Three systems interact in one-shot reinforcement learning

Amy R. Zou (amyzou@berkeley.edu)
Department of Psychology
University of California, Berkeley

Anne G. E. Collins (annecollins@berkeley.edu)
Department of Psychology
Helen Wills Neuroscience Institute
University of California, Berkeley

Abstract

Human adaptive decision-making recruits multiple cognitive processes for learning stimulus-action (SA) associations. These proceses include reinforcement learning (RL), which represents gradual estimation of values of choices relevant for future reward-driven decisions, episodic memory (EM), which stores precise event information for long-term retrieval, and working memory (WM), which serves as flexible but temporary, capacity-limited storage. However, we have limited understanding of how these systems work together. Here, we introduce a new one-shot RL task to disentangle their respective roles. In 16 independent 8-trial blocks, 144 participants used one-shot rewards to learn 4 new SA associations per block. Each block provided one chance to obtain feedback for pressing one of two keys for each stimulus (trials 1–4), followed by a chance to use this feedback to make a choice in a short-term association task (trials 5–8; no feedback), primarily targeting WM. In a subsequent testing phase designed to assess long-term retention through RL or EM, all 64 stimuli were shown in randomized order and subjects were asked to press the correct key for each, without feedback. Trials 5–8 revealed WM-dependent strategy effects on choice accuracy, as well as a role for both RL and EM when WM is overwhelmed. Testing phase accuracy depended on feedback interacting with initial presentation order, revealing signatures of both RL and EM in learning from one-shot rewards. Computational modeling suggests that a mixture model combining RL and EM components best fits group-level testing phase behavior. Our results show that our new protocol can identify signatures of each of the three memory systems' contributions to reward-based learning. With this approach, we create new possibilities to better understand how each integrates a single bit of information, what their exact contributions to choice are, and how they interact.

Keywords: Reinforcement learning; episodic memory; computational modeling

Introduction

Learning is important to daily life, enabling humans to make rewarding decisions and efficiently adapt to new situations. Learning is often studied within the framework of reinforcement learning (RL) (Sutton & Barto, 2018; Eckstein, Wilbrecht, & Collins, 2021), which assumes that agents learn from past outcomes to estimate expected values of different actions and use these values to inform future decisions. RL algorithms have been successful at describing both behavior and brain function (Eckstein et al., 2021).

However, recent evidence has shown that beyond RL, other cognitive processes that store information in a different format also contribute to learning. One such process is working memory (WM) (Collins & Frank, 2012; Collins, 2018; Yoo & Collins, 2021), which is a temporary, capacity-limited, and effortful but extremely flexible form of short-term memory storage. Previous studies have shown how learning can be slowed by the *load effect* characteristic of WM capacity/resource limitations (Oberauer et al., 2018), and that WM interferes with RL computations (Collins, 2018; Collins & Frank, 2018). However, what exact information is stored in WM on a trial-by-trial basis remains unclear. A second cognitive process that contributes to learning alongside RL is episodic memory (Bornstein, Khaw, Shohamy, & Daw, 2017), a hippocampal/medial temporal lobe-dependent system that stores very precise information for long-term retrieval (Ritchey, Montchal, Yonelinas, & Ranganath, 2015). Key characteristics of episodic memory (EM) are its temporal and context sensitivity, which contribute to serial position effects where retrieval is more successful at the beginning of an episode (*primacy effect*) or more recently (*recency effect*) (Ebbinghaus, 1913). Past studies have leveraged these features to identify EM contributions to learning, but disentangling its precise contribution from RL also remains challenging. Furthermore, EM also appears to interact with WM in different experiments (Murty, FeldmanHall, Hunter, Phelps, & Davachi, 2016; Poldrack & Packard, 2003; Wimmer, Braun, Daw, & Shohamy, 2014).

Thus, our goal was to design a new approach that simultaneously identifies contributions of all three systems in a single learning context to better qualify them individually and in interaction. We designed a novel experiment where participants learned stimulus-action (SA) associations from a single instance of feedback. We investigated how this learning depended on factors that may differently impact RL, WM, and EM (e.g., reward, load, and primacy/recency). We predicted that WM, RL, and EM would contribute to SA association learning, while long-term retention would be predominantly driven by RL and EM. While the whole task allows for an investigation of the role of WM, we focused our computational modeling efforts toward RL and EM as a preliminary step.

Methods

Participants

The research was approved by UC Berkeley's Institutional Review Board. We recruited young adults through Amazon Mechanical Turk (MTurk; $n = 160$, 49 women, age = 31.9±4.84 years) and the university's psychology undergraduate student population (RPP; $n = 95$, 74 women, age=20.7±2.45 years). MTurk participants received monetary

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

compensation ($6/hour), and student participants received course credit. All participants provided informed consent before completing the task online (approximately 12 minutes long).

We excluded participants based on poor training phase performance: if they responded to more than 10 trials (15%) unreasonably fast (under 200 ms), performed below chance on more than 5 out of the 16 blocks, or missed more than 5 trials (9%). Following exclusion criteria, we analyzed a final sample size of 66 for MTurk (22 female, 31.8±5.42 years) and 78 for RPP (59 female, 20.8±2.68 years).

Task design

Participants completed a one-shot RL task (Fig. 1), where the goal was to determine the correct key responses to visual stimuli. The task consisted of a training phase and a testing phase. The training phase comprised 16 blocks (stimuli were categorically independent across blocks), each with 8 trials. Participants' goal was to use one-shot rewards to learn 4 SA associations in each block. On trials 1–4, they saw 4 distinct images presented sequentially, responded to each by pressing one of two keys ("J" or "K") and immediately received a truthful, positive (+1) or negative (0) feedback. On trials 5–8, they saw the same 4 images in a different order and needed to press the correct key given what they learned on trials 1–4. Here, a trial was considered correct if the participant repeated the rewarded choice or avoided the unrewarded choice, and incorrect otherwise. They did not receive immediate feedback on trials 5–8, but were instead told at the end of the block how many of the four they answered correctly. Participants could rest for up to 30 s between training blocks.

The training phase was designed to assess how various features of the task design, like feedback and stimulus presentation order, impacted how SA associations were learned. Although positive and and negative feedback were of equal use in identifying the correct response key, values are unlikely to be updated equivalently in light of positive and negative outcomes (Katahira, 2018). There is strong evidence that values are updated asymmetrically in RL-based learning, making feedback valence sensitivity an important marker of RL. Feedback could also impact WM or long-term EM storage, as participants may choose to prioritize positive information in WM or only remember the SA event (but not the outcome) in EM. Similarly, stimuli encountered on trial 1, and those seen on trial 4 then trial 5 may benefit respectively from *primacy* and *recency* effects. We evenly pseudo-randomized the sequence and distribution of trial features encountered in the first 4 trials to analyze their impact on learning in a balanced manner. Finally, a key feature of this task is that participants may adopt different strategies to learn the correct key, affecting how the three components of RL, WM, and EM interact with each other. Variability in strategic approach could lead to differences in what information is used by WM for learning (see "Exploration strategy" below).

In the testing phase, participants saw all 64 stimuli again (4 images for each of 16 blocks) in a randomized order, and were

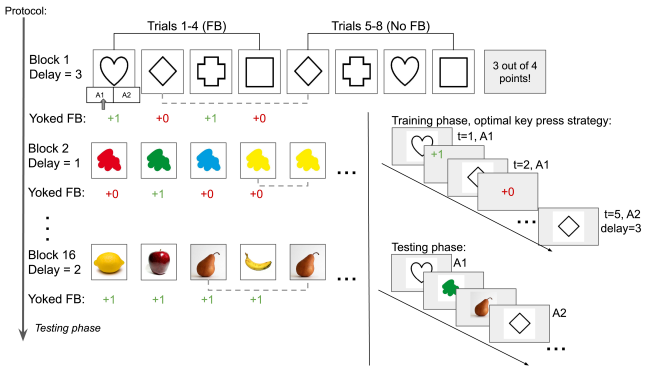

Figure 1: Experiment protocol. Participants first completed 16 independent training blocks where they learned stimulus-action associations from one-shot rewards, then a testing phase probing long-term retention of all previously learned associations.

told to press the correct key for each image shown. Again, they received no immediate feedback for each choice (to avoid further learning), but were told that their performance was still tracked (for motivation), and shown how many total points they collected throughout this phase.

We aimed to test long-term retention in the testing phase—specifically, contributions of RL and EM—as we expected that randomization and delay in stimulus presentation would wash out WM effects. Thus, manipulation of features like feedback and presentation order related to RL and EM processes would allow us to better understand their effects on long-term retention of learned SA associations.

Participants viewed the instructions and completed a practice block prior to starting the true task. They could practice 3 times, and were allowed to move on with a score above chance (50% correct); otherwise the experiment ended. We implemented two attention checks for the MTurk population to ensure data quality (Gureckis et al., 2016): one at the end of the practice trials during the instructions, and one halfway through the training phase. The experiment ended early for participants who did not reach performance higher than chance (50%) at these cutoff points.

Behavioral analysis

Exploration strategy. We observed a common action selection pattern that emerged from exploratory behavior on trials 1–4 of the training phase, where participants had no prior information (Mohr et al., 2018). Multiple participants adopted a *same-key strategy*, pressing the same-key for all trials in the first half of a block. This strategy reduces the WM load for computing the correct action following feedback. A participant who presses different keys on trials 1–4 needs to track the initial key press, the outcome, and compute a key switch if the action was wrong. However, if they pressed the same key for all trials, an unrewarded outcome would always signal pressing the other key. Thus, the same-key strategy

reduces WM load by collapsing the number of dimensions of features participants need to track, as they no longer need to remember the initial action by trial to generate the correct response. As a subset of participants used this strategy consistently throughout the training phase, we divided the sample into those who used the same-key strategy for at least half of the training phase, i.e., only using the strategy after block 9 or earlier, and those who did not. As differences in strategy use may introduce variation in WM contributions, we limited our model-based analyses to only those in the same-key strategy group ($n = 75$), since our modeling goal was to investigate EM and RL processes behind testing phase behavior.

Statistical analysis Our primary outcome of interest was choice accuracy. To evaluate the effect of different experimental manipulations, we ran one mixed effects logistic regression separately for data from trials 5–8 in the training phase, and another for all trials in the testing phase. Predictors included the feedback received for this stimulus during trials 1–4 (FB; 1/0), presentation order (order; 1–4), block number (1–16), and the strategy used (dummy coded as 1 if the same-key strategy was used, 0 otherwise). For the training phase, we included two interactions with strategy: one for FB and one for presentation order. For the testing phase, we included the corresponding training phase performance of this stimulus, as well as an interaction between FB and presentation order. For both phases, we included random effects across participants.

Computational models

We focused our computational modeling on the testing phase to model only the contributions of RL and EM. We did not model WM, and to further reduce variance related to WM use, we only included the subset of participants who used the same-key strategy. To attempt to capture signatures of either process in the testing phase behavior, we built 5 computational models that capture different assumptions about their separate or mixed contributions.

Our first models capture different three nested RL models. Each of these captures a different RL process by which an agent could learn to maximize its cumulative reward based on outcomes resulting from taking action on stimuli in its environment.

RL: In our base RL model, we implemented a classic delta-rule learning algorithm, where the expected value Q at time t of an action a for a stimulus s is updated by the reward prediction error—the difference between outcome r and prior expectations of $Q(a,s)$—scaled by learning rate α ($0 < \alpha < 1$):

$$Q_t(a,s) = Q_{t-1}(a,s) + \alpha\,(r_{t-1} - Q_{t-1}(a,s)) \qquad (1)$$

After Q-values were initialized at 0.5, we assume that Q-values are updated only when explicit feedback is received as a reward r_t on trials 1–4 of training.

Q-values are converted to action policy via softmax:

$$P(a|s) = \frac{\exp(\beta\, Q_t(a,s))}{\sum_i \exp(\beta\, Q_t(a_i,s))} \qquad (2)$$

All RL models used the same softmax equation to calculate action probabilities. Because there is only one learning trial, the learning rate and softmax inverse temperature parameter β would not be jointly recoverable, as their product will uniquely influence the choice policy on trials 5–8. Thus, we fixed β to 10 and only estimated the learning rate(s).

RL2a: This model extends the base RL model by including separate learning rates for gains and losses (α when $r_t = 1$ and α_{neg} when $r_t = 0$), capturing an often observed effect of feedback valence in RL (Katahira, 2018).

RLRe: This model extends the RL2a to include an additional parameter *re*, used to iteratively update expected values of previously seen stimuli/action on following trials. In this way, the RLRe model implements a form of offline rehearsal of previously experienced trials in a block:

$$Q(a_i,s_i,b) = Q(a_i,s_i,b) + re \cdot \alpha\,(r_i - Q(a_i,s_i,b)) \qquad (3)$$

For trial $i < t$. When reward $r_i = 0$:

$$Q(a_i,s_i,b) = Q(a_i,s_i,b) + re \cdot \alpha_{neg}\,(r_i - Q(a_i,s_i,b)) \qquad (4)$$

This model allows us to consider the possibility of an RL-family model that might nevertheless exhibit temporal order effects.

EM model: Our descriptive EM model quantitatively implements a memory process of probabilistic storage and retrieval of a trial's information. It features within-block primacy and feedback-dependent encoding/retrieval accuracy.

To model potential primacy effects, we assume that the probability of retrieving a memory of a trial, $p(ret)$, depends on when in the block it was presented (i.e., presentation order s at storage), via an exponentially decaying memory parameterized by time constant τ:

$$p(ret) = \exp(-\tau(s-1)) \qquad (5)$$

Furthermore, we assume that encoding and retrieval accuracies are not perfect, and can potentially be corrupted in a feedback-dependent way. Specifically, probabilities of correct retrieval $0 < m < 1$ (when FB=1) and $0 < m_{neg} < 1$ (when FB=0) are combined with probability of retrieving $p(ret)$ to compute the final probability of choosing the correct action, $p(cor)$, under the assumption that choice is random (0.5) if the memory was not stored or retrieved:

$$p(cor) = p(ret)\cdot m + (1 - p(ret))\cdot 0.5 \qquad \text{if FB = 1} \quad (6)$$

$$p(cor) = p(ret)\cdot m_{neg} + (1 - p(ret))\cdot 0.5 \qquad \text{if FB = 0} \quad (7)$$

Note that values of $m_{neg} < 0.5$ mean that memory encoding is worse than chance, capturing the possibility that participants store the wrong memory. This could show up as participants remembering an unrewarded action instead of the

correct one. We tested a simpler model where $m = m_{neg}$, but did not present it here as it did not account for the data well.

Mixture model: Finally, we included a mixture model, RLEM, that allows for RL and EM processes to jointly contribute to testing phase behavior. The model's policy is a weighted sum of the RL2a model's policy p_{RL} and the EM model's policy p_{EM}, as follows:

$$p(a|s) = \rho\ p_{RL}(a|s) + (1-\rho)\ p_{EM}(a|s) \tag{8}$$

Where ρ is the mixture weight parameter $(0 < \rho < 1)$.

Modeling methods

The goal of our computational modeling was to better understand the mechanisms driving long-term retention in the testing phase; accounting for short-term memory use during the training phase was beyond the scope of this paper. Thus, we limited our modeling to capturing RL and EM contributions to learning in the testing phase, and our modeling dataset to participants who used the same-key strategy ($n = 75$) to keep WM effects consistent between subjects and remove further confounds introduced by different WM strategies in the training phase. We fit our models to testing phase choice data, conditioned on information learned during training. The models were updated on trials 1–4 of each block of the training phase; then we evaluated how well the policies predicted the actions selected during the testing phase.

We used maximum likelihood estimation to fit our models via MATLAB's `fmincon` function with 10 random starting points (Wilson & Collins, 2019). We used AIC for model comparison, and validated parameter identification and model comparison procedures with simulation studies (Wilson & Collins, 2019). We simulated data from each model of interest and fitted all models on that simulated data to determine recoverability of the original simulated model. We calculated the exceedance probability (Rigoux, Stephan, Friston, & Daunizeau, 2014), or the probability that each model generated its own data, for each model.

For model validation, we generated 10 simulated datasets using parameters obtained from model fitting. This allowed us to qualitatively compare experimental result patterns to those obtained from simulations (Palminteri, Wyart, & Koechlin, 2017).

Results

Behavioral Results

Learning The RPP and MTurk groups both performed above chance early in the training phase (Block 1 accuracy around 80%). Because both groups performed similarly in both the training (MTurk: $M = 0.862, SD = 0.094$; RPP: $M = 0.862, SD = 0.094$; independent t-test: $t(142) = -0.18, p = 0.86$) and testing phases (MTurk: $M = 0.601, SD = 0.109$; RPP:$M = 0.602, SD = 0.111$; independent t-test: $t(142) = 0.13, p = 0.89$), we combined both groups into one dataset in all further analyses.

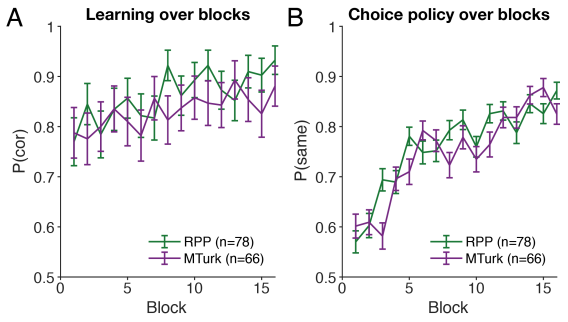

Figure 2: Task performance in the training phase. A) Participants from both groups perform similarly in trials 5–8, improving their choice accuracy ($P(cor)$) across training blocks. B) Participants from both groups adopt the same-key strategy ($P(same)$) over training blocks (MTurk: $\rho = 0.912$; RPP: $\rho = 0.905$, both $p < 0.001$).

Participants demonstrated meta-learning effects, improving mean accuracy for trials 5–8 across independent blocks (Fig. 2a; block regressor $\beta = 0.0363, p < 0.001$ in Fig. 2a), and learned to use a choice policy that limited WM load throughout the training phase (Fig. 2b).

A mixed effects logistic regression revealed significant main effects of strategy ($\beta = 1.71, p < 0.001$), presentation order (order; $\beta = 0.0806, p = 0.0323$), and initial feedback received (FB; $\beta = 0.757, p < 0.001$), as well as interactions between strategy and FB ($\beta = -0.788, p < 0.001$) and strategy and order ($\beta = -0.147, p = 0.008$) (Fig. 3a; Fig. 4, red). However, while FB and order had significant effects on training phase performance, using the same-key strategy diminished these effects (Fig. 3a), indicating that WM storage was not sensitive to FB or order when under minimal load.

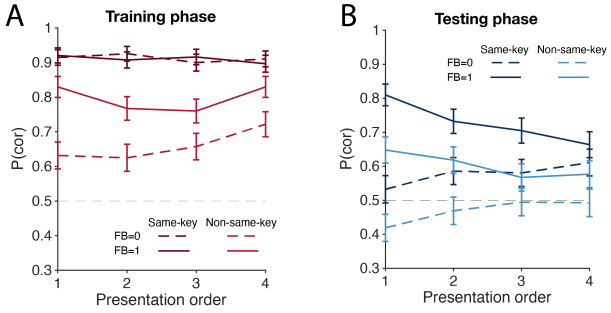

Figure 3: Trial-level performance in each phase as a function of initial feedback received, presentation order, and WM strategy. A) In the training phase, choice strategy interacts with initial feedback received and presentation order. B) In the testing phase, choice strategy has a main effect, while initial feedback and presentation order interact.

Testing We extended the previous analysis to the testing phase (Fig. 3b; Fig. 4, blue). Controlling for whether par-

ticipants made the correct choice in trials 5–8 (Train Correct; $\beta = 0.336, p < 0.001$), training phase choice strategy ($\beta = 0.281, p < 0.001$) still had a significant effect on testing phase performance, hinting at deep interactions between short-term strategies and long-term memory systems RL and EM.

Testing phase performance revealed signatures of markers associated with both RL and EM (Fig. 3b). Accuracy was higher for initially rewarded SA associations, a *positivity bias* likely reflecting the involvement of feedback-based learning via RL to learn stimulus-action-outcome (SAO) associations. There was a slight long-term recency effect (effect of block; $\beta = 0.0191, p < 0.001$). Accuracy also varied as a function of within-block temporal order; within-block primacy effects emerged, potentially indicating the role long-term EM for learning SA associations (Ebbinghaus, 1913). Main effects were significant for FB ($\beta = 1.44, p < 0.001$), presentation order ($\beta = 0.100, p < 0.001$), and their interaction ($\beta = -0.307, p < 0.001$). This interaction captures the performance difference between rewarded and unrewarded trials at t = 1, hinting at the possibility that participants retrieved SA events, rather than SAO events, determining long-term memory retrieval in a feedback-independent way.

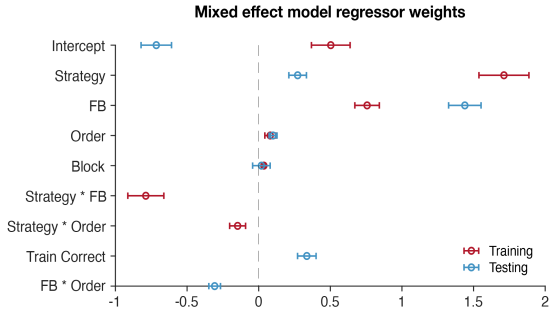

Figure 4: Regressor weights from linear mixed effects models for training (red) and testing (blue) phases. Points represent estimates of main effects and interactions, bars represent standard error. All effects are significant (at $p < 0.05$).

Modeling results

Model comparison We first evaluated model-fitting results at the subject-level (yielding one set of parameters per subject) via AIC score comparison (Fig. 5a). The EM model (3 parameters) had the lowest AIC score, while the mixture model RLEM (6 parameters) performed the worst. We suspected that this outcome was due to insufficient trial data at the subject level, leading to overfitting by RLEM, which had 3 additional parameters. Indeed, simulations with the winning EM model were unable to capture the key behavior of failing to store unrewarded SAO associations on t=1, showing the limitations of this model comparison approach (Palminteri et al., 2017). In fitting the models at the group-level, where we combined all data into one dataset to generate one set of parameters for the whole group, the AIC score comparison showed that RLEM performed best, confirming that it was previously disadvantaged by a shortage of trials (Fig. 5b). Future work will use hierarchical modeling to overcome this issue while accounting for individual differences (Baribault & Collins, 2021).

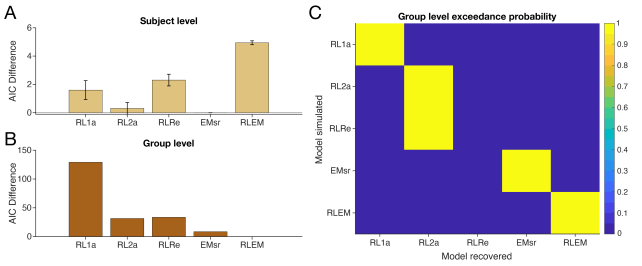

Figure 5: Model comparison and validation of comparison. A) and B) show AIC score differences between the 5 models fit at the subject-level and group-level, respectively. C) Confusion matrix of exceedance probabilities indicate most models are identifiable.

Model validation Simulations of the RLEM best captured the qualitative trends observed in the real data, replicating the main effects of feedback and the crucial performance gap at t=1 (Fig. 6). As expected, the base RL model failed to capture feedback or order effects, the more complex RL models replicated the feedback effect but not order effects, and the EM model failed to capture the interaction. A model recovery studied verified the interpretability of our parameter recovery results, which was successful for RLEM other models. Furthermore, exceedance probabilities from model recovery showed that most models, including the winning RLEM model, were identifiable (Fig. 5c). Fit model parameters confirmed an RL asymmetric learning rate bias ($\alpha = 0.55, \alpha_{neg} = 0.33$), a greater likelihood of encoding SA vs. SAO for negative feedback ($m = 0.72, m_{neg} = 0.41$), and a greater contribution of EM than RL after a single feedback ($\rho = 0.38$); individual parameter fits showed similar findings. While our EM models did not account for experiment-wide memory effects seen in model-independent analyses (block effect in Fig. 4), simulations of a variant EM model with such a global recency mechanism failed to capture key behaviors.

Discussion

Our model-independent analyses suggest that our task is able to parse out distinct contributions of three systems in a one-shot reward learning paradigm. The training phase was designed to incorporate and counterbalance task features with selective effects on each of the three systems—valenced feedback targeting RL's asymmetrical learning mechanism, stimulus presentation order targeting EM's within-block primacy effect, and a load around the WM limit depending on strategy use—to evaluate their contributions and interactions. Indeed, WM effects emerged in the training phase where short-term memory effects should be more dominant: efficiency storage

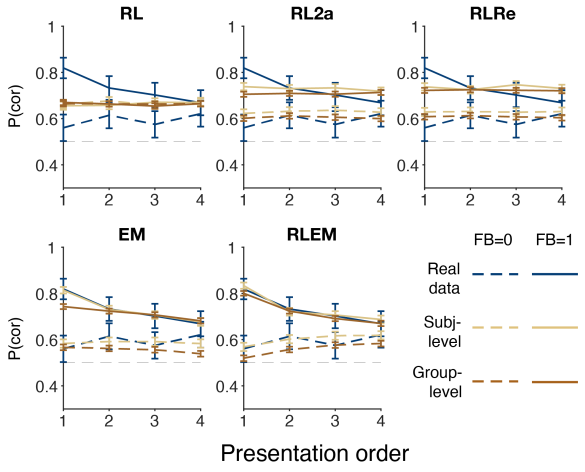

Figure 6: Model validation results. Simulations of all 5 models; RLEM replicates qualitative trends observed in real data (grey lines) at the group and subject-level.

of SA associations in WM via the same-key strategy closes feedback- and order-sensitive performance gaps indicative of other systems. Thus, training performance appears to mostly rely on close to perfect WM when the same-key strategy is used, but relies on a mixture of all three components otherwise. Future work is needed to better qualify this short term contribution of RL and EM in the non-same-key strategy blocks.

In the testing phase, there should be little direct contribution of WM because it is beyond both the capacity and temporal scope of WM. Nevertheless, there is evidence for WM interacting with RL and EM, as efficient WM use in training mildly improves testing phase performance, controlling for other factors. Regardless of choice policy, however, performance in both groups was impacted by the initial feedback received and its interaction with presentation order. The improved performance for rewarded stimuli is non-trivial as participants should have access to the same information in this two-alternative forced choice design, independent of feedback. This bias can be interpreted in a number of ways. First, it could be a marker of RL function, which has frequently been shown to be positively biased (Palminteri et al., 2017; Katahira, 2018; Xia, Master, Eckstein, Wilbrecht, & Collins, 2020; Master et al., 2020), though not always (Sugawara & Katahira, 2021). Second, this could reflect an EM effect, where only the SA components of the association was stored and subsequently retrieved regardless of whether the action was correct.

The interaction between presentation order and reward effect supports the second interpretation, whereby EM contributes to the feedback effect. A purely RL-dependent reward effect should be identical across all training phase presentation trials 1-4. However, this effect was stronger at t=1 where EM storage/retrieval was improved by within-block primacy effects, potentially due to EM's potential failure to bind reward to event information, or a lower likelihood of EM storage of unrewarded info. Our model comparison results suggest that long-term retention observed in the testing phase is described by a mixture model that captures the feedback-dependent process of RL and the temporally-sensitive and feedback-dependent process of EM, also supporting this interpretation.

One study limitation is a shortage of trials to properly fit the mixture model on individuals' performance. This forced us to sacrifice either the statistical power necessary for model comparison or the ability to account for individual differences. As we continue our work with this approach, we will instead use Bayesian hierarchical modeling methods, which will enable us to have maintain sufficient estimation power while simultaneously capturing individual differences. This will allow us reconcile the conflicting subject- and group-level accounts in follow-up experiments designed to increase the number of trials for modeling the training phase.

Another limitation lies in our limited ability to study WM contributions. First, we could counterbalance the possible feedback sequences evenly with possible presentation delay durations at only trial 5. This is a minor issue, however, as our priority for the no-feedback training trials was on trial 5, as further trials would be confounded by decision-making processes beyond the scope of this paper. Second, because most participants eventually learned to use the same-key strategy which enabled perfect WM performance (Fig. 3a, dark red), there remained relatively few trials in which we could carefully analyze and model the roles of RL and EM in short-term decision-making on trials 5–8 (Fig. 3a, dark red). Future experiments are needed to better answer this question. Accordingly, we focused our computational on the long-term memory and learning systems in the testing phase as they would be better understood. We further controlled for WM confounds by only modeling data from participants who employed the same strategy that decreased WM load in a consistent way, thereby decreasing potential variance in WM use). Nevertheless, we also conducted model-fitting on the remaining non-same-key strategy participants. While we replicated patterns in model comparison, behavioral simulations, and model recovery, the distribution of fitted RLEM parameter values differed between the strategy subgroups (e.g., for mixture ρ, same-key subjects $M = 0.42, SE = 0.05$ vs. $M = 0.27, SE = 0.03$; independent t-test: $t(142) = 3.66, p < 0.001$). Our current analyses cannot aptly account for individual differences (visible in certain subject-level simulations). Future work would address the role of WM and individual differences in the testing phase.

Disentangling how multiple systems contribute to adaptive decision-making is essential to better understanding where individual differences come from, in healthy individuals, across development and in clinical populations. By demonstrating how these three systems contribute distinct characteristics to learning from one-shot rewards, our study represents an important step in this direction.

References

Baribault, B., & Collins, A. (2021, December). *Troubleshooting Bayesian cognitive models: A tutorial with matstanlib* (preprint). PsyArXiv. doi: 10.31234/osf.io/rtgew

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017, December). Reminders of past choices bias decisions for reward in humans. *Nature Communications*, *8*(1), 15958. doi: 10.1038/ncomms15958

Collins, A. G. E. (2018, October). The Tortoise and the Hare: Interactions between Reinforcement Learning and Working Memory. *Journal of Cognitive Neuroscience*, *30*(10), 1422–1432. doi: 10.1162/jocn_a_01238

Collins, A. G. E., & Frank, M. J. (2012, April). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis: Working memory in reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1024–1035. doi: 10.1111/j.1460-9568.2011.07980.x

Collins, A. G. E., & Frank, M. J. (2018, March). Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proceedings of the National Academy of Sciences*, *115*(10), 2502–2507. doi: 10.1073/pnas.1720963115

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology.* (H. A. Ruger & C. E. Bussenius, Trans.). New York: Teachers College Press. doi: 10.1037/10011-000

Eckstein, M. K., Wilbrecht, L., & Collins, A. G. (2021, October). What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. *Current Opinion in Behavioral Sciences*, *41*, 128–137. doi: 10.1016/j.cobeha.2021.06.004

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2016, September). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842. doi: 10.3758/s13428-015-0642-8

Katahira, K. (2018, December). The statistical structures of reinforcement learning with asymmetric value updates. *Journal of Mathematical Psychology*, *87*, 31–45. doi: 10.1016/j.jmp.2018.09.002

Master, S. L., Eckstein, M. K., Gotlieb, N., Dahl, R., Wilbrecht, L., & Collins, A. G. (2020, February). Disentangling the systems contributing to changes in learning during adolescence. *Developmental Cognitive Neuroscience*, *41*, 100732. doi: 10.1016/j.dcn.2019.100732

Melrose, R. J., Zahniser, E., Wilkins, S. S., Veliz, J., Hasratian, A. S., Sultzer, D. L., & Jimenez, A. M. (2020, February). Prefrontal working memory activity predicts episodic memory performance: A neuroimaging study. *Behavioural Brain Research*, *379*, 112307. doi: 10.1016/j.bbr.2019.112307

Mohr, H., Zwosta, K., Markovic, D., Bitzer, S., Wolfensteller, U., & Ruge, H. (2018, November). Deterministic response strategies in a trial-and-error learning task. *PLOS Computational Biology*, *14*(11), e1006621. doi: 10.1371/journal.pcbi.1006621

Murty, V. P., FeldmanHall, O., Hunter, L. E., Phelps, E. A., & Davachi, L. (2016, May). Episodic memories predict adaptive value-based decision-making. *Journal of Experimental Psychology: General*, *145*(5), 548–558. doi: 10.1037/xge0000158

Nyberg, L., Sandblom, J., Jones, S., Neely, A. S., Petersson, K. M., Ingvar, M., & Backman, L. (2003, November). Neural correlates of training-related memory improvement in adulthood and aging. *Proceedings of the National Academy of Sciences*, *100*(23), 13728–13733. doi: 10.1073/pnas.1735487100

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., … Ward, G. (2018, September). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885–958. doi: 10.1037/bul0000153

Palminteri, S., Wyart, V., & Koechlin, E. (2017, June). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. doi: 10.1016/j.tics.2017.03.011

Poldrack, R. A., & Packard, M. G. (2003, January). Competition among multiple memory systems: converging evidence from animal and human brain studies. *Neuropsychologia*, *41*(3), 245–251. doi: 10.1016/S0028-3932(02)00157-4

Rigoux, L., Stephan, K., Friston, K., & Daunizeau, J. (2014, January). Bayesian model selection for group studies — Revisited. *NeuroImage*, *84*, 971–985. doi: 10.1016/j.neuroimage.2013.08.065

Ritchey, M., Montchal, M. E., Yonelinas, A. P., & Ranganath, C. (2015, January). Delay-dependent contributions of medial temporal lobe regions to episodic memory retrieval. *eLife*, *4*, e05025. doi: 10.7554/eLife.05025

Sugawara, M., & Katahira, K. (2021, December). Dissociation between asymmetric value updating and perseverance in human reinforcement learning. *Scientific Reports*, *11*(1), 3574. doi: 10.1038/s41598-020-80593-7

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Wilson, R. C., & Collins, A. G. (2019, November). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, e49547. doi: 10.7554/eLife.49547

Wimmer, G. E., Braun, E. K., Daw, N. D., & Shohamy, D. (2014, November). Episodic Memory Encoding Interferes with Reward Learning and Decreases Striatal Prediction Errors. *Journal of Neuroscience*, *34*(45), 14901–14912. doi: 10.1523/JNEUROSCI.0204-14.2014

Xia, L., Master, S., Eckstein, M., Wilbrecht, L., & Collins, A. (2020). Learning under uncertainty changes during adolescence. In *Cogsci.*

Yoo, A., & Collins, A. (2021, December). How Working Memory and Reinforcement Learning Are Intertwined: A

Cognitive, Neural, and Computational Perspective. *Journal of Cognitive Neuroscience*, 1–17.