Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory

Anne G. E. Collins[a,b,1] and Michael J. Frank[c]

[a]Department of Psychology, University of California, Berkeley, CA 94720; [b]Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720; and [c]Department of Cognitive, Linguistic and Psychological Science, Brown Institute for Brain Sciences, Brown University, Providence, RI 02912

Edited by James L. McClelland, Stanford University, Stanford, CA, and approved January 24, 2018 (received for review December 8, 2017)

Learning from rewards and punishments is essential to survival and facilitates flexible human behavior. It is widely appreciated that multiple cognitive and reinforcement learning systems contribute to decision-making, but the nature of their interactions is elusive. Here, we leverage methods for extracting trial-by-trial indices of reinforcement learning (RL) and working memory (WM) in human electro-encephalography to reveal single-trial computations beyond that afforded by behavior alone. Neural dynamics confirmed that increases in neural expectation were predictive of reduced neural surprise in the following feedback period, supporting central tenets of RL models. Within- and cross-trial dynamics revealed a cooperative interplay between systems for learning, in which WM contributes expectations to guide RL, despite competition between systems during choice. Together, these results provide a deeper understanding of how multiple neural systems interact for learning and decision-making and facilitate analysis of their disruption in clinical populations.

reinforcement learning | working memory | EEG | computational model | dynamics

When learning a new skill (like driving), humans often rely on explicit instructions indicating how to perform that skill. However, for many problems, these instructions may be too numerous to keep in working memory (WM), and one needs to focus on a subset of them while acquiring large portions of skills by trial and error, or reinforcement learning (RL): "Practice makes perfect." Previous research showed that dual cognitive and incremental RL systems contribute to learning across a range of situations, even when explicit instructions are not provided, and stimulus–response contingencies must be acquired solely by reinforcement (1–7).

This body of work is motivated by theoretical considerations suggesting that RL and cognitive systems optimize different trade-offs. The RL process statistically integrates reinforcement history to estimate the expected value of choices, in accordance with "model-free" algorithms that guarantee convergence, but are slow and inflexible (8). This process is widely thought to be implemented in cortico-basal ganglia loops and their innervation by dopaminergic signals (9, 10). In contrast, the cognitive system facilitates more flexible and rapid learning, but is limited by WM capacity, is subject to forgetting, and is evidenced by differential efficiency of learning in simple and complex environments (6). The WM system is a primitive for more "model-based" or goal-directed cognitive processes and is thought to depend on prefrontal cortex among other regions (1, 3, 11).

Although it is well established that multiple systems contribute to learning, their interactions are poorly understood. Most models assume that distinct systems compete for influence over behavioral output. However, the nature of their interaction during learning, in terms of how one system's updating of learned knowledge influences another's, is far less clear. Recent studies have shown that reward prediction error (RPE) signals—canonical neuroimaging signatures of model-free RL—are more strongly represented (3), and behavioral value learning is actually enhanced, under high compared with low WM load (7). However, there are multiple forms of interaction that could give rise to these effects, which were not possible to disambiguate in previous work (3, 7).

Here, we combined computational modeling, electro-encephalography (EEG), and decoding to provide insight into this issue. Specifically, EEG allowed us to interrogate within-trial dynamics of the two systems and how they are combined to converge on a single decision and interpret an outcome. We used computational modeling to quantify variables involved in RL and WM and decoding to identify their signatures in EEG. First, we confirmed that EEG markers of reward expectation at decision onset are negatively coupled with markers of RPE in the subsequent feedback (FB) period within the same trial, as predicted by axiomatic tenets of RL, but never directly shown in neural data. Second, we predicted that we would see markers of RL processing earlier than those of WM in the neural signal, given that the latter process is more cognitively costly. Finally, we investigated whether the two systems update learned knowledge independently or if they influence each other. As noted above, earlier work has hinted that WM and executive functions might interfere with, or modify, RL computations (3, 4, 7, 12), but the nature of these interactions remains elusive. We leveraged model-informed within- and across-trial analyses of EEG decoding signals to arbitrate between three possibilities: independent processes, inhibition of RL by WM, or a cooperative contribution of WM to RL expectations. We show evidence for the latter type of interaction between WM and RL, whereby the

Significance

A major factor that improves learning in artificial agents is the use of multiple algorithms in parallel to benefit from their complementary strengths across different environments. The human brain performs a similar optimization, balancing the use of resource-intensive but immediately accessible information in working memory and a more reliable slow but steady reinforcement learning to build habits. These parallel computations are evident in neural signal dynamics that unfold across both short- and long-term time scales, which reveal that the two processes compete for decisions but cooperate for learning. These findings further our understanding of human learning and may inspire better artificial learners.

Author contributions: A.G.E.C. and M.J.F. designed research; A.G.E.C. performed research; A.G.E.C. contributed new reagents/analytic tools; A.G.E.C. analyzed data; and A.G.E.C. and M.J.F. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: annecollins@berkeley.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1720963115/-/DCSupplemental.

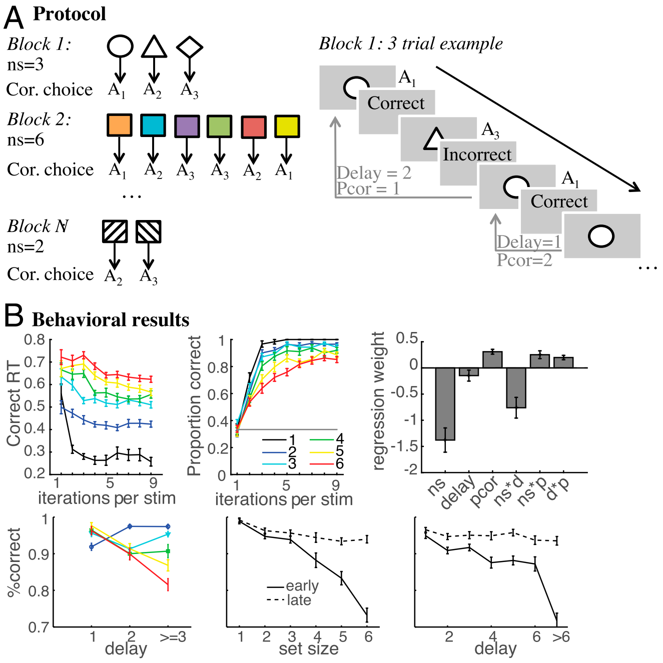

Fig. 1. Experimental protocol and behavioral results. (*A*) In each block, participants use deterministic reward FB to learn which of three actions to select for each stimulus image. The set size (or number of stimuli; *ns*) varies from one to six across blocks. (*B*, *Upper Left* and *Upper Center*) Reaction times (RT) and performance learning curves for each set size as a function of number of iterations of a stimulus (stim). (*B*, *Upper Right*) Logistic regression weights show contributions of WM (smaller set sizes and smaller delays facilitate performance) and RL [incremental effects of previous correct trials (*pcor*) for a stimulus] and their interactions. *B*, *Lower* shows that these interactions are mediated by greater effects of delay in high set sizes (*Left*) and reduced effects of both set size and delay as learning progresses from early to late in a block (*Center* and *Right*), suggestive of a transition from WM to RL (1).

RL process is counterintuitively weakened when the learning environment is least complex (i.e., WM load is lowest).

Results

To parse out contributions of RL and WM to learning, we used our RLWM task (Fig. 1*A*) (1–3, 7) while recording EEG (*Materials and Methods*). Participants learned via reinforcement to select one of three actions for each visual stimulus. WM demands were manipulated by varying across blocks

Behavioral results from 40 participants replicated previous findings implicating separable RL and WM systems, with the relative contribution of WM decreasing with learning. First, participants were more likely to select the correct choice as the number of previous correct (*pcor*) trials accumulated (Fig. 1*B*) a basic marker of incremental RL [$t(39) = 9.1$, $P < 10^{-4}$]. Second, correct performance was more rapidly attained in lower set sizes and declined with increasing set sizes [*ns*; $t(39) = -5.4$, $P < 10^{-4}$] and delays [$t(39) = -3.1$; $P = 0.004$], with delay effects amplified under high loads ($t = -4.2$, $P = 0.0002$), consistent with contributions of a capacity- and maintenance-limited WM system. Finally, interactions between the three factors showed that set-size and delay effects decreased with learning (t values > 3.2, $P < 0.003$), confirming a shift from WM to RL with experience (Fig. 1*B*) (1–3, 7).

Trial-by-Trial Decoding of Model-Based Indices of RL and WM. We used our previously developed computational model to quantitatively estimate the contributions of RL and WM to each participant's behavior. The model included a standard model-free RL module, which estimated the expected "*Q*" value of stimulus–action pairs and incrementally updated those values on each trial in proportion to the RPE. This module was complemented by a WM module that assumed perfect memory of the last trial's stimulus–action–outcome transition, but had limits on both capacity *K* (number of items that can be held in mind, such that the probability of recall $P = K/ns$) and on maintenance (memory for transitions is decayed on each subsequent trial, due to forgetting/updating of intervening items). Model selection confirmed that the RLWM model quantitatively fit participants' behavior better than other models that assumed only a single process (Fig. 2), and simulations of the RLWM model captured participants' patterns of behavior (Fig. 2, *Right* and Fig. S1). We then extracted trial-by-trial estimates of the expected *Q* value and RPE from the RL module (thus factoring out WM contributions to behavior), as a quantity of interest for model-based analysis of EEG.

To investigate the contributions of RL and WM in the neural signals, we leveraged a trial-by-trial decoding approach to analyzing the EEG data (13). We used a regression approach to simultaneously extract the effect of multiple variables of interest on the EEG signal at all time points and electrodes, using correction for multiple comparisons, while controlling for other factors (such as reaction times), and separating out the role of correlated predictors. We identified clusters of electrodes and time points that showed significant sensitivity to each predictor. The main predictors of interest were the set size, the delay, and, from the model, the expected *Q* value (for stimulus-locked analysis) and RPE (for FB-locked analysis).

In stimulus-locked EEG, this analysis yielded significant and widespread effects of all three main regressors and, similar to behavior, an interaction of set size with delay, indicative of WM (Fig. 3*A* and Figs. S2 and S3). Notably, neural markers of *Q* values appeared substantially earlier (starting at ∼230 ms after stimulus onset) than those for set size (peaking at ∼600 m; Fig. 3), supporting the existence of two separable processes sensitive to RL and WM within a trial. Moreover, the early signal modulated the scalp voltage distribution in the same way (Fig. 3*C*) for increasing *Q* values (when the RL system had learned more) and increasing delays (when the WM system was less likely to contain the relevant information), and thus putatively signaled the early recruitment of the RL system. For FB-locked analysis, we observed robust effects of RPE, and

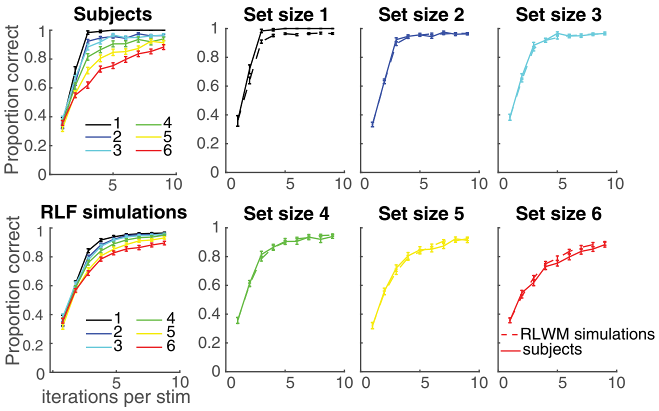

Fig. 2. Model validation. (*Left*) Simulations of the RLF model (a standard RL model with forgetting) with fit parameters do not capture behavior appropriately. (*Right*) Simulations of the RLWM model with fit parameters captures learning curves in most set sizes. Simulations were run 1,000 times per subject. stim, stimulus.

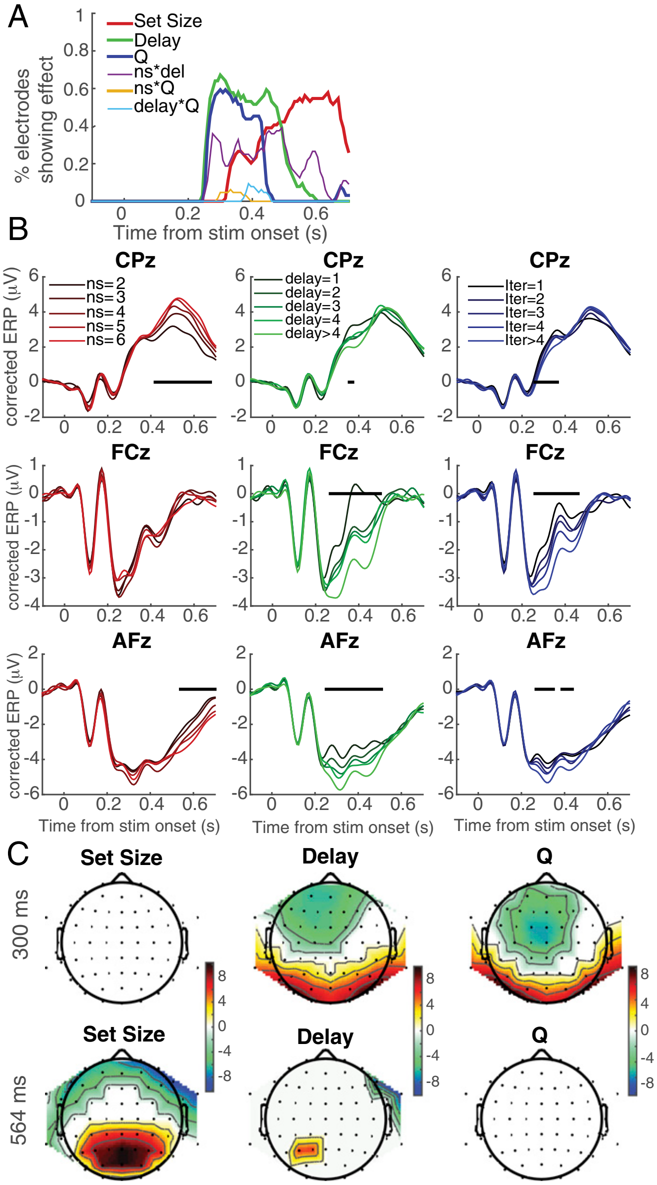

Fig. 3. EEG decoding of RL and WM effects during choice. (*A*) Proportion of electrodes showing a significant effect ($P < 0.05$ cluster corrected with $P < 0.001$ for cluster formation threshold), for each predictor in a multiple-regression analysis, against time from stimulus (stim) onset. *Q* values and delays are decoded early (~230 ms), whereas set-size *ns* is decoded later (peaking at ~600 ms). (*B*) Corrected event-related potentials (ERPs) are plotted to visualize effects accounting for each regression factor [set-size *ns*, *delay*, stimulus iterations *iter* for three electrodes (CPz, FCz, and AFz)]. (*C*) Scalp topography at an early (300 ms) and late (564 ms) time point, plotting significance-thresholded average regression weights for the three main predictors. White is nonsignificant; warm colors are positive; cold colors are negative.

RPE-modulated by delay, but only very weak effects of set size and delays (Fig. 4 and Fig. S4).

Testing Axiomatic Indices of RL Signals and Interactions with WM. We next leveraged these quantities conveying *Q* value and RPE signals at distinct time points to test a central axiomatic tenet of neural RL theories (14), which, to our knowledge, has not been directly evaluated in neural signal. If neural signals on individual trials truly reflect the latent model variables of expected value and RPE, they should provide more informed estimates of those quantities than those inferred from model fits to behavior alone. Thus, variance in trial-wise stimulus-locked *Q*-value signals should (negatively) predict variance in subsequent FB-locked RPE signals in the same trial (i.e., greater expectations should be met with diminished surprise), over and above the behavioral RPE (Fig. 5 *A* and *B*). Indeed, while (by definition) the behavioral RPE accounted for most of the variance in the FB-locked RPE signal [$t(38) = 10.9$, $P < 10^{-4}$], increases in trial-wise neural metrics of expected *Q* value were associated with lower neural indices of RPE [$t(38) = -2.08$, $P = 0.045$], as expected from the computation of $RPE = reward - Q$.

If the RL and WM processes are independent, neural indices of RL should be independent of set size. Conversely, if RL processing is degraded when the task becomes more difficult, one might expect that these indices would degrade with set size. Instead, we observed strong evidence for the opposite effect: RL indices were actually enhanced under high load, for both stimulus- and FB-locked activity [Fig. 6*A*: $t(38) = 2.4$, $P = 0.02$; Fig. 6*B*: $t(38) = 4.5$, $P = 10^{-4}$]. This result is inconsistent with independent RL and WM processes and instead suggests an interaction between them; but what is the form of this interaction?

Distinguishing Competitive from Cooperative Accounts via EEG Temporal Dynamics. Recent neuroimaging and behavioral findings (refs. 3 and 7; see also ref. 4) suggested RL–WM interactions during learning, but could not distinguish whether this interaction was cooperative or competitive. Under the competitive hypothesis, the WM system would compete with, and hence hinder, the RL system from learning when WM is reliable (i.e., during low set sizes; Fig. 6 *B*, *Upper*). Under the cooperative hypothesis, the WM system would instead inform reward expectations needed to compute RPE. Hence, within low set sizes, RPEs would be smaller compared with those expected by pure RL (Fig. 6 *C*, *Upper*). Thus, both hypotheses could account for the blunted RL signaling observed in low set sizes (Fig. 6*A*). However, they make qualitatively different predictions regarding

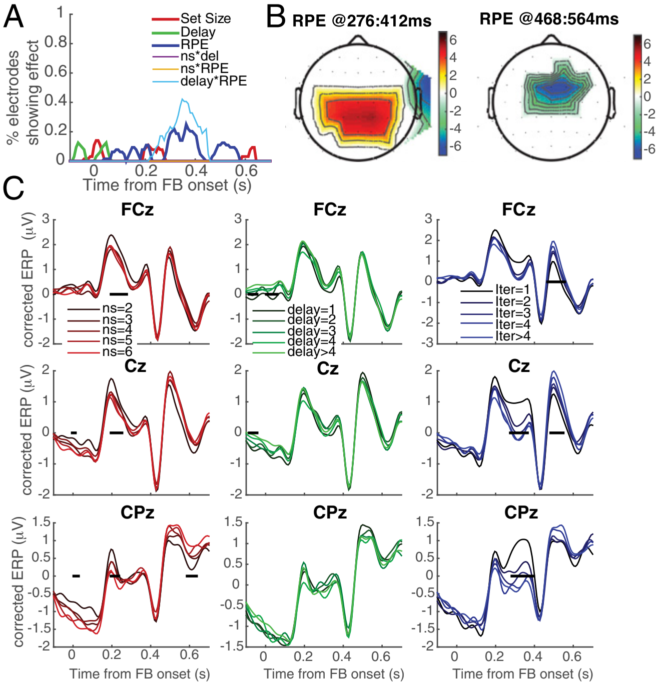

Fig. 4. EEG decoding of RL and WM effects during FB. (*A*) Proportion of electrodes showing a significant effect for each predictor. (*B*) Scalp topographa at two time points for the RPE regressor. (*C*) Corrected event-related potentials (ERPs). Iter, iterations. See Fig. 3 for details.

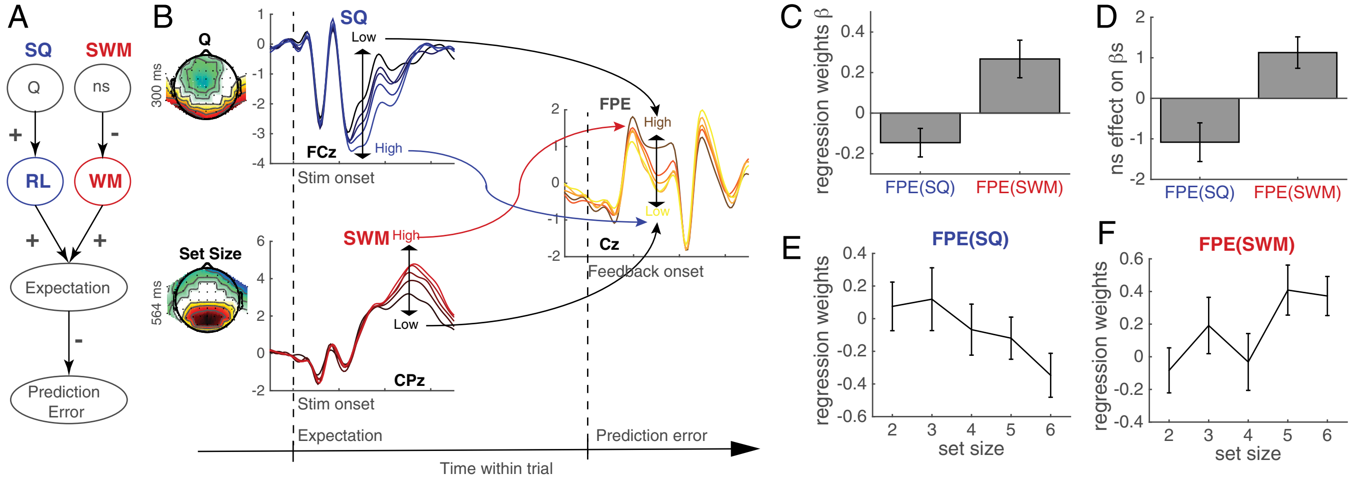

Fig. 5. Within-trial dynamics support RL and WM contributions to learning. (*A*) A basic tenet of RL is that reward expectation is negatively correlated to the RPE, through the formula *RPE = reward − expectation*. Expectations can be informed by both RL and WM systems, where RL accumulates reward experience to estimate value *Q*, and WM is less reliable with set size (*ns*), reducing WM contributions to expectation. (*B*) Schematic of the trial-by-trial prediction method. For each trial, we computed a stimulus-locked index of RL-related activity (*SQ*) and of WM-related activity (*SWM*), using similarity to spatiotemporal masks obtained from the multiple-regression analysis (Fig. 3). The model in *A* predicts that trial-by-trial variability in *SQ* negatively predicts neural responses to prediction error during FB (FPE) of that same trial, with the opposite effect of *SWM*, after controlling for the behavioral RPE. Scalp topographies of SQ and SWM at early and late time points are displayed to the left of each index. (*C*) Trial-by-trial variance in FPE is significantly and oppositely accounted for by variability in *SQ* and *SWM*. (*D–F*) These effects are amplified in high set sizes, in which the RL system is relatively more potent (1–3).

the dynamic changes in the RPE signal across trials (Fig. 6 *B*, *Lower*). Specifically, the cooperative account predicted that RPE signals decline more rapidly in low set sizes (due to the contributions of the fast-learning WM system to expectations). In contrast, the competitive model predicted the opposite pattern: When WM dominates in low set sizes, it suppresses

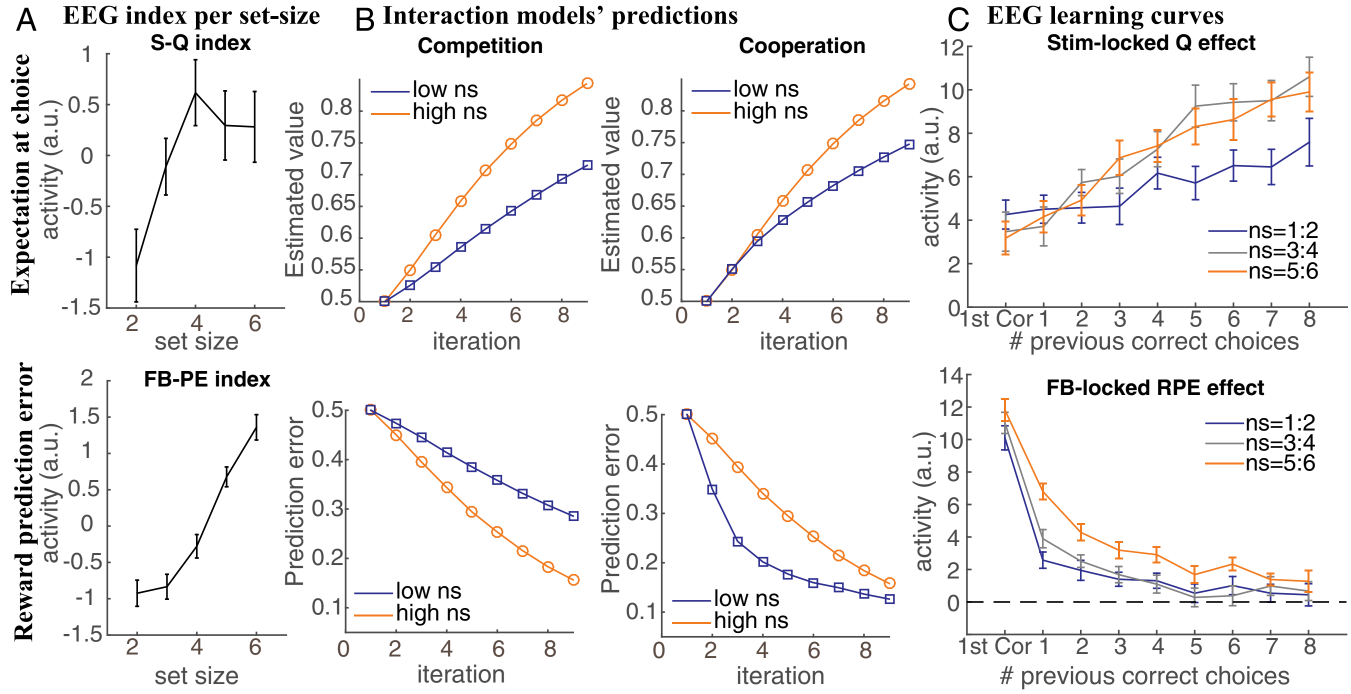

Fig. 6. Distinguishing cooperative vs. competitive RL–WM interactions via temporal dynamics of EEG decoding, for stimulus-locked expectation (*Upper*) and FB-locked RPEs (*Lower*). (*A*) Neural markers of RL at both stimulus presentation (*S-Q*) and FB (*FB-PE*) are amplified with higher set size, suggesting that WM and RL are not independent. (*B*) Computational model simulations show that both a competitive and a cooperative account can account for blunted RL computations with lower set sizes (*Upper*). However, the competitive account predicts that RPEs decline more rapidly across stimulus iterations in high (orange) compared with low (black) set sizes, whereas the cooperative model predicts the opposite pattern. (*C*, *Upper*) Stimulus (stim)-locked indices of RL increase with learning (by definition), and do so more slowly in low set sizes, consistent with the observed interactions and with both computational models. (*C*, *Lower*) FB-locked indices of RPE decrease with learning, and do so more slowly in high set sizes, consistent with the cooperative and refuting the competitive model.

learning in the RL system, and, hence, RL RPEs persist for more trials (compared with high set sizes, where RL learning is unhindered). We tested these hypotheses by plotting the average stimulus- and FB-locked RL indices as a function of the number of previous correct choices per stimulus, thus obtaining "neural learning curves." As predicted by both models, learning curves for neural measures of expected Q values were blunted under low load (Fig. 6 *C*, *Upper*). However, supporting the cooperative, but not competitive, model, the neural signal of RPEs declined more rapidly under low set sizes than high set-sizes [Fig. 6 *C*, *Lower*; $t(38) = 3.93$, $P = 0.0003$].

To further test this hypothesis, we investigated within-trial temporal dynamics indicative of cooperation. In particular, we asked whether trial-by-trial variance in neural markers of WM during stimulus presentation could predict RPE-related variance in the subsequent FB period. Specifically, if WM cooperated with RL expectations to compute RPEs, then a high perceived WM load decoded by the stimulus-locked WM index should be predictive of a larger FB-locked RPE (Fig. 5*A*). We tested this in a multiple-regression model that simultaneously accounted for variance in Q-value signals, WM signals, and RPE estimated by fits to behavior alone.

Results showed that, indeed, neural metrics of WM (corresponding to higher perceived load) during stimulus onset were associated with larger FB-locked RPE signals [Fig. 5*C*; $t(38) = 2.89$, $P = 0.006$]. These positive effects contrasted with the negative effects of Q-value signals (which support prediction error computations), and were predicted by the cooperation model, whereby higher perceived load was indicative of degraded WM representations, and hence RPE signals were enhanced relative to when WM is intact. Fig. S5 shows that this result held when regressors were orthogonalized. Fig. 5 *C*–*F* show that these effects held within individual objective set sizes, suggesting that the neural index of perceived load was more predictive than objective load, and, indeed, were even more pronounced under high set sizes, when WM and RL were more likely to jointly contribute to behavior [SQ: $t(38) = -2.27$, $P = 0.03$; SWM: $t(38) = 2.9$, $P = 0.006$]. Finally, we confirmed the robustness of this double dissociation using a bootstrapping method. Specifically, we shuffled the masks used to obtain trial-level indices of SQ and SWM and found that the more similar shuffled masks were to Q-value mask, the more they predicted decreased neural RPEs, and the more similar they were to WM masks, the more they predicted increased neural RPEs ($P < 0.0006$).

Discussion

Our findings support a growing literature implicating multiple separable neural processes contributing jointly to human instrumental learning and decision-making (1–4, 7). While many imaging studies implicate interactive systems with strong debates about how they are arbitrated for choice (3, 4, 7, 15–17), they have not resolved the specific nature of these interactions, either for choice or for learning. Our model-based decoding of single-trial EEG dynamics revealed a cooperative interaction between WM and RL systems during learning, in addition to their competitive interaction for choice.

Specifically, our multiple-regression analysis of EEG signals identified two separable spatiotemporal networks sensitive to dissociable aspects of learning. Early on during the choice period, the neural signal was sensitive to reward history, a marker of model-free RL, whereas a later-onset signal was sensitive to set size, a marker of WM load. These results seem to indicate a shift from an early recruitment of a fast, automated RL process to a more deliberative WM (18)—a conclusion supported by our finding that the early signals were more strongly recruited when WM would be weaker (with increased intervening trials) and thus favoring RL recruitment for decision-making.

Within-trial correlations between choice and FB dynamics confirmed, first, that these were neural signatures of RL and, second, that the RL system was informed by WM for learning. First, trial-by-trial variations of signal encoding Q values during expectations were negatively predictive of variation in that same trial of signals encoding RPE, providing evidence for the notion that neural RPE signals compute *reward* – Q *value*. These data provided axiomatic evidence (14) for a central but heretofore untested account of neural RL via within-trial dynamics. Second, in contrast to the neural signature of Q, we found that those indicative of higher WM load during expectations were positively related to subsequent RPE signals. These findings provided dissociable signals related to WM and RL expectation that exhibit differential effects on RPE signals as predicted by the cooperation model. Moreover, both of these findings accounted for variance in RPE signals over and above those that could be predicted based on model fits to behavior alone, providing further confirmation that they are related to the computations of interest and evidence that neural markers can be used as a more direct lens into value and decision computations (19–22). These results could not be attributed to generic (negative) autocorrelation in the neural signal; indeed, shuffling of electrode and time clusters confirmed that the degree to which neural measures of expectation influenced subsequent RPE signals was directly related to their similarity to specific WM and RL masks.

While much past research has argued for competition between multiple systems during choice (1, 4, 15, 16), these studies usually still assume that the systems learn independently, or even compete for learning (23). By contrast, our previous behavioral and fMRI findings hinted that the RL computations were not independent of WM and, indeed, that the RL process was actually stronger in more difficult task settings, under high WM load (3, 7). Our findings here showed that, indeed, trial-wise EEG markers of the RL process were stronger with increasing WM load, during both decision and FB periods. However, weaker RL signals under low load might be explained by multiple forms of interaction, and previous studies could not identify the nature of these interactions. Thus, the current study strived to confirm the role of WM in reducing RL signals under low load (as other mechanisms could be considered) and to pinpoint the nature of this role. The dynamic decoding analysis used here clearly favors a cooperative mechanism during learning, whereby WM use can augment expectations of reward within RL, and thereby reduce subsequent RPEs. These findings directly contrast with the predictions by a competitive account of learning, in which reliable WM signals would suppress RPEs within an independent RL system (Fig. 6). Second, we showed that trial-by-trial variability in WM signaling in the neural signal at the time of decision predicted variability in the RL signal during subsequent FB. Together, these findings strongly support our proposed cooperative mechanism, which was not possible in previous studies. It is important to note that our finding contributes to a growing literature showing that multiple systems contributing to learning are not fully separable and that their computations may be more integrated than previously thought.

Of course, although ascending monoaminergic systems may affect cortical EEG responses to RPEs (24), one limitation of EEG is that it cannot directly index subcortical markers of RPE, such as striatum, and we cannot rule out the possibility that such signals would reflect a "purer" RL system that is protected from WM influences. However, we have shown in the same paradigm using fMRI that canonical striatal RPE signals are also influenced by WM (3). Moreover, the RL markers we do observe in EEG signals are signatures predicted from canonical models. Finally, other putatively striatal-dependent behavioral measures support the same interaction, whereby reward value learning is counterintuitively enhanced in high set sizes (7). This could have reflected competitive or cooperative dynamics, but cannot be explained by pure RL. In sum, multiple lines of evidence suggest that the pure RL signal is biased.

It is interesting to note that the "cooperative" mechanism here interfered with the RL computation. By decreasing the magnitude of the RPE before the estimate of the Q value has converged, it slowed the learning of the RL Q values, and thus diminished RL computations overall, as observed in the neural signal (Fig. 6*C*). This mechanism predicted that statistical learning of expected reward values would be degraded under low load, a phenomenon we observed behaviorally in a variant of this task using multiple reward outcomes (7). However, while WM might hinder RL in this task, this interaction may be useful in general, allowing WM to be used judiciously for tasks that are less well learned and the RL system to take over when it has accumulated sufficient information. Indeed, since the RL computations occur earlier in the trial, if they are sufficiently reliable, the learner might learn to use only RL and not recruit WM, as observed over the course of learning (1, 18).

To conclude, our results contribute to a better understanding of human learning. First, they show evidence of separable neural processes of WM and RL contributing to learning and competing for decisions in the EEG signal. Second, they provide trial-by-trial evidence for computation of RPEs in the EEG signal related to the RL process. Third, they provide evidence for a cooperative interplay between WM and RL systems for learning, despite a competitive dynamic during choice. Identifying the neural correlates of the multiple systems that jointly contribute to human learning and decision making is crucial to better understanding dysfunction (2, 7, 25). Our results are thus an important step toward better understanding of learning in healthy and patient populations.

Materials and Methods

Subjects. We collected data for 40 subjects (28 female, ages 18–29), and all were included in the behavioral analyses. One subject was excluded from EEG analysis due to technical problems with the EEG cap. All participants were compensated for their participation and gave informed, written consent as approved by the Human Research Protection Office of Brown University.

Experimental Protocol. Subjects performed a learning experiment in which they used reinforcement FB to figure out which key to press for each presented visual stimulus (Fig. 1). The experiment was divided into 22 blocks, with new visual stimuli in each block. After stimulus presentation, subjects selected one of three keys to press with their right hand. FB indicated truthfully whether they had selected the correct action for the current stimulus. Blocks varied in the number of stimuli that participants learned concomitantly (the set size *ns*) between one and six. See *SI Materials and Methods* for details of the experimental paradigm.

Computational Modeling. We used a version of the RLWM model (1) to account for subjects' behavior and disentangle roles of WM and RL to choices. See *SI Materials and Methods* for modeling details.

EEG Analyses. EEG was recorded from a 64-channel Synamps2 system (0.1–100 Hz bandpass; 500 Hz sampling rate). EEG preprocessing followed standard procedures, and multiple-regression analyses followed similar techniques as in ref. 26; see details in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank Julie Helmers for help with data collection, as well as Matthew Nassar, Samuel McDougle, Nicholas Franklin, and Andrew Westbrook for comments on this manuscript. This work was funded by NSF Grant 1460604 (to M.J.F. and A.G.E.C.).

1. Collins AGE, Frank MJ (2012) How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *Eur J Neurosci* 35:1024–1035.
2. Collins AGE, Brown JK, Gold JM, Waltz JA, Frank MJ (2014) Working memory contributions to reinforcement learning impairments in schizophrenia. *J Neurosci* 34:13747–13756.
3. Collins AGE, Ciullo B, Frank MJ, Badre D (2017) Working memory load strengthens reward prediction errors. *J Neurosci* 37:4332–4342.
4. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69:1204–1215.
5. Doll BB, Duncan KD, Simon DA, Shohamy D, Daw ND (2015) Model-based choices involve prospective neural activity. *Nat Neurosci* 18:767–772.
6. Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND (2013) Working-memory capacity protects model-based learning from stress. *Proc Natl Acad Sci USA* 110:20941–20946.
7. Collins AGE, Albrecht MA, Waltz JA, Gold JM, Frank MJ (2017) Interactions among working memory, reinforcement learning, and effort in value-based choice: A new paradigm and selective deficits in schizophrenia. *Biol Psychiatry* 82:431–439.
8. Sutton RS, Barto AG (1998) *Reinforcement Learning* (MIT Press, Cambridge, MA).
9. Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* 16:1936–1947.
10. Collins AGE, Frank MJ (2014) Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychol Rev* 121:337–366.
11. D'Esposito M, Postle BR (2015) The cognitive neuroscience of working memory. *Annu Rev Psychol* 66:115–142.
12. Doll BB, Hutchison KE, Frank MJ (2011) Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *J Neurosci* 31:6188–6198.
13. Collins AGE, Frank MJ (2016) Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. *Cognition* 152:160–169.
14. Rutledge RB, Dean M, Caplin A, Glimcher PW (2010) Testing the reward prediction error hypothesis with an axiomatic model. *J Neurosci* 30:13525–13536.
15. Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* 8:1704–1711.
16. Lee SW, Shimojo S, O'Doherty JP (2014) Neural computations underlying arbitration between model-based and model-free learning. *Neuron* 81:687–699.
17. Poldrack RA, et al. (2001) Interactive memory systems in the human brain. *Nature* 414: 546–550.
18. Viejo G, Khamassi M, Brovelli A, Girard B (2015) Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. *Front Behav Neurosci* 9:225.
19. Niv Y, Edlund JA, Dayan P, O'Doherty JP (2012) Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *J Neurosci* 32:551–562.
20. Frank MJ, et al. (2015) fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *J Neurosci* 35:485–494.
21. Kahnt T, Heinzle J, Park SQ, Haynes J-D (2011) Decoding the formation of reward predictions across learning. *J Neurosci* 31:14624–14630.
22. Turner BM, van Maanen L, Forstmann BU (2015) Informing cognitive abstractions through neuroimaging: The neural drift diffusion model. *Psychol Rev* 122:312–336.
23. Doya K, Samejima K, Katagiri K, Kawato M (2002) Multiple model-based reinforcement learning. *Neural Comput* 14:1347–1369.
24. Holroyd CB, Coles MGH (2002) The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychol Rev* 109: 679–709.
25. Huys QJM, Maia TV, Frank MJ (2016) Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat Neurosci* 19:404–413.
26. Collins AGE, Cavanagh JF, Frank MJ (2014) Human EEG uncovers latent generalizable rule structure during learning. *J Neurosci* 34:4677–4685.
27. Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464.