data. To test whether models capture the key features of the behavior (e.g., learning curves), we simulated each model with fit parameters for each subject, with 100 repetitions per subject then averaged to represent this subject's contribution. In order to account for initial biases, we assume that the model's choice at first encounter of a stimulus is identical to the subjects, while all further choices are randomly selected from the model's learned values and policies.

fMRI recording and preprocessing:

Whole-brain imaging was performed on a Siemens 3T TIM Trio MRI system equipped with a 32-channel head coil. A high-resolution T1-weighted 3D multi-echo MPRAGE image was collected from each participant for anatomical visualization. Functional images were acquired in one run of 1,920 volume acquisitions using a gradient-echo, echo planar pulse sequence (TR 2 s, TE 28 ms, flip angle 90, 40 interleaved axial slices, 192 mm field of view with 3x3x3 mm voxel size). Stimuli were presented on a BOLD screen display device (http://www.crsltd.com/tools-for-functional-imaging/mr-safe-displays/boldscreen-24-lcd-for-fmri/) located behind the scanner and made visible to the participant via an angled mirror attached to the head coil. Padding around the head was used to restrict head motion. Participants made their responses using an MRI-compatible button box.

Functional images were preprocessed in SPM8 (http://www.fil.ion.ucl.ac.uk/spm). Before preprocessing, data were inspected for artifacts and excessive variance in global signal (functions: tsdiffana, art_global, art_movie). Functional data were corrected for differences in slice acquisition timing by resampling slices to match the first slice. Next, functional data were realigned (corrected for motion) using B-spline interpolation and referenced to the mean functional image. Functional and structural images were normalized to Montreal Neurological Institute (MNI) stereotaxic space using affine regularization followed by a nonlinear transformation based on a cosine basis set, and

then resampled into 2x2x2 mm voxels using trilinear interpolation. Lastly, images were spatially smoothed with an 8 mm full-width at half-maximum isotropic Gaussian kernel.

GLMs:

A temporal high-pass filter of 400 seconds (.0025 Hz) was applied to our functional data in order to remove noise but preserve power from low-frequency regressors. Changes in MR signal were modeled using a general linear model (GLM) approach. Our GLM included six onsets regressors, one for correct trials corresponding to each set size (1-6). Each onset was coded as a boxcar of 2 seconds in length that encompasses stimulus presentation, response, and feedback. Each onset regressor was modulated by a Prediction Error parametric regressor. We modeled Error trials, No Response trials, and instructions (1 instruction screen at the beginning of each block, 18 total and each 10 seconds in length) as separate regressors. Note that error trials across all set sizes were binned into one regressor due to the low number of error trials in low set sizes. Finally, we included nuisance regressors for the six motion parameters (x, y, z, pitch, roll, yaw) and a linear drift over the course of the run. SPM-generated regressors were created by convolving onset boxcars and parametric functions with the canonical hemodynamic response (HRF) function and the temporal derivative of the HRF. Beta weights for each regressor were estimated in a first-level, subject-specific fixed-effects model. For group analysis, the subject-specific beta estimates were analyzed with subject treated as a random effect. At each voxel, a one-sample t-test against a contrast value of zero gave us our estimate of statistical reliability. For whole brain analysis, we corrected for multiple comparison using cluster correction, with a cluster forming threshold of $p<.001$ and an extent threshold calculated with SPM to set a family-wise error cluster level corrected threshold of $p < .05$ (127 for PE>fixation; 267 for PE*set size interaction). Note that these appropriately high cluster forming threshold ensures that parametric assumptions are valid and the rate of false positives are appropriate (Eklund et al. 2016; Flandin & Friston 2016).

ROIs:

Fronto-parietal network: As we did not have specific regional predictions regarding the WM component of learning, we defined broad fronto-parietal networks as ROIs that have been previously associated with a wide range of tasks involving cognitive control. Specifically, our first control network ROIs were defined by using left and right anterior

dorsal premotor cortex (prePMd: 8mm sphere around -38 10 34, (Badre & D'Esposito 2007) as seeds in two separate "resting state" (task-free) seed-to-voxel correlation analyes in the CONN toolbox (https://www.nitrc.org/projects/conn/), and using the corresponding whole-brain connectivity to left and right prePMd, as our control network ROI. In order to confirm the robustness of our findings, we then ran a larger frontoparietal network ROI defined from a functionally neutral group (Yeo et al. 2011), along with a functionally defined ROI of the multiple demands network from (Fedorenko et al. 2013). All three of these frontoparietal ROIs yielded similar outcomes, thus confirming the robustness of our findings. We report here the results from the Yeo et al ROI as the widest, most neutral ROI.

The striatum ROI was defined based on univariate activity for prediction error ($p < .001$, uncorrected), masked by AAL definitions for putamen, caudate, and nucleus accumbens (Marsbar AAL structural ROIs: http://marsbar.sourceforge.net/download.html). We note that this ROI definition would be biased for assessing the effect of RPE in the striatum. However, this is not our goal as the relationship of RPE and striatum is established both in general from the prior literature, and in this study based on the corrected whole brain analysis (see Results). Rather, this ROI will be used to test the effects of set size and the interaction of set size with RPE, within regions maximally sensitive to RPE. As the set size variable is uncorrelated with that of RPE, this ROI definition does not bias either of these analyses.

For each ROI, a mean time course was extracted using the MarsBar toolbox (http://marsbar.sourceforge.net/). The GLM design was estimated against this mean time series, yielding parameter estimates (beta weights) for the entire ROI for each regressor in the design matrix.

Whole brain Contrasts: We focus on two main contrasts: 1) positive effect of RPE; 2) positive interaction of RPE and set size, to determine whether WM processes influence RPE signaling and whether such interactions relate to behavior. The first contrast is defined by considering the sum of the beta weights across all set sizes: $\Sigma_{i=1:6}\ \beta_{PE(i)}$; we test whether this contrast value is significantly positive. The second contrast takes the linear contrast of the beta weights across set sizes by the set size: $\Sigma_{i=1:6}\ (i-3.5)*\beta_{PE(i)}$; testing whether this contrast is positive signals a linear increase of RPE with set size.

We also tested the opposite contrasts, as well as the linear effect of set size $\Sigma_{i=1:6}$ (i-3.5)*β_i

Interaction between set-size and RPE: To investigate individual differences in the interaction between set size and RPE, we assessed ROI markers of this interaction. We computed this in one of three ways, each reflecting different assumptions: (A) a linear contrast of set-size on RPE regression weight; (B) a contrast of high set size (4-6) vs. low set size (1-3) on RPE regression weights (in case of a step function for e.g. above vs. below capacity sets), and (C) Spearman rho of RPE weights across set-sizes, which does not require linearity and is less susceptible to outliers than linear regression. Despite slightly different assumptions, all three measures are highly correlated (all rhos>0.8,p<10-4) and yielded qualitatively similar results. Because we observe that results neither show linear changes across set sizes, nor a step function, we report results using the measure defined as option C.

Results:

Behavior:

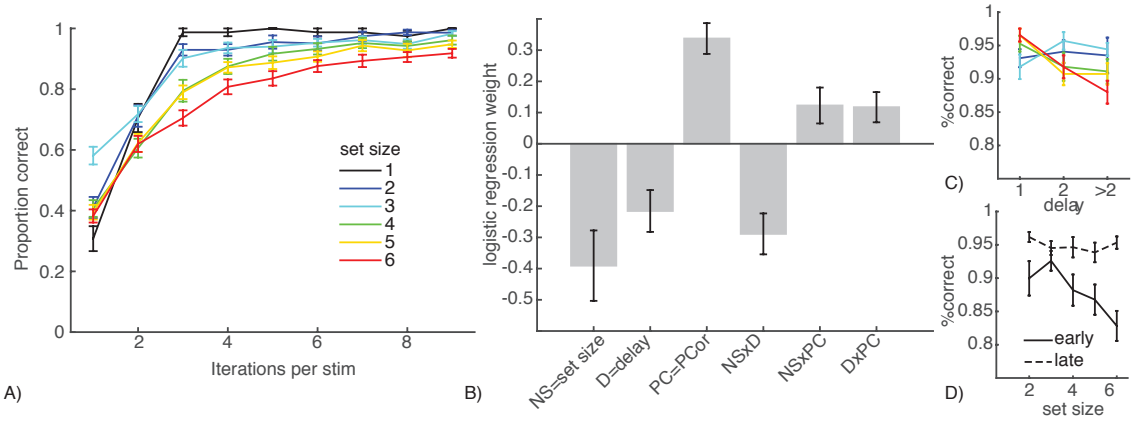

Figure 2: Behavioral Results. A) Proportion of correct choices as a function of how many times a specific stimulus was encountered (i.e., learning curves), for each set size. B) Logistic regression on factors that contribute to accuracy for a given image, including set size (NS), delay since last previous correct choice for a given image (D), PCor (number of previous correct choices for that image), and their interactions. C) Illustration of the interaction between delay and set size. D) Illustration of the interaction between set size and PCor – early indicates PCor<4, late indicate PCor>6. Error bars indicate standard error of the mean.

Behavioral results replicate our previous findings (Collins & Frank 2012; Collins et al. 2014; Figure 2). Learning curves showed strong differences as a function of set size, despite the same number of encounters for each stimulus. Logistic regression analysis of subjects choices (Fig. 2B) showed main effects of reward history, delay and load, indicating that subjects were more likely to select the correct action with more previous correct experience for a given stimulus ($t(25)=6.8$, $p<10\text{-}4$), and less likely to be correct with increasing set size ($t(25)=-3.4$, $p=.002$) and increasing delay (intervening trials since their last correct choice on this stimulus) ($t(25)=-3.2$, $p=.004$). There were also interactions between all pairs of factors, such that the delay effect was stronger in high load ($t(25)=-4.4$,$p=.0002$, Fig. 2C), and the effects of load and delay both decreased with more correct reward history ($ts>2.1$, $ps<.05$, Fig. 2D). The latter interaction is expected given the RLWM model's prediction that behavior transitions from WM (which is more sensitive to delay and load) to RL as a function of learned reliability.

Model fitting:

Model fitting also confirmed our previous findings, showing that a computational model including two modules (RL and WM) explained subjects' behavior better than variants of a model assuming a single RL or WM process. Specifically, RLWM provided a significantly better AIC than RL6 ($t(25)=3.9$, $p=0.001$) and RL ($t(25)=-6.6$,$p<10^{-4}$), and individual AICs favored RLWM for a significant number of subjects (21/26 for RL6, sign test $p=0.002$; 23/26 for RL, $p<10^{-4}$). Model simulations show that a simple RL model cannot capture the behavior as well as RLWM or RL6, but note that RL6 needs too many parameters to appropriately capture behavior. Pure working memory models assuming changes in decay with experience, or interference, also cannot capture behavior as well as our hybrid RLWM model (Fig. 3E)

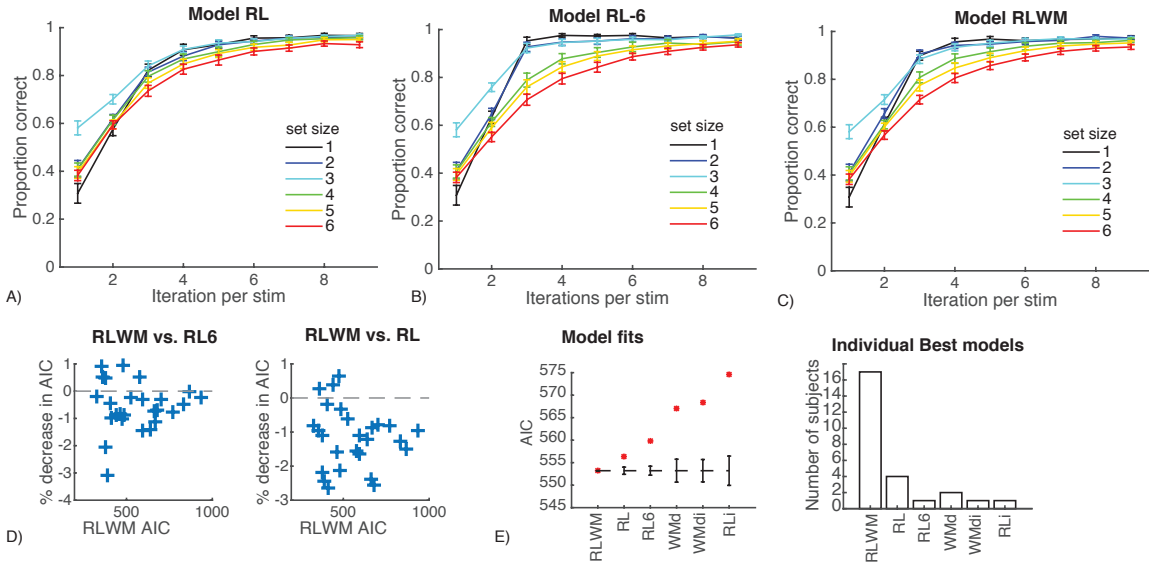

Figure 3: Model Validation. A-C) Proportion of correct responses as a function of how many times a specific stimulus was encountered, for each set size, for simulation of different models with individually fit parameters. Models were simulated 100 times per subject then averaged within subjects to represent this subject's contribution. Error bars indicate standard error of the mean across subjects. A) simple RL model including decay and different sensitivity to gains/losses. B) Identical model to A, with learning rate varying per set size. C) Model incorporating both RL and WM. D) Model comparisons show a significantly lower AIC for RLWM than RL6 or RL, for a significant number of subjects. Each cross indicates a single subject. E) Model comparison to other potential models show best fit for RLWM (see methods for other model names).

Imaging results:

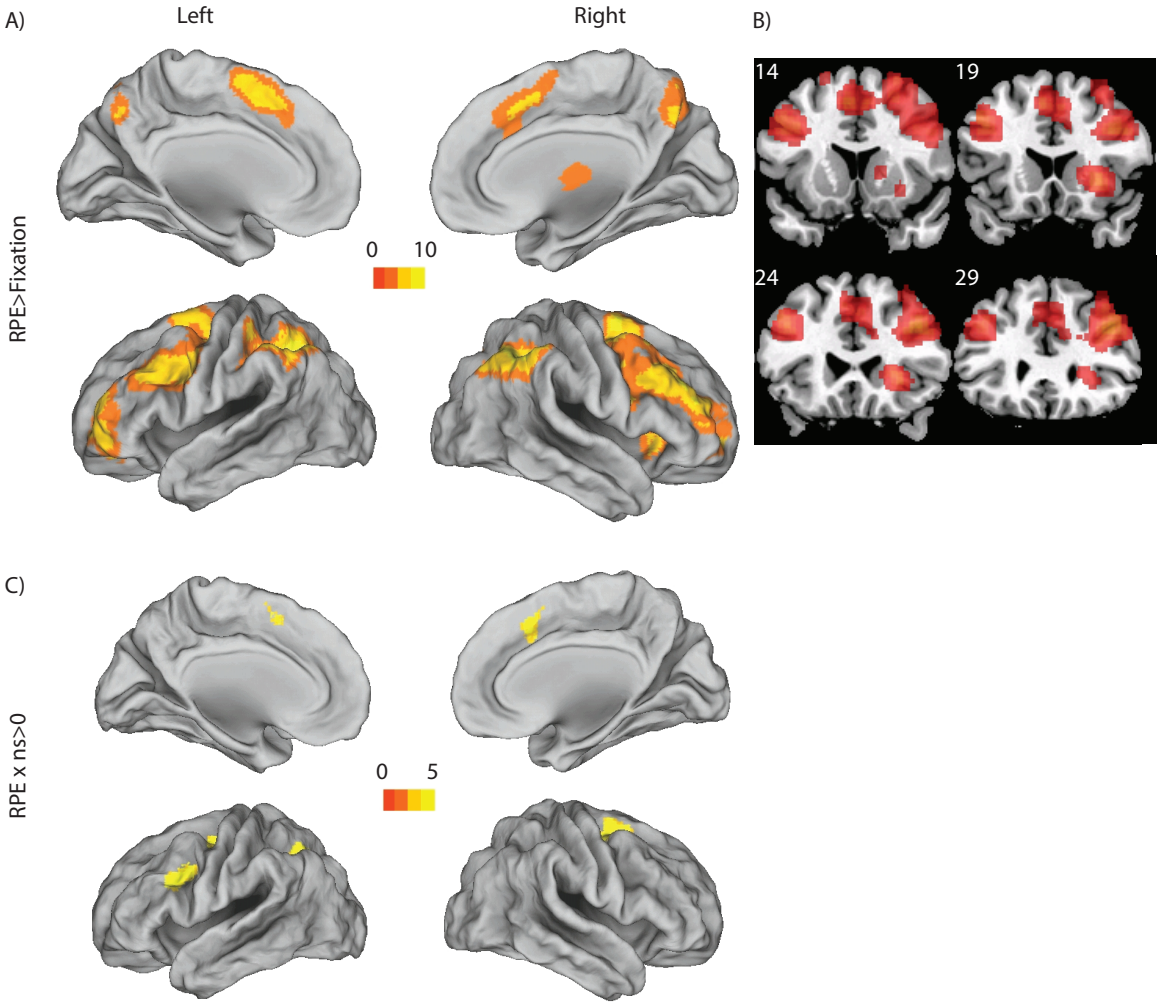

Figure 4: Whole brain effects of RPE and RPExns. A-B) Regions positively correlated with RPE (p<.05 cluster corrected). C) Regions showing a positive interaction of RPE with set size.

Whole brain analysis showed increasing activity with set size in bilateral precuneus and decreasing activity in a network including bilateral superior frontal gyrus, bilateral angular gyrus and bilateral supramarginal gyrus (table 3), confirming that the set size manipulation is effective at differentially engaging large brain networks.

Whole brain analysis showed a distributed network that positively correlated with the parametric reward prediction errors (RPE) regressor. We verified RPE-related activation in the right caudate nucleus and thalamus (See table 1 for full results, figure 4B), as expected from the literature. Notably, the RPE network also includes regions of bilateral prefrontal and parietal cortex commonly observed in cognitive control tasks.

We next tested whether the RPE signal was homogeneous across set sizes in striatum, as implicitly expected if striatal RL is independent of WM. To the contrary. we found a significant positive interaction of set size with RPE (t(20)=2.4, p=0.026; figure 5B) in the striatal ROI (see Methods). Note that this interaction reflects a *stronger* effect of RPE on the striatal BOLD signal at higher set-sizes (i.e., under more cognitive load). This finding supports the hypothesis that WM interacts with RL, showing blunted RL signals in low set sizes (i.e., within the capacity of WM).

Next, we investigated whether other brain regions showed the same modulation of RPE signaling by WM load. Whole brain analysis showed a positive linear interaction of set size with RPE in left lateral prefrontal cortex and parietal cortex (MNI coordinates -38, 20, 28; table 2). Further investigation within an independent fronto-parietal network ROI (Yeo et al. 2011) showed both a strong main effect of prediction error (t(20)=6.9,p<10-4) and a significant interaction of set-size with RPE in the fronto-parietal ROI (t(20)=2.3,p=0.03), a pattern similar to the striatum ROI. Again, RPE signaling was larger with more WM load, possibly reflective of a common neuromodulatory signal in striatum and cortex influenced by cognitive demands.

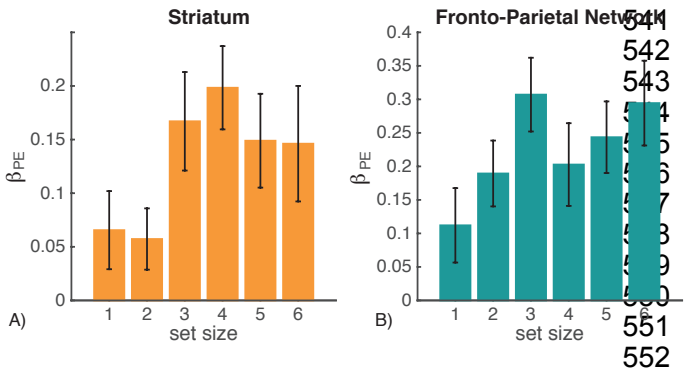

Figure 5: Striatum and Fronto-parietal ROIs show increased RPE effects in higher set sizes. Average beta coefficient for RPE regressor per set size for A) Striatal ROI and B) Fronto-parietal Network ROI defined by Yeo et al. Error bars indicate standard error of the mean.

Link to behavior:

We hypothesized that the weaker RPE signals observed in low set-sizes might reflect an interaction between WM and RL systems. Specifically, this may reflect the greater use of WM, instead of RL, at low set sizes. This strategy could be because low set sizes do not require RPE signaling: the most recent stimulus-action-outcome can be accessed from memory. Thus, we predicted that those subjects relying more on WM would exhibit a stronger neural interaction effect (i.e., they would show less homogeneity in their RPE signals across set-sizes). To index WM contributions to choice, we use the

computational model-inferred weight of the WM module, averaged over all trials. Indeed, we found that greater WM contributions to choices was significantly related to the set-size effect on RPE signaling, both in striatum (rho=0.55, p=.01), and the fronto-parietal ROI (rho=0.49; p=0.02; figure 6 left). Moreover, subjects who continued to rely on WM with experience (i.e., exhibiting less transition to RL) also showed greater set-size effects on RPE signaling in FP (rho=-0.46, p=0.03) and marginally, in striatum (rho=-0.41; p=0.06; figure 6 middle). This may be due to the fact that for participants with higher overall reliance on WM, the WM module is more reliable, and thus WM use decreases less over learning. Indeed, the two indexes were negatively correlated (rho=-0.69, p<10-3). The results were partly accounted for by differences in model fit capacity parameter: subjects with higher capacity showed significantly stronger nsxRPE interaction in FP (rho=0.46, p=0.03), and marginally so in striatum (rho=.41, p=0.06;). Finally, we confirmed this effect was independent of the fit of the RLWM model by using the logistic regression, and specifically the effect of set-size on accuracy (note that this measure was, as expected, related to the one obtained by the computational model: Spearman rho=-.42, p=0.05). Indeed, the effect of set-size on accuracy was marginally related to the set-size by RPE interaction in striatum (rho=-0.4,p=0.06, figure 6 right) and FP (rho=-0.41,p=0.06). Again, neural interactions were stronger for those subjects exhibiting a stronger negative effect of set-size on behavior.

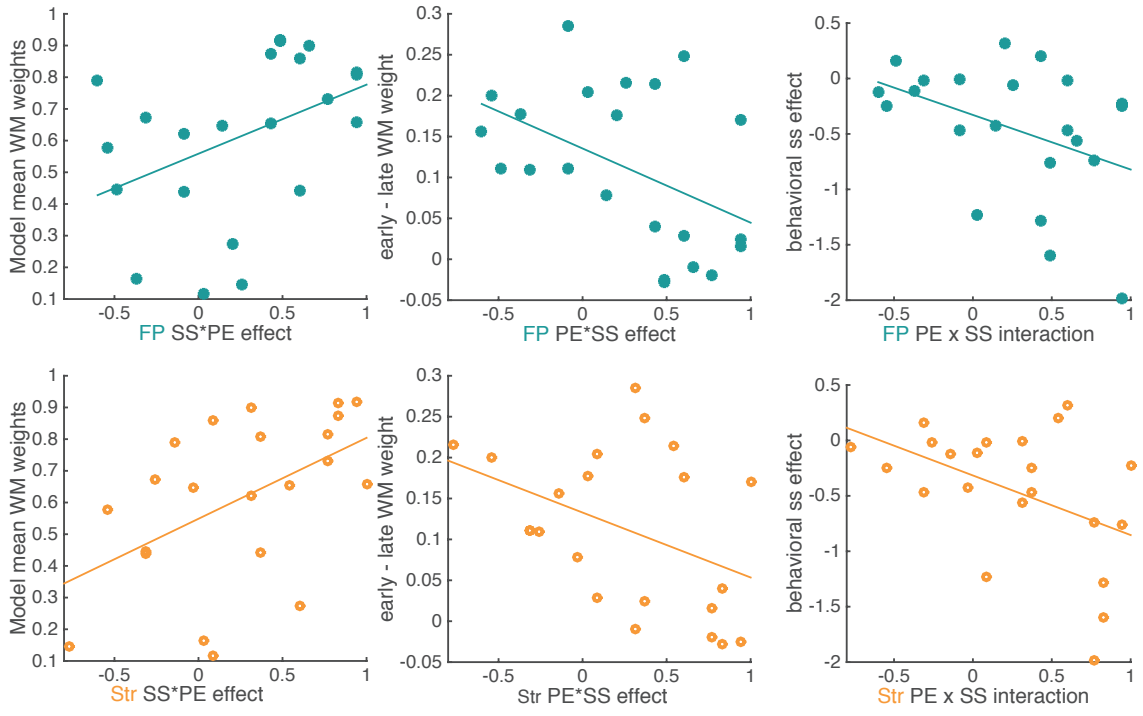

Figure 6: Effect of set size on RPE in the fMRI signal is related to individual differences in behavior. Left: average model-inferred mixture weight assigned to working memory over RL ("Model mean WM weight") is significantly related to a stronger effect of set size in fronto-parietal ROI (ρ=0.49, p=0.02) and in the striatum (ρ=0.55, p=0.01). Middle: decrease in working memory weight from early (first 3 iterations) to late in a learning block (last 3 iterations) is significantly related to fMRI effect in FP ROI (ρ=-0.46, p=0.03), and marginally so in striatum (ρ=-0.41, p=0.06). Right: the behavioral set-size effect is measured as the logistic regression weight of the set-size predictor; stronger behavioral effect is marginally related to a stronger neural effect in FP ROI (ρ=-0.41, p=0.059) and in striatum ROI (ρ=0.4, p=0.063).

Discussion:

We combined computational modeling and fMRI to investigate the contributions of two distinct processes to human learning: reinforcement learning and working memory. We replicated our previous results (Collins & Frank 2012; Collins et al. 2014) showing that these jointly play a role in decisions: computational models assuming a single learning process (either WM or RL) could not capture behavior adequately. We also replicated the widespread observation that the striatum and lateral prefrontal cortex are sensitive to reward prediction errors, a marker of RL.We made the novel observation that RL and WM are not independent processes, with the most commonly studied RL signal blunted under low WM load. Further, we found that the degree of interaction was related to individual differences in subjects' use of WM: the more robustly subjects used WM for learning, the more they showed WM effects on RL signals.

The process of model-free reinforcement learning, as both a class of machine learning algorithms and as the neural network function implemented via dopamine-dependent plasticity in cortico-basal ganglia networks, is characterized by integration of rewards over time to estimate the value of different options, and a value dependent policy. Our behavioral results replicate our previous work showing that even in simple instrumental learning, we cannot account for human learning based only on the integrated history of reward. Instead, the influences of load and delay/intervening trials show that working memory also contributes to learning. That this influence decreases with experience supports a model where RL and WM modules are dynamically weighted according to their success in predicting observed outcomes.

We used computational modeling to disentangle the contributions of RL and WM to learning and to assess neural indicators of their interactions. We extracted the reward prediction error signal from the RL module, and confirmed in a model-based whole brain

fMRI analysis that striatum was sensitive to prediction errors, as established from a large literature (Pessiglione et al. 2006; Schönberg et al. 2007), as was a large bilateral fronto-parietal region (Daw et al. 2011). However, we found in both regions that sensitivity to RPE was modulated by set size, the number of items that subjects learned about in a given block. Specifically, the RPE signal was weaker in lower set sizes, in which subjects' learning was closest to optimal, and thus likely to mostly use WM. Thus, as noted in our earlier studies (Collins & Frank 2012; Collins et al. 2014), WM contributions to learning can confound measures typically attributed to RL. While the previous findings were limited to behavioral, genetic and computational model parameters, here we report for the first time that even neural RPE signals are influenced by WM. These results also imply that in other studies that do not manipulate WM load during learning, the contribution of WM to learning may yield inflated or blunted estimates of the pure RL process.

We further found that individual differences in the degree to which set-size modulated RPE signals correlated with the degree to which subjects relied on WM in their behavioral learning curves. Specifically, subjects with more robust use of WM showed more reliably blunted RPE signals in lower set sizes, supporting the interpretation that WM use induces weaker RPEs in the RL system. Further supporting this interpretation, we observed that subjects who continued to use WM with learning (i.e., showing less transition to RL) exhibited larger effects of set-size on RPE signaling.

One might expect to observe more reliable indicators of neural computations with easier tasks; our findings show the opposite. These results thus strongly hint at a mechanism by which WM and RL interact beyond the competition for control of action (Poldrack et al. 2001), and specifically at a mechanism by which WM interferes with RL computations. How might this interference occur? One possibility is that the two processes compete not only for guiding action, but also more generally, for example based on their reliability in a given environment. Such interference would mean that in conditions in which WM performs better that RL (eg. early in learning for low set sizes), WM inhibits the whole RL mechanism and thus weakens its characteristic neural signals, such as RPEs. Another possible explanation for the observed interference is cooperative interaction, where WM modifies the reward expectations in the RL system. This would lead – when WM was working well – to higher expectations than would be computed by

pure RL, and thus to weaker RPEs. Future research will need to distinguish these possibilities. There may be other interpretations of the change in RPE signaling with set size, besides our interpretation as an interaction between the RL and WM processes. However, given that behavioral fits strongly implicate separate WM and RL processes in learning (see above and previous studies), and that WM is sensitive to load in other paradigms with similar profiles, this remains the most parsimonious explanation. Note that this interaction also makes other behavioral predictions suggesting that reinforcement value learning is actually enhanced under high WM load; we have recently confirmed this prediction using a novel task building on this line of work (Collins et al, submitted).

Our results are related to recent work on sequential decision making and learning that highlighted the role of a model-free module (similar to our RL model), and of a model-based module, responsible for representing stimulus-action-outcome transitions and using them to plan decisions (Doll et al. 2015). This latter module has been linked to cognitive control and is weakened under load (Otto et al. 2013), suggesting that it may require WM. Moreover, both WM use in the current task and model-based processing in the sequential task are related to the same genetic variant associated with prefrontal catecholaminergic function (Collins & Frank 2012; Doll et al. 2016). Notably, (Daw et al. 2011) showed that RPEs in the striatum were modulated by model-based values, a result that may support our collaborative hypothesis. However, we demonstrate such interaction even in paradigms that are traditionally thought to involve purely “model-free” RL. As there is no sequential dependence between trials, learning in our paradigm does not require learning a transition model or planning. Indeed, we could adequately capture learning curves for individual set-sizes using a purely model-free RL model (Collins & Frank 2012; Collins et al. 2014), with decreasing learning rates across set sizes, but this model has more parameters than RLWM and cannot capture the nuanced effects of e.g., delay and set-size interactions. Thus, our results show that learning in very simple environments that appear to require purely model-free learning still recruits executive functions, with working memory contributing to learning and interfering with the putative dopaminergic RL process. Our results show a similar pattern of RPE activations for subcortical and lateral prefrontal cortex areas, a common finding in published studies (e.g. Badre & Frank 2012; Frank & Badre 2012), possibly reflecting a common dopaminergic input to both regions (Bjorklund & Dunnett 2007).

We investigated the role of working memory using set-size as a proxy. However, this leaves open some questions and may limit some of our interpretations. In particular, set-size affects the overall load of working memory, but is also predictive of higher delays between repetitions of the same stimulus. While our analyses tease apart load from delay, the delay itself comprises both a temporal component (number of seconds over which working memory could decay passively), and a discrete component (number of intervening trials that may interfere with working memory). Our paradigm did not manipulate those two factors to make them maximally decorrelated, and cannot distinguish their relative contributions to the effect of delay on behavior. Furthermore, by focusing on set size as the marker of WM, we cannot distinguish between a "tonic", or slowly tuned interference of WM in RL computation, vs. a more "phasic", trial-by-trial adjustment of their role and interaction between them. A target for future research is increasing the experimental paradigm's capacity to carefully disentangle delay from load, allowing us to better understand the dynamics of interactions between RL and WM.

We focused on WM as an alternative learning mechanism from RL, with an a priori interest in regions of the cognitive control network in lateral frontal and parietal cortices. However, regions involved in long-term memory (LTM), such as medial temporal lobe (MTL) and hippocampus, could also play an important role: rote memorization of explicit rules is in the prime domain of LTM, others have shown trade-offs for learning between LTM and striatal based learning (Poldrack et al. 2001), and WM itself is often difficult to distinguish from LTM (Ranganath & Blumenfeld 2005; D'Esposito & Postle 2015). Our results are consistent with LTM having a role in learning: indeed, we observe a negative correlation between RPE and activation in a network of regions including MTL (table 2), indicating higher activation early in learning (Poldrack et al. 2001). However, computational modeling shows that the second learning component we extract is capacity limited, supporting our interpretation of this component as mainly WM. Nevertheless, future research is needed to more carefully dissociate the role of WM from LTM in reinforcement learning.

Learning is a key factor in humans improving their abilities, skills, and fitting to our quickly changing environments. Understanding what distinct cognitive and neurological components contribute to learning is thus essential, in particular to study differences in

719 learning across individuals. Many neurological and psychiatric disorders include learning
720 impairments (Huys et al. 2016). To precisely understand how learning is affected by
721 these conditions, we must be able to reliably extract separable cognitive factors,
722 understand how these factors interact, and link them to their underlying neural
723 mechanisms. Our results provide a first step toward clarifying how we trade off working
724 memory and integrative value learning to make decisions in simple learning
725 environments, and how these processes may interfere with each other.

Badre, D. & D'Esposito, M., 2007. Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *Journal of cognitive neuroscience*, 19(12), pp.2082–99.

Badre, D. & Frank, M.J., 2011. Mechanisms of Hierarchical Reinforcement Learning in Cortico-Striatal Circuits 2: Evidence from fMRI. *Cerebral cortex (New York, N.Y. : 1991)*, pp.1–10.

Bjorklund, A. & Dunnett, S.B., 2007. Dopamine neuron systems in the brain: an update. *Trends in Neurosciences*, 30(5), pp.194–202.

Botvinick, M.M., Niv, Y. & Barto, A.C., 2009. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, 113(3), pp.262–80.

Burnham, K.P. & Anderson, D.R., 2002. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach (Google eBook)*, Springer.

Collins, a. G.E. et al., 2014. Working Memory Contributions to Reinforcement Learning Impairments in Schizophrenia. *Journal of Neuroscience*, 34(41), pp.13747–13756.

Collins, A. & Koechlin, E., 2012. Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making J. P. O'Doherty, ed. *PLoS Biology*, 10(3), p.e1001293.

Collins, A.G.E. & Frank, M.J., 2013. Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, 120(1), pp.190–229.

Collins, A.G.E. & Frank, M.J., 2012. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *The European journal of neuroscience*, 35(7), pp.1024–35.

Collins, A.G.E. & Frank, M.J., Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive.

D'Esposito, M. & Postle, B.R., 2015. The Cognitive Neuroscience of Working Memory. *Annual Review of Psychology*, 66(1), pp.115–142.

Daw, N.D. et al., 2011. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6), pp.1204–15.

Daw, N.D. & Doya, K., 2006. The computational neurobiology of learning and reward. *Current opinion in neurobiology*, 16(2), pp.199–204.

Doll, B.B. et al., 2015. Model-based choices involve prospective neural activity. *Nature neuroscience*, (February), pp.1–9.

Doll, B.B. et al., 2016. Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning. *Journal of Neuroscience*, 36(4), pp.1211–1222.

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(28), pp.7900–7905.

Fedorenko, E., Duncan, J. & Kanwisher, N., 2013. Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 110(41), pp.16616–21.

Flandin, G. & Friston, K.J., 2016. Analysis of family-wise error rates in statistical parametric mapping using random field theory.

Frank, M.J., Seeberger, L.C. & O'reilly, R.C., 2004. By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science (New York, N.Y.)*, 306(5703), pp.1940–3.

Guitart-Masip, M. et al., 2012. Go and no-go learning in reward and punishment: interactions between affect and effect. *NeuroImage*, 62(1), pp.154–66.

Hart, a. S. et al., 2014. Phasic Dopamine Release in the Rat Nucleus Accumbens Symmetrically Encodes a Reward Prediction Error Term. *Journal of Neuroscience*, 34(3), pp.698–704.

Huys, Q.J.M., Maia, T. V & Frank, M.J., 2016. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 19(3), pp.404–413.

Jocham, G., Klein, T. a & Ullsperger, M., 2011. Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(5), pp.1606–13.

Montague, P.R., Dayan, P. & Sejnowski, T.J., 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 16(5), pp.1936–47.

Nassar, M.R. & Frank, M.J., 2016. Taming the beast: Extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences*, 11, pp.49–54.

Otto, a R. et al., 2013. Working-memory capacity protects model-based learning from stress. *Proceedings*

of the National Academy of Sciences of the United States of America, 110(52), pp.20941–6.
Pessiglione, M. et al., 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, 442(7106), pp.1042–5.
Petrides, M., 1985. Deficits on conditional associative-learning tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia*, 23(5), pp.601–614.
Poldrack, R. a et al., 2001. Interactive memory systems in the human brain. *Nature*, 414(November), pp.546–550.
Ranganath, C. & Blumenfeld, R.S., 2005. Doubts about double dissociations between short- and long-term memory. , 9(8).
Schönberg, T. et al., 2007. Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(47), pp.12860–7.
Schultz, W., 2002. Getting formal with dopamine and reward. *Neuron*, 36(2), pp.241–63.
Schwarz, G., 1978. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2), pp.461–464.
Sutton, R.S. & Barto, A.G., 1998. *reinforcement learning*, MIT Press.
Yeo, B.T.T. et al., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 106, pp.1125–1165.

Figure Legends

Figure 1: Experimental Protocol. At the beginning of each block, subjects were shown for 10 s the set of stimuli they would see in that block. In this example, Block 1 uses color patches for stimuli and has a set size n_s =*2*; Block n uses shapes and has n_s =*6.* Each trial included the presentation of a stimulus for 0.5s followed by a blue fixation cross until subject pressed one of three buttons, or up to 1.5s after trial onset. Button press caused the fixation cross to turn white. Feedback was presented for 1s, and came 1.5s after trial onset. Feedback consisted of the words correct or incorrect in green and red, respectively. The inter-trial interval consisted of a white fixation cross with jittered duration to allow trial by trial event-related analysis of fMRI signal. Blocks set sizes varied between 1 and 6, and the order was randomized across subjects.

Figure 2: Behavioral Results. A) Proportion of correct choices as a function of how many times a specific stimulus was encountered (i.e., learning curves), for each set size. B) Logistic regression on factors that contribute to accuracy for a given image, including set size (NS), delay since last previous correct choice for a given image (D), PCor (number of previous correct choices for that image), and their interactions. C) Illustration of the interaction between delay and set size. D) Illustration of the interaction between set size and PCor – early indicates PCor<4, late indicate PCor>6. Error bars indicate standard error of the mean.

Figure 3: Model Validation. A-C) Proportion of correct responses as a function of how many times a specific stimulus was encountered, for each set size, for simulation of different models with individually fit parameters. Models were simulated a 100 times per subject then averaged within subjects to represent this subject's contribution. Error bars indicate standard error of the mean across subjects. A) simple RL model including decay and different sensitivity to gains/losses. B) Identical model to A, with learning rate varying per set size. C) Model incorporating both RL and WM. D) Model comparisons show a significantly lower AIC for RLWM than RL6 or RL, for a significant number of subjects. Each cross indicates a single subject. E) Model comparison to other potential models show best fit for RLWM (see methods for other model names).

Figure 4: Whole brain effects of RPE and RPExns. A-B) Regions positively correlated with RPE (p<.05 cluster corrected). C) Regions showing a positive interaction of RPE with set size.

Figure 5: Striatum and Fronto-parietal ROIs show increased RPE effects in higher set sizes. Average beta coefficient for RPE regressor per set size for A) Striatal ROI and B) Fronto-parietal Network ROI defined by Yeo et al. Error bars indicate standard error of the mean.

Figure 6: Effect of set size on RPE in the fMRI signal is related to individual differences in behavior: Effect of set size on RPE in the fMRI signal is related to individual differences in behavior. Left: average model-inferred mixture weight assigned to working memory over RL ("Model mean WM weight") is significantly related to a stronger effect of set size in fronto-parietal ROI (r=0.49, p=0.02) and in the striatum (r=0.55, p=0.01). Middle: decrease in working memory weight from early (first 3 iterations) to late in a learning block (last 3 iterations) is significantly related to fMRI effect in FP ROI (r=-0.46, p=0.03), and marginally so in striatum (r=-0.41, p=0.06). Right: the behavioral set-size effect is measured as the logistic regression weight of the set-size predictor; stronger behavioral effect is marginally related to a stronger neural effect in FP ROI (r=-0.41, p=0.059) and in striatum ROI (r=0.4, p=0.063).

fMRI activations from Prediction Error contrasts

Table 1: Main effect of RPE

All clusters reliable at p < .05 corrected. Coordinates are the center of mass in MNI.

A) Contrast: Main Effect of RPE > Fixation

Region	BA	Extent (voxels)	*x*	*y*	*z*	Peak *t*-val
Right Angular Gyrus	7	3202	34	-60	42	10.83
	40		46	-52	44	9.2
Right Inferior Parietal Gyrus	40		42	-42	40	10.21
Left Superior Parietal Gyrus	7	3317	-30	-54	44	10.43
Left Angular Gyrus	40		-46	-48	56	10.32
Left Inferior Parietal Gyrus	40		-42	-42	42	9.45
Right Superior Frontal Sulcus	6	12409	20	2	62	9.64
Right Middle Frontal Gyrus	46		38	36	30	8.84
Left Superior Frontal Gyrus	6		-24	-6	62	8.15
Left Middle Frontal Gyrus	11	1686	-30	56	4	7.78
Left Lateral Orbital Gyrus	46		-40	56	-2	6.96
Left Anterior Orbital Gyrus	11		-24	44	-14	6.42
Right Putamen		955	28	22	0	6.99
Right Thalamus			12	-10	10	5.15
Right Pallidum			12	0	6	4.36
Right Precuneus	7	731	6	-64	40	6.32
	7		8	-66	58	5.21

B) Contrast: Main Effect of RPE < Fixation

Region	BA	Extent (voxels)	*x*	*y*	*z*	Peak *t*-val
Right Superior Occipital Gyrus	18	9715	16	-92	24	10.22
Left Superior Occipital Gyrus	18		-16	-96	18	8.73
Right Inferior Lingual Gyrus	30		-10	-48	-6	8.9
Left Cingulate Gyrus (subgenual)	11	2264	-4	28	-12	8.52
	25		-2	18	-8	7.34

Left Superior Frontal Gyrus	10		-8	58	2	7.24
Left Middle Temporal Gyrus	20	2543	-56	-8	-18	6.69
Left Supramarginal Gyrus	48		-36	-36	22	6.26
Left Superior Temporal Gyrus	38		-34	8	-20	6.24
Right Precentral Sulcus	4	1248	26	-30	66	6.21
Right Postcentral Gyrus	4		36	-26	72	6.13
Right Precentral Gyrus	4		52	-12	58	5.62
Right Superior Temporal Gyrus	38	336	30	10	-28	6.08
Right Middle Temporal Gyrus	21		50	2	-26	5.56
	21		58	0	-24	4.76
Right Cingulate Gyrus	23	516	6	-20	44	5.64
Right Superior Frontal Gyrus	6		12	-18	62	5.03
Right Cingulate Sulcus	4		10	-16	54	4.56
Right Superior Temporal Gyrus	48	935	54	-4	4	5.6
Right Lateral Fissure	48		50	4	-6	5.11
Right Lateral Fissure/Insular Gyrus	48		40	-14	20	5.04

Table 2: set-size * RPE interaction

Contrast: RPE Parametric Increasing With Set Size

Region	**BA**	**Extent (voxels)**	***x***	***y***	***z***	**Peak *t*-val**
Left Superior Precentral Sulcus	44	725	-46	10	36	5.69
Left Inferior Frontal Sulcus	48		-38	20	28	5.16
Left Middle Frontal Gyrus	6		-32	2	38	4.57
Right Superior Frontal Gyrus	6	689	18	4	54	5.42
	32		6	22	46	4.3
Left Superior Frontal Gyrus	6		-6	10	50	4.08
Left Intraparietal Sulcus	7	463	-26	-66	44	5.28
	7		-30	-58	46	5.24
	19		-26	-68	34	4.59

Table 3: Main effect of set size

A) **Contrast: Set Size Parametric Increasing**

Region	BA	Extent (voxels)	*x*	*y*	*z*	Peak *t*-val
Left Precuneus	7	1948	-6	-72	44	6.98
Left Angular Gyrus	40		-32	-50	36	6.4
Right Precuneus	7		12	-70	44	5.22

B) Contrast: Set Size Parametric Decreasing

Region	BA	Extent (voxels)	*x*	*y*	*z*	Peak *t*-val
Right Superior Frontal Gyrus	9	1344	14	58	34	6.64
Left Superior Frontal Gyrus	9		-12	46	42	4.69
	10		-4	58	28	4.5
Left Supramarginal Gyrus	40	447	-64	-44	34	6.31
Left Angular Gyrus	39		-52	-70	28	5.83
	40		-60	-52	40	4.52
Right Angular Gyrus	40	255	58	-52	44	5.41
	22		62	-54	28	4.41
Right Supramarginal Gyrus	40		64	-46	36	5.19
Right Superior Frontal Gyrus	8	239	14	20	62	5.15
	9		10	38	52	4.42
Left Superior Frontal Sulcus	8	255	-24	22	58	4.9
Left Middle Frontal Gyrus	46		-24	18	40	4

A)	K	α	ϕ_{WM}	ρ	ϕ_{RL}	ε	pers
Mean (std)	4.08 (0.98)	0.07 (0.13)	0.29 (0.31)	0.86 (0.18)	0.05 (0.05)	0.03 (0.03)	0.34 (0.31)
Median	4	0.03	0.18	0.94	0.05	0.03	0.25
Min - max	2-5	0.01-0.5	0-1	0.42-1	0-0.21	0-0.14	0.02-1

B)	K	α	ϕ_{WM}	ρ	ϕ_{RL}	ε
α	ns					
ϕ_{WM}	ns	0.77				
ρ	ns	-0.65	-0.77			
ϕ_{RL}	ns	0.83	0.69	-0.62		
ε	ns	ns	ns	ns	ns	
pers	ns	ns	ns	ns	ns	ns

Table 4) RLWM model fit parameters. A) Parameter statistics B) Correlation between parameters. ns indicates non-significant correlation ($p<.05$, corrected for multiple comparisons.