

# Journal of Experimental Psychology: General

## **How Working Memory and Reinforcement Learning Interact When Avoiding Punishment and Pursuing Reward Concurrently**

Peter F. Hitchcock, Joonhwa Kim, and Michael J. Frank

Online First Publication, September 1, 2025. <https://dx.doi.org/10.1037/xge0001817>

### CITATION

Hitchcock, P. F., Kim, J., & Frank, M. J. (2025). How working memory and reinforcement learning interact when avoiding punishment and pursuing reward concurrently. *Journal of Experimental Psychology: General*. Advance online publication. <https://dx.doi.org/10.1037/xge0001817>

# How Working Memory and Reinforcement Learning Interact When Avoiding Punishment and Pursuing Reward Concurrently

Peter F. Hitchcock<sup>1, 2</sup>, Joonhwa Kim<sup>3, 4</sup>, and Michael J. Frank<sup>2, 3</sup>

<sup>1</sup> Department of Psychology, Emory University

<sup>2</sup> Department of Cognitive and Psychological Sciences, Brown University

<sup>3</sup> Carney Institute for Brain Science, Brown University

<sup>4</sup> Department of Neuroscience, Brown University

Humans learn adaptive behaviors via a durable but incremental reinforcement learning (RL) system and a fast but fleeting working memory (WM) system. Past work parsing these systems focused on reward learning alone; hence, little is known about how they interact while simultaneously learning to avoid punishment and whether arbitrating between these demands is disrupted by psychiatric symptoms. We administered a novel reward/punishment RL-WM task to an online sample oversampled for depression and anxiety symptoms ( $N = 298$ ;  $n = 275$  after quality control). Participants avoided punishment during initial learning, yet poorly retained this avoidance. Computational modeling captured this pattern via the fleeting WM system facilitating punishment avoidance, while the durable RL system retained little about punishment. Our task also included two test phases interleaved with learning, which permitted a targeted examination of past findings that WM blunts the RL system. When RL-based retention was tested midway through learning, we indeed found evidence of blunting. Yet, after learning resumed—leading to further prediction errors—blunting was no longer evident in a final test phase. However, individual differences moderated this effect: Some individuals were especially susceptible to blunting; for others, WM actually facilitated retention. Finally, task performance was largely spared as a function of depression/anxiety and trait rumination. Overall, our findings demonstrate that—when seeking to attain reward and avoid punishment concurrently—the WM system can facilitate short-term punishment avoidance while the RL system retains little about punishment, reveal individual differences in the extent to which WM blunts RL, and demonstrate intact behavior under internalizing-disorder symptoms.


## Public Significance Statement

We developed a novel task to disentangle how reinforcement learning and working memory contribute to navigating one of life's fundamental challenges: learning to pursue reward and simultaneously avoid punishment. We found that working memory temporarily facilitates punishment avoidance, yet such avoidance is poorly retained. We also found, on the one hand, that there were substantial individual differences among participants in the extent to which working memory blunted reinforcement learning, yet, on the other hand, depression/anxiety and trait rumination symptoms had little impact on task performance.

**Keywords:** depression and anxiety, individual differences, punishment, reinforcement learning, working memory

**Supplemental materials:** <https://doi.org/10.1037/xge0001817.supp>

Wilma Bainbridge served as action editor.

Peter F. Hitchcock  <https://orcid.org/0000-0001-7606-5132>

The study was preprinted on the Open Science Framework at <https://osf.io/preprints/psyarxiv/82pyz>. Results from this article were presented at the American College of Neuropsychopharmacology 2023 Annual Meeting. All participants gave informed consent prior to the study, and the study was approved by the Brown University Institutional Review Board.

Peter F. Hitchcock was supported by National Institute of Mental Health and National Institutes of Health Grant F32MH123055. Michael J. Frank was supported by National Institute of Mental Health Grants P50MH119467 and R01MH084840-08A. The authors thank the Frank Lab for helpful discussions and Preeti Nagalamadaka for help coding the task. This study used the

Oscar high-performance cluster at Brown University.

Peter F. Hitchcock played a lead role in conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualization, writing—original draft, and writing—review and editing. Joonhwa Kim played a supporting role in investigation, project administration, and software. Michael J. Frank played a lead role in supervision and a supporting role in conceptualization, funding acquisition, methodology, and writing—review and editing.

Correspondence concerning this article should be addressed to Peter F. Hitchcock, Department of Psychology, Emory University, Psychology and Interdisciplinary Sciences Building, 36 Eagle Row, Atlanta, GA 30322, United States. Email: [ptrhitchcock@gmail.com](mailto:ptrhitchcock@gmail.com)

Telling a joke, walking a high wire, choosing a restaurant, raising a delicate subject with a friend—life requires us to constantly navigate situations in which our actions may lead either to reward or to punishment. How do humans learn how to behave in such situations? Part of the answer appears to be that they rely on an evolutionary conserved reinforcement learning (RL) system, centered on the basal ganglia, which learns via an incremental yet durable trial-and-error process (Masset & Gershman, in press; Niv, 2009; O'Doherty et al., 2017). However, the operations of this system appear to be augmented in humans, such that it closely interacts with more recently evolved attention, memory, and cognitive-control systems that facilitate flexible and rapid adaptation across a wide range of situations (Bornstein et al., 2017; Bornstein & Norman, 2017; Collins & Frank, 2012, 2013; Collins & Shenhav, 2022; Daw et al., 2011; Gershman & Daw, 2017; Hitchcock & Frank, 2024; Leong et al., 2017; Molinaro & Collins, 2023; Niv, 2019; Niv et al., 2015; Rmus et al., 2021; Yoo & Collins, 2022).

A noteworthy example of such interaction is that of the RL and working memory (WM) systems, which have been investigated via an RL-WM task manipulating WM load and delay in the context of trial-and-error learning (Collins & Frank, 2012). Behavioral, genetic, functional imaging, electroencephalography, computational modeling, pharmacologic, and individual-difference research with this task has led to a variety of insights, including that there are distinct behavioral and neural signatures of each system (Collins, 2018; Collins & Frank, 2012, 2018; Rac-Lubashevsky et al., 2023; Westbrook et al., 2024). The task has allowed learning impairments in patient populations to be disentangled (e.g., Cheng et al., 2024; Rutherford et al., 2023). For instance, people with schizophrenia had often been thought to exhibit RL deficits, which can in fact be attributed mostly to the WM system (Collins, Albrecht, et al., 2017; Collins et al., 2014).

Notably, this research has identified a key trade-off between WM and RL. While learning is strongly facilitated when it can be accomplished by WM, it is also less durable in such cases (Collins, 2018; Collins, Albrecht, et al., 2017; Collins, Ciullo, et al., 2017; Rac-Lubashevsky et al., 2023). This leads to a striking paradoxical effect. Namely, when participants learn under low WM demand—and thus show high performance via WM rapidly storing the correct options—their retention in a later test phase is actually markedly degraded. In contrast, when they learn under demands exceeding WM capacity, and RL is thus strongly enlisted to “pitch in” during initial learning, performance is initially impeded due to relatively lower WM contributions, but later retention is improved due to the RL system having been enlisted into the learning process (Collins, 2018; Collins, Albrecht, et al., 2017; Collins, Ciullo, et al., 2017; Rac-Lubashevsky et al., 2023). In short, the past pattern of both behavioral and neural findings with the RL-WM task suggests that WM facilitates fast acquisition, yet at the cost of undermining durable RL (Yoo & Collins, 2022). Yet, these findings appear difficult to reconcile with other research suggesting that WM and top-down processes actually *enhance* RL (Cavanagh et al., 2010; Daniel et al., 2020; Doll et al., 2009, 2011; Farashahi et al., 2017; Frank & Claus, 2006; Geana et al., 2022; Gold et al., 2012; Hernaes et al., 2018, 2019; Hitchcock, Forman, et al., 2022; Hitchcock & Frank, 2024; Leong et al., 2017; Radulescu et al., 2016; see also Collins, 2024; Miller et al., 2019).

Another open question is how WM and RL interact in the ubiquitous situation where some actions lead to reward and others to punishment. Much RL research has found that this system neglects punishment compared to reward (Collins, Ciullo, et al., 2017; Collins et al., 2014; Gershman, 2016; Master et al., 2020; Palminteri, 2023; Palminteri & Lebreton, 2022; see also Collins, 2024). Yet, what role WM plays in this process is unclear, because no past work (of which we are aware) employed a task wherein RL and WM contributions could be disentangled when concurrently learning to attain reward and avoid punishment.

Disentangling these contributions might have important implications for mental health. A recent simulation meta-analysis found that individuals with depression and/or anxiety disorders (vs. healthy controls) showed an apparently elevated RL punishment learning rate (Pike & Robinson, 2022). If this elevation is due to an alteration within the canonical RL system, this finding could have crucial treatment implications. By analogy, the gradual denervation of midbrain dopaminergic neurons in Parkinson's disorder leads to an asymmetry in RL from negative rather than positive prediction errors, which is related to alterations in striatal reward prediction error (RPE) signaling and can be partially remediated by dopaminergic medication (Frank et al., 2004). Pike and Robinson's (2022) findings suggest a similar core learning difference might lie at the root of depression and anxiety.

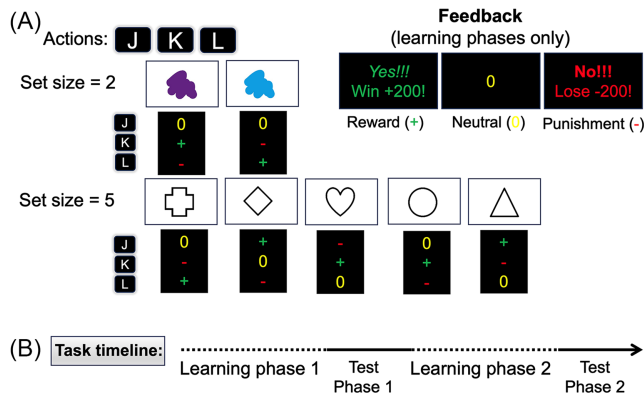
Yet, as noted, past work with the RL-WM task showed that trial-and-error learning differences in schizophrenia, which might appear to arise from RL, can in fact be attributed to WM (Collins, Albrecht, et al., 2017; Collins et al., 2014). This raises the question of whether apparent RL-based differences in punishment learning might instead arise from differences in WM allocation. Indeed, it is quite plausible that people with depression and anxiety symptoms preferentially allocate WM to punishment. These symptoms correlate extremely highly with trait neuroticism (Griffith et al., 2010), which is the disposition to react negatively to threat, frustration, or loss (Lahey, 2009).

Differences in WM allocation might have quite different treatment implications than differences in RL. For instance, the tendency to become explicitly preoccupied by threat, frustration, or loss might be readily targeted in psychotherapy, rather than necessarily requiring treatments (e.g., pharmacological ones) targeting the (primarily subcortical) canonical RL system directly (indeed, see Brown et al., 2021, for evidence that cognitive-behavioral therapy influences loss processing during trial-and-error learning—there, a bias in negative-outcome valuation).

One key innovation in the RL-WM task has been the incorporation of a test (retention) phase that is administered after a substantial break from initial learning. At this point, WM representations should have long since decayed. Thus, the test phase offers a relatively pure assessment of what has been retained by RL (Collins, 2018; Rac-Lubashevsky et al., 2023). Here, we leverage the combination of learning and test phases in a design with not only rewards but also punishments. This enables testing whether punishment avoidance during learning (and any differences as a function of psychiatric symptoms therein) is actually due to RL, based on whether they persist into the test phase.

We designed a novel approach/avoidance RL-WM task (Figure 1) to address the above open questions and administered it to a

**Figure 1**  
The Rew-Pun RL-WM Task



**Note.** (A) As in the standard RL-WM task, participants had to learn stimulus-action-outcome contingencies in blocks with varying numbers of unique stimuli, thereby manipulating set size and thus WM demand. Participants tried to maximize reward and minimize loss by selecting the best possible action (the Keys J, K, or L) in response to each image. Whereas, in the standard task, for each image one action was rewarding and the other two actions were nonrewarding, in our version, one action led to reward, another to neutral, and the last to punishment (image-action-outcome mappings were randomized, as shown). As in the standard task, these mappings were deterministic and stationary. This figure shows the feedback screens from the learning phases of the task; participants also completed test phases, which had the same structure but with feedback withheld. (B) Task timeline: Participants completed half of all learning trials, underwent a surprise test phase, completed the second half of learning trials, and underwent a second surprise test phase. RL = reinforcement learning; WM = working memory. See the online article for the color version of this figure.

relatively large online sample ( $N = 298$ ;  $n = 275$  after quality control),<sup>1</sup> who were oversampled for depression and anxiety symptoms (see the Method section). We hypothesized that individuals with depression/anxiety symptoms would show heightened punishment avoidance during initial acquisition, which would be attributable to WM, but not during a later test phase when information is no longer in WM and performance thus relies primarily on RL. Given our prediction that symptomatic participants would preferentially allocate WM resources toward avoiding punishment, we predicted—due to the aforementioned findings that higher WM engagement can paradoxically undermine RL—that they would actually show worse long-term retention of punishment contingencies (potentially forging a link to stress generation in internalizing disorders, which could arise from a systematic decrement in adaptive punishment RL; see also Beltzer et al., 2023; Conway et al., 2012; Hitchcock, Forman, et al., 2022; Reilly et al., 2019; Whitmer et al., 2012; Wierenga et al., 2022). Moreover, given poor performance on executive function tasks in depression and anxiety disorders (Abramovitch et al., 2021; Grahek et al., 2019; Snyder, 2013), we predicted decreased overall WM contribution to learning as a function of depression and anxiety symptoms. We also assessed whether highly ruminative individuals would preferentially allocate WM to punishing contingencies (see also Sharp et al., 2022), as well as show lower overall WM contribution due to off-task rumination (Hitchcock, Forman, et al., 2022; Rutherford et al., 2023; Whitmer et al., 2012).

## Method

### Participants

Participants from the United States ages 18–65 ( $N = 298$ ) were recruited via Prolific and provided informed consent via a form approved by the Brown University Institutional Review Board. Participants were compensated \$9.50 United States dollars/hr. Following consent, they completed the Rew-Pun RL-WM Task followed by questionnaires.

In quality control checks similar to those used in prior work (Hitchcock, Forman, et al., 2022; Hitchcock & Frank, 2024), we excluded participants who, during the learning phase, performed worse than 65% in Set Size 1 and/or 50% of Set Size 2 trials and/or had more than 10% invalid trials (i.e., timeouts or wrong keys; described below). We also excluded one participant who failed two attention checks embedded in posttask self-reports (see below). After these exclusions, the sample size was  $n = 275$ . The demographic composition of participants (via their demographics included in their Prolific profile) was as follows: age  $M (SD) = 38.48 (11.42)$ ; 50.36% female, 48.19% male, <1% data expired, and 1.09% preferred not to say; and 4.35% Asian, 9.42% Black, 1.45% data expired, 6.89% mixed, 4.35% other, and 73.55% White.

We oversampled for depression and anxiety symptoms using the following strategy: One Prolific substudy was run in which participants were eligible if they had answered “Yes” to the item “Do you experience anxiety?” on their Prolific profile; another was run where they were eligible if they had answered “Yes” to the item “Do you experience depression?”; a final was run with an unselected sample (after quality control,  $n_s = 73, 75$ , and 127 [ $n = 275$  total] were retained from each substudy). No other symptom-related eligibility criteria were imposed on the subgroups. For instance, the unselected sample was truly unselected and was not required to have answered “No” to the depression or anxiety questions. Supplemental Figure S1 confirms that this simple oversampling strategy led to greater depression and anxiety symptoms in the substudies comprising participants who ticked those boxes. As described below, statistically, we took a continuous/dimensional approach to analyzing psychiatric symptoms; hence, the purpose of plotting these distributions is simply to show that our strategy indeed oversampled depression and anxiety relative to unselected sampling; no subsequent analyses use these groupings or differentiate by substudy. Descriptive statistics for symptoms were as follows: depression symptoms (via Beck’s Depression Inventory–II):  $M (SD) = 14.21 (13.96)$ , range = 0–58; anxiety symptoms (via Generalized Anxiety Disorder–7):  $M (SD) = 5.86 (5.78)$ , range = 0–21.

<sup>1</sup> Our sample size was in line with our target of approximately 300 participants. Given our new version of the RL-WM task focused on reward pursuit/punishment avoidance and that we would need to develop a new variant of the task’s computational model to capture punishment avoidance (see below), it was not possible to estimate an effect size for punishment avoidance individual differences from the literature to use for power analysis. Instead, we sought a sample size that was larger than most past studies of reward and punishment processing in depression/anxiety (e.g., Pike & Robinson, 2022, analyzed 27 such studies with  $N = 3,085$  total participants—an average of  $n = 114.26$  participants per study) and much larger than most past studies of the RL-WM task.

## The Reward–Punishment RL and WM Task

Like the standard RL-WM, the task was a deterministic trial-and-error learning task wherein WM demand was systematically manipulated by varying the set size from 1 to 5 in different blocks (Figure 1A depicts a set size of 2, where participants had to learn adaptive actions in response to only two images in that block, vs. a set size of 5, where they had to learn adaptive actions in response to five images; image set to set size assignments were randomized). Participants were fully instructed about the task, including that actions would deterministically lead to reward, neutral, or punishment for different images and that their goal was to “GAIN reward and AVOID loss.”

Before each block, participants viewed a 7-s display showing the images that they would learn about that block. Then the block began—comprising a sequence of trials in which an image was presented, the participant responded with an action (they pressed the Key J, K, or L), feedback was displayed on a subsequent screen, and then a fixation cross was displayed for 500 ms. However, unlike the standard RL-WM task—where one action led to reward and the other two to nonreward—here each action led to reward, neutral, or punishment feedback (feedback displays are shown in Figure 1A). The addition of a demand to avoid loss was expected to increase the effective WM load for a given set size compared to the standard task (given that participants might allocate WM resources to avoid punishment and not just pursue reward). Thus, we used a set size range of 1 to 5, whereas past studies with the standard task have sometimes used a smallest set size of 2 or 3 and/or a highest set size of 6 (e.g., Collins, 2018; Collins, Ciullo, et al., 2017).

Participants completed 300 learning trials, which comprised two different sets of 1 through 5 set size stimuli, with each stimulus learned about 10 times ( $[1 + 2 + 3 + 4 + 5] \text{ images} = 15 \times 2 \text{ stimulus sets} \times 10 \text{ iterations with each image} = 300 \text{ learning trials}$ ). The primary purpose of having two stimulus sets (i.e., two sets of Set Sizes 1 through 5) was to increase the number of trials per subject; the sets did not differ in any way, and analyses and figures below do not distinguish between sets (for instance, performance at Set Size 5 reflects average performance in both stimulus sets).

Participants also completed 120 test trials, which comprised four iterations with each image from each stimulus set ( $[1 + 2 + 3 + 4 + 5] \text{ images} = 15 \times 2 \text{ stimulus sets} \times 4 \text{ iterations} = 120 \text{ test trials}$ ). Test trials came as a surprise; that is, participants were not informed that they would later be tested on the stimuli when they initially learned about them. The test trials had the same structure as the learning trials except that feedback was withheld: After the participant responded with an action, a blank screen was shown for 400 ms followed by a 500-ms fixation cross. (Prior to test trials, participants were told to “Try to remember the best key for each image. ... But don’t worry if you can’t consciously remember—many people feel that and still perform fine! Just PAY ATTENTION to each image and TRY YOUR BEST to select the correct—and avoid the worst—key!”)

The task timeline (Figure 1B) was as follows: Participants first completed half of the learning trials (five iterations of learning each for all images—i.e., those from all set sizes in both stimulus sets), they then completed a surprise test phase (two iterations of testing on all images), they then resumed learning (a final five iterations of learning for each image), and finally, they completed a second surprise test phase (a final two iterations of testing on each image).

The purpose of the test phases was to examine what had been durably learned by the RL system, after many intervening trials since initially learning about those stimuli, thereby ensuring that short-lived and interference-prone WM representations were no longer available. This goal would be undermined if, for example, a participant learned about the Set Size 3 stimuli in Stimulus Set 2 at the end of the first learning phase and then tested on those same stimuli only shortly after at the beginning of the first test phase. Thus, we ensured that the order of stimulus sets, as well as the order of the set sizes within them, was preserved across phases. Thus, in this example, the participant would first test on the entire set of the Stimulus Set 1 stimuli (completing 1 through 5 set size blocks in the same order as during prior learning) and then, once in Stimulus Set 2, would not test on its Set Size 3 images until the end.

In both learning and test phases, stimuli appeared for 1,400 ms followed by a timeout message (“RESPOND FASTER!”) if there was no response or an “INVALID RESPONSE | USE J K or L!” message if the wrong key was entered. In either case, the trial was repeated until a valid response was given.

Of note, we developed this task variant, in which two distinct test phases were employed, while planning the experiment by simulating several different potential experimental designs via an adapted RL-WM computational model (Collins & Frank, 2012; Master et al., 2020). The simulations showed this design increased simulated agents’ error rates, chiefly because simulated agents completed Test Phase 1 and then began Learning Phase 2 at a point when (according to the model) WM representations have decayed, leading them to rely only on the incremental RL system, which had not yet many opportunities to learn. As described in the Results section, empirically we indeed found a precipitous increase in error at these points (as predicted by the model)—which was desirable given our interest in examining neutral preference (i.e., punishment avoidance) within the subset of error trials (see the Behavioral Measures section below).

Prior to completing the task itself, participants viewed practice instructions to learn how the task worked (what keys to press; that feedback was deterministic and rewarding, neutral, or punishing, etc.). They then completed a practice block, which had the same structure as a Set Size 2 block; to move on to the main task, they had to demonstrate their understanding of the task by selecting the rewarding action in 80% of trials in this block. The practice-block data were not analyzed.

## Posttask Self-Reports

After completing the tasks, participants completed the following self-report questionnaires.

### *Beck’s Depression Inventory–II*

Depression symptoms were measured via the Beck’s Depression Inventory–II (Beck et al., 1996), which is a well-validated 21-item assessment comprising depression symptoms. Symptoms are rated from 0 to 3, where higher numbers reflect more severe symptoms. Cronbach’s  $\alpha$  was .96 for this measure in our sample.

### *Generalized Anxiety Disorder–7*

Anxiety symptoms were measured via the Generalized Anxiety Disorder–7 (Spitzer et al., 2006), which is a well-validated seven-item



assessment concerning the frequency with which symptoms of generalized anxiety disorder occurred over the past 2 weeks (0 = *not at all*, 4 = *nearly every day*). Cronbach's  $\alpha$  was .94 for this measure in our sample.

### ***Rumination and Response Scale–Short Form***

Trait rumination was measured via the short-form version of the Response Styles Questionnaire (Nolen-Hoeksema & Morrow, 1991), which excludes items from the original questionnaire that concern depression symptoms. The Rumination and Response Scale–Short Form has 10 items that concern the frequency of engaging in brooding or reflective pondering (1 = *I almost never respond this way* to 10 = *I almost always respond this way*). Cronbach's  $\alpha$  was .90 for this measure in our sample.

### ***Measuring Depression/Anxiety Symptoms***

Given our expectation that depression and anxiety (which were correlated at  $r = .80$  in our sample) would lead to similar decrements in WM usage and greater avoidance of punishing over neutral actions (see also Pike & Robinson, 2022), we created a single depression–anxiety outcome measure by taking the sum of Beck's Depression Inventory–II and Generalized Anxiety Disorder–7  $z$ -scores divided by 2 ( $z$ -scoring ensured that each measure contributed equally to the sum). This strategy is consistent with our aim of employing a depression/anxiety measure reflecting symptoms that encompass these constructs, rather than only what is shared between them (as would be reflected in a factor on which items jointly load derived from factor analysis or a shared trait measure such as neuroticism).

### ***Attention Checks***

To encourage careful responding to questionnaires and attempt to flag careless/insufficient effort responders (Zorowitz et al., 2023), we informed participants prior to beginning self-reports that two attention-check items had been included, and we embedded in two questionnaires an attention check in which participants were instructed to select a specific response for one question (“not at all” and “almost never”; however, see Zorowitz et al., 2023, and limitations in the Discussion section regarding recent evidence that this approach is not as sensitive as others).

## **Analyses**

### ***Behavioral Measures***

In this article, “proportion correct” refers to the proportion of trials in which the rewarding (i.e., optimal) action was selected. “Neutral preference” refers to—within the subset of error trials—the proportion of times the neutral action was selected minus the number of times the worst (i.e., punishing) action was selected (hence, a value of 0 would correspond to no neutral preference). Hence, it gives a measure of punishment avoidance.

### ***Statistical Data Analyses***

Data were cleaned, organized, and analyzed using R (Version 4.3.1; R Core Team, 2023). To analyze the effects of set size on

proportion correct and reaction time in the learning and test phases, we used logistic and linear mixed-effects models with linear terms to capture linear effects and orthogonal polynomial terms (Mirman, 2014) to statistically test an inverted U effect, using the lme4 package (Bates et al., 2014), with  $p$  values estimated via Satterthwaite's method for approximating degrees of freedom. The variance inflation factor for all multivariate regression models was  $<1.7$ , suggesting no issues with collinearity. To allow interpreting estimates and the uncertainty in them (irrespective of whether they were statistically significant), we ran Bayesian mixed-effects regression models to examine the relationships between depression/anxiety and rumination symptom scores and behavioral/computational model-derived task metrics (and when relating computational model-derived metrics to task behavioral measures, for comparison). To estimate the relationship between trial-wise behavioral measures and symptoms (see Supplemental Figure S7), we used hierarchical Bayesian mixed-effects regression models estimated via the “brms” package in R (Version 2.21.0; Bürkner, 2017) using default priors and using five Markov Chain Monte Carlo chains with 4,000 samples each, 2,000 of which were warm-up samples, thus giving 10,000 samples for inferences; more samples were run in if the model did not converge with a lower number. To estimate the relationship between task parameters and symptoms and behavioral summaries (with one data point per participant), we used the “stan\_glm” function from the “rstanarm” package in R (Version 2.32.1; Carpenter et al., 2017) with default priors. No divergence, maximum tree depth reached, or estimated Bayesian fraction of missing information warnings were generated for any model; maximum tree depth was left at its default setting of 10 within the function.  $\hat{R}$  for all Bayesian models used for inference was below 1.1, suggesting no issues with convergence. Significance was defined by whether 90% of samples were above/below 0.

We used “maximal” mixed-effect model structures (Barr et al., 2013) when these converged and were not singular; when they were, we stepwise reduced the model and/or used hierarchical Bayesian mixed-effects regression models as alternatives to frequentist ones. The code notebook “key-results-paper.html” in the GitHub repository accompanying this article (<https://github.com/peter-hitchcock/rlwm-rew-pun-analysis>) displays all regression model specifications. All predictors in regression models were  $z$ -scored.

### ***Computational Modeling***

Building on past modeling with the standard RL-WM task (Collins, 2018; Collins & Frank, 2012; Master et al., 2020; Rac-Lubashevsky et al., 2023; Westbrook et al., 2024), we developed a computational model that was able to capture all key features of the data.

The full model is described in the Supplemental Equations. Here, we focus on elements that are key for understanding the results. The model comprised WM and RL systems that each represented action values, with RL values learned as,

$$Q(s, a)_{t+1}^{\text{RL}} = Q(s, a)_t^{\text{RL}} + \alpha_t \delta_t, \text{ where} \quad (1)$$

$$\delta_t = r_t - Q(s, a)_t^{\text{RL}},$$

where  $Q(s, a)_t^{\text{RL}}$  is the RL system's value for a specific action for a given image at trial  $t$ ,  $\delta_t$  is the RPE of that trial, and  $r_t$  is the reward of that trial (which could take values  $-1$  for punishment, 0 for neutral,

$$Q(s, a)^{\text{WM}'} = \begin{cases} \text{non-pun}_{\text{bonus}}, & \text{for the chosen action} & \text{if } r_t = 0 \\ Q(s, a)^{\text{WM}} + \text{non-pun}_{\text{bonus}}, & \text{for all non-chosen actions} & \text{if } r_t = -1. \\ Q(s, a)^{\text{WM}} - \text{non-pun}_{\text{bonus}}, & \text{for all non-chosen actions} & \text{if } r_t = 1 \end{cases} \quad (4)$$

or 1 for reward). The RL rate  $\alpha_t$  on a given trial varied depending on whether the RPE was positive or negative—specifically, it was respectively proportional to separately estimated parameters  $\alpha^+$  or  $\alpha^-$  in these cases. For all trials except the first stimulus iteration in each phase (at which point there should be no WM load), the learning rate could be further blunted on a given trial by an  $\text{RL}^{\text{off}}$  parameter,

$$\alpha_t = (1 - \text{RL}_t^{\text{off}}) \times \alpha_t, \quad (2)$$

where

$$\text{RL}_t^{\text{off}} = \begin{cases} \text{RL}^{\text{off}} & \text{if } \text{delay}_t = 0 \\ \frac{\text{RL}^{\text{off}}}{\text{delay}_t} & \text{otherwise} \end{cases}, \quad (3)$$

where  $\text{delay}_t$  refers to the number of images that intervened since the image had last been encountered. In short, the effective learning rate not only varied depending on the sign of the RPE but was also decreased by an  $\text{RL}^{\text{off}}$  parameter that varied between participants; this parameter itself varied trial-wise so that it exerted less blunting effect on trials with a higher delay between trials, given that higher delay should make WM representations weaker and less accessible, and thus the RL system should be more likely to be enlisted.<sup>2</sup> (Of note, delays tended to be longer at higher set sizes, given that trial images were randomly ordered; hence, the effective RL rate tended to be higher at set sizes that tended to exceed WM capacity—consistent with neural evidence that RL  $Q$  values increase more rapidly in higher set sizes, Collins & Frank, 2018; Rac-Lubashevsky et al., 2023, predicting better test-phase performance, Rac-Lubashevsky et al., 2023.)

Unlike the RL system that incrementally learned via a low learning rate, the WM system stored recent outcomes, which then quickly decayed; participants also varied in WM capacity and usage (see Supplemental Equations).

To enable punishment avoidance via the WM system, we also modeled the WM system as ascribing a bonus to actions that could be inferred to be nonpunishing,

(see Equation 4 above)

That is, when an action led to a neutral outcome, it was given a bonus value because at least it was not punishing; when an action was punishing, the remaining two actions were given a bonus because they could be inferred not to be punishing; and when an action led to a reward outcome, the remaining two actions were decreased in value because they might be punishing, thereby leading to a relative increase in the rewarding value. Overall, this rule allows the WM system to eschew punishing actions, including those that have not been directly experienced but which can be inferred from the task structure—consistent with evidence for updating even nonchosen options in RL tasks (Ben-Artzi et al., 2023; Biderman et al., 2023; Biderman & Shohamy, 2021).

Of note, as is standard in this task and as described in the Supplemental Equations, the inverse temperature parameter partly

determining the stochasticity of choice (in combination with an  $\epsilon$  parameter modeling attention lapsing/fully random choice; see Supplemental Equations) was fixed at a high value ( $\beta = 100$ , reflecting highly deterministic choice in this task with deterministic contingencies). In a model where this parameter was allowed to vary as a free parameter, we confirmed that it fit at high values for all subjects ( $Mdn = 99.99$ , range = 48.54–100) and that model fit was worse for the model where this parameter was estimated rather than fixed ( $\Delta\text{AIC} = 537.84$ , where AIC is the Akaike information criterion, Akaike, 1998, a model-comparison metric that penalizes for model complexity, which was used given that this model had an extra free parameter).

### Model Fitting, Validation, Parameter Recovery Procedures, and Modeling Strategy

We estimated model parameters using maximum likelihood estimation, which allowed for adequate to high parameter recovery (range = .67–.90 for the eight parameters in our model; Supplemental Figure S2; of note, the recovered  $\text{RL}^{\text{off}}$  parameter was correctly below/above its median value in 78% of cases). More specifically, we minimized the sum of negative log likelihoods of each participant's empirical choices across both the learning and test phases using the “solnp” function in the “Rsolnp” package in R (Ghalanos & Theussl, 2011). Test-phase choices were modeled by simply using the softmax-predicted choice probabilities based only on the corresponding RL  $Q$  values up to that point (i.e., the RL  $Q$  values after the first five learning iterations per image for Test Phase 1 and after all learning experiences for the final test phase; see “Supplemental Equations”; also see Frank et al., 2007, and related studies for fitting to test). Because there was no feedback in the test phase, no further updates to the  $Q$  values were made. Because the optimizer can sometimes find local minima, we ran optimization 40 times and took the run with the lowest negative log likelihood.

We validated our computational model's predictions against the empirical data by simulating the task with the inferred parameter estimates from the optimization procedure; specifically, we ran 50 iterations simulating the full experimental data set (i.e., each participant's best fitting parameters were used to simulate their experimental data, for all participants, and this was repeated 50 times; the simulation plots in the Results section show average values from this procedure as well as points representing individual simulation iterations). We statistically tested whether simulations could capture an inverted U effect observed empirically in the first test phase and first learning trial after test (described below) by running the same

<sup>2</sup> A cooperative model (as implemented in Collins, 2018; Collins & Frank, 2018)—in which WM contributed to the RL reward prediction error proportional to a free parameter  $\eta$ , such that  $\delta_t = r_t - [\eta Q(s, a)_t^{\text{RL}} + (1 - \eta) Q(s, a)_t^{\text{WM}}]$  (in lieu of scaling of  $\alpha$  and with no modulation proportional to delay)—that had the same number of free parameters as the model described in the text provided a worse fit to the data ( $\Delta$  negative log likelihood = 395.26).

model with orthogonal polynomial terms (Mirman, 2014) that we used to test for an inverted U effect in the empirical data (as described above) and finding the proportion of simulated data sets in which there was a statistically significant negative estimate for the quadratic term (providing evidence for an inverted U).

We tested how well we were able to recover parameters from behavioral data by drawing parameter estimates from a multivariate Gaussian distribution with parameter estimates simulated based on their empirically estimated values, following the simulation and then maximum likelihood estimation procedures just described, and then correlating the values estimated on the simulation data against the ones that had truly generated them.

Given extensive past research on the RL-WM task, we followed the same modeling strategy as other applied studies employing it (e.g., Cheng et al., 2024; Master et al., 2020; Rac-Lubashevsky et al., 2023; Westbrook et al., 2024). Namely, we did not conduct extensive model comparison to establish basic effects (such as that both WM and RL modules were needed, which is already well established from past work, e.g., Collins, 2018; Collins & Frank, 2012; Collins & Frank, 2018); instead, we focused on targeted comparisons and adaptations to the new design. Specifically, the Results section will demonstrate how variation in an RL-blunting parameter captures key individual differences among participants; show the mis-specification of models that do not include a non-pun<sub>bonus</sub> parameter, use the same RL rate for negative prediction errors (PEs) as for positive PEs, and use a stimulus-response (S-R) rule instead of RL rule (see below); and demonstrate worse model fit for a model that uses an alternative cooperative RL rule rather than the blunting rule described above. Moreover, our model includes free parameters—representing weightings on specific modules (such as a parameter dictating the extent of WM contribution) or the contribution of specific processes (such as the extent to which the RL system was blunted under WM load)—that were fit with bounds that allow them to drop out of the model (resulting in no WM contribution or no RL blunting, in these examples). Supplemental Figure S3 depicts histograms of all estimated parameters and thereby visualizes for each parameter the proportion of participants for whom the parameter was at a value such that it dropped out or nearly dropped out of the model.

Of note, we took a dimensional approach to assessing psychiatric symptoms, rather than recruiting psychiatric groups (e.g., participants with major depressive disorder vs. healthy controls); hence, we examined how parameters continuously related to symptoms rather than conducting a comparative test on groups (cf. Bayesian *t* tests; Piray et al., 2019).

### Testing a S-R Rather Than RL System

We also compared our primary model to one inspired by recent work by Collins (2024), which found across six data sets with the RL-WM task, as well as a probabilistic version from McDougle and Collins (2021), that data in the learning phase of this task could be best explained by a combination of WM and an S-R learning mechanism that is more passive than RL. In particular, the S-R system functions to simply increase the probability of repeating the same response after taking an action in a given situation, irrespective of the outcome. Collins (2024) highlighted that the mixture of these systems can produce similar learning

curves to an RL-WM agent, because the WM system can guide the S-R system to ingrain adaptive actions (see also Miller et al., 2019; and similar to how top-down influences are thought to accelerate RL-based learning in aforementioned studies and models, e.g., Cavanagh et al., 2010; Frank & Claus, 2006; Geana et al., 2022; Gold et al., 2012; Hernaus et al., 2018, 2019; Hitchcock & Frank, 2024). These latter models, ranging from biologically detailed neural networks to algorithmic models, assume that learning is a product of multiple systems, including WM, RL, and S-R learning. Collins (2024), however, found that there was no need for any value-based RL at all when fitting data from the RL-WM task.

Yet, Collins (2024) focused only on the learning phase, wherein it may be difficult to arbitrate between RL versus S-R contributions, given that both are thought to contribute, but where behavior is largely dominated by WM. In contrast, the test phase is thought to be largely independent of WM, because test choices require responding to stimuli that had been learned across several blocks, after substantial delays and intervening stimuli, and the set of which far exceeds WM capacity. Several past studies also showed that test-phase performance in this and other RL tasks indexes the degree to which people learned preferences for probabilistically more-rewarded stimuli, even when such preferences could not be solved within WM or even S-R alone (e.g., Collins, Ciullo, et al., 2017; Doll et al., 2011, 2016; Frank et al., 2004, 2007; Rac-Lubashevsky et al., 2023). Hence, we took advantage of our design having substantial test-phase data (120 trials/participant) and compared the computational model described above to an adaptation of it based on Collins (2024). (Of note, our study was not originally designed to arbitrate between S-R and RL accounts, but the rich test-phase data in our task allowed for a serendipitous opportunity to do so; likewise, Collins, 2024, analyzed previously collected data sets that were not originally designed to arbitrate between the accounts.) In our adaptation, the RL system in Equation 1 was replaced with a type of choice kernel (CK) for the chosen action in a given situation,

$$\text{CK}(s, a)_{t+1} = \text{CK}(s, a)_t + \alpha_t^{\text{CK}} \delta_t, \text{ where} \quad (5)$$

$$\delta_t = 1 - \text{CK}(s, a)_t,$$

given that Collins's (2024) best model was one where the system that replaced RL was insensitive to outcome in terms of its prediction error. Collins's (2024) most successful model nevertheless let this system and the WM learning rate (that was otherwise set at 1) be downregulated for negative outcomes by a shared bias parameter; hence, in our adaptation, we let:

$$\alpha_t^{\text{WM}} = \begin{cases} 1 & \text{if } r_t = 1 \\ \text{bias} & \text{otherwise} \end{cases}, \quad (6)$$

$$Q(s, a)_{t+1}^{\text{WM}} = Q(s, a)_t^{\text{WM}} + \alpha_t^{\text{WM}} [r_t - Q(s, a)_t^{\text{WM}}], \quad (7)$$

$$\alpha_t^{\text{CK}} = \begin{cases} \alpha^{\text{CK}} & \text{if } r_t = 1 \\ \alpha^{\text{CK}} \times \text{bias} & \text{otherwise} \end{cases}. \quad (8)$$

The CK system's values were initialized at 1/3 to match Collins (2024; however, model fit was very similar when using 0 initializations, as in the RL system in our primary model). WM values were also initialized at 1/3, matching Collins (2024) and



our primary model. The CK system did not update during the test phases.

To offer a direct S-R analog to the aspect of our model that captured individual differences in paradoxical set size effects within our task (see the Results section), we let this system's learning rate be further downregulated by a parameter:

$$\alpha_t^{\text{CK}} = \alpha_t^{\text{CK}} \times (1 - \text{CK}_t^{\text{off}}), \quad (9)$$

which inversely scaled with the delay between trials via the rule described in Equation 3. This model therefore substituted three parameters ( $\alpha^{\text{CK}}$ , bias,  $\text{CK}^{\text{off}}$ ), all in the range [0, 1], for the two RL rates and  $\text{RL}^{\text{off}}$  parameter in our primary model. A mixture of the choice-kernel (S-R) system and WM then drove choice exactly analogous to how a mixture of RL and WM did in our primary model.

As with our primary model, to avoid local minima, we ran optimization 40 times for this model and took estimates from the run with the lowest negative log likelihood. We compared this model and our primary model, which were identical except for the changes just described, in terms of their ability to capture key features of the data (model validation) and model fit via negative log likelihood (no penalty term was added because both models had eight free parameters).

## Transparency and Openness

The code used to produce the results and figures is available at [github.com/peter-hitchcock/rlwm-rew-pun-analysis](https://github.com/peter-hitchcock/rlwm-rew-pun-analysis). Data that can be used to reproduce the results is available in the "public-data" folder within the repository. The task was coded in Honeycomb (Provenza et al., 2022). The study was not preregistered.

## Results

### Accuracy and Reaction Time Parametrically Vary With Set Size During Learning

We replicated a parametric effect of set size during learning that has been well-established through the standard task, namely that participants are slower to acquire the optimal rewarding action in higher set sizes (Figure 2A, left—white region)—Phase 1: set size  $\beta$  ( $SE$ ) =  $-.540$  (.021), stimulus iteration  $\beta$  ( $SE$ ) =  $1.07$  (.024), Set Size  $\times$  Stimulus Iteration  $\beta$  ( $SE$ ) =  $-.372$  (.018), all  $ps < 2e-16$ . Novel to our task was that, after five iterations learning about each stimulus, a break was taken for testing on all stimuli, after which learning resumed (see the Method section). Notably, there was a precipitous decline in performance in the lower set sizes when learning resumed, whereas the break had less detrimental performance in higher set sizes, consistent with the notion that WM was responsible for superior performance in lower set sizes, yet was no longer available after the break (Figure 2A, left—change from Stimulus Iterations 5 to 6)—set size  $\beta$  ( $SE$ ) =  $-.291$  (.029), before versus after break regressor  $\beta$  ( $SE$ ) =  $-.522$  (.023), Set Size  $\times$  Before Versus After Break  $\beta$  ( $SE$ ) =  $.373$  (.023), all  $ps < 2e-16$ ; here, the first two terms capture deleterious main effects of set size and the break on performance, respectively, whereas the interaction indicates that the postbreak decline disproportionately influenced the lowest set sizes (see Supplemental Figure S4 for plot of regression predictions).

After the break, participants had five more opportunities with each stimulus to learn with feedback, and the parametric effect of set size quickly reestablished (Figure 2A, left—gray-shaded region)—Phase 2: set size  $\beta$  ( $SE$ ) =  $-.287$  (.029), stimulus iteration  $\beta$  ( $SE$ ) =  $.699$  (.021), Set Size  $\times$  Stimulus Iteration  $\beta$  ( $SE$ ) =  $-.258$  (.020), all  $ps < 2e-16$ .

Simulations from our computational model of RL-WM interactions (Supplemental Equations; Figure 2B, right) captured these patterns, including the difference in learning slope by set size in Phase 1, the magnitude of decline at the start of Phase 2 (other than for Set Size 1 where empirically the decline was more precipitous), the reestablishment of the set size effect in Phase 2, and the asymptotic difference in performance between set sizes.

Consistent with models that assume more difficult choices take longer (e.g., to accumulate evidence toward a value-based decision, including one learned by RL and WM; see McDougale & Collins, 2021), reaction times showed the mirror-image parametric effect as proportion correct (Figure 2B, top left)—Phase 1: set size  $\beta$  ( $SE$ ) =  $88.066$  (2.12), stimulus iteration  $\beta$  ( $SE$ ) =  $-38.97$  (1.62), Set Size  $\times$  Stimulus Iteration  $\beta$  ( $SE$ ) =  $18.14$  (1.02), all  $ps < 2e-16$ ; Phase 2: set size  $\beta$  ( $SE$ ) =  $57.94$  (1.74), stimulus iteration  $\beta$  ( $SE$ ) =  $-40.31$  (1.13), Set Size  $\times$  Stimulus Iteration  $\beta$  ( $SE$ ) =  $19.61$  (.874), all  $ps < 2e-16$ . Notably, there was an effect of set size on reaction time even when the same stimulus was repeated across successive trials and a reward had been received on the last trial (Figure 2B, top right), set size  $\beta$  ( $SE$ ) =  $67.56$  (1.67),  $p < 2e-16$ , as well as on the first stimulus iteration of Phase 1 (Figure 2B, bottom left), set size  $\beta$  ( $SE$ ) =  $37.57$  (3.03),  $p < 2e-16$ . The latter is noteworthy because it occurs before any RL or WM values could have been acquired, and thus indicates choice speed is influenced by proactive strategies (e.g., raising the decision threshold when participants know they have more to learn; of note, participants saw the number of stimuli that they would learn about before each trial began; see the Method section). Strikingly, after the break, participants were much *slower* to respond to Set Size 1 (Figure 2B, bottom right) than the other set sizes, even though they had responded by far the fastest in this set size before the break—again consistent with them no longer having access to WM that had been available before the break and consistent with prior reports that participants show poor retention of S-R information from low set sizes when tested later (Rac-Lubashevsky et al., 2023).

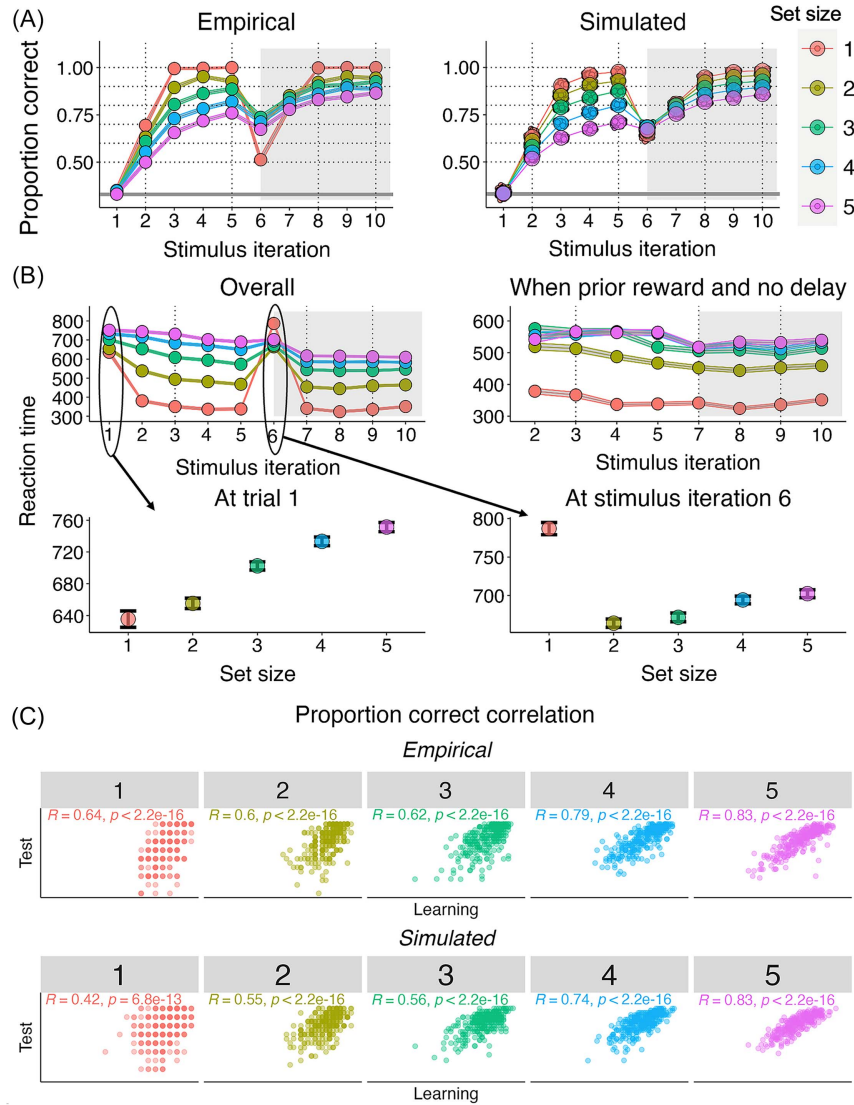
Because our RL-WM model assumes that low set sizes are more reliant on WM available during learning but not test, whereas high set sizes exceed WM capacity and hence require enlisting the durable RL system during learning, the model should predict a higher correlation between performance in the learning and test phases as a function of increasing set size. This prediction is indeed borne out empirically and recapitulated by the model, wherein the correlation between learning and test phase performance was higher in high set sizes (4 and 5) than low set sizes (1 and 2) in all 50 simulations (higher than expected by chance;  $p < 5e-15$ , binomial test; Figure 2C).

### Set Size Led to an Inverted U Effect on Retention Controlled by an RL-Off Parameter

Past work with the standard RL-WM task suggested that, when WM dominates learning in lower set sizes, it reduces prediction

**Figure 2**

*Learning Phases—Proportion Selecting the Correct (Rewarding) Action and Response Time*



*Note.* (A: Left) Empirical proportion correct (i.e., selecting the rewarding action) as a function of set size in Phases 1 and 2 (white and gray shading). Points represent means and shading  $\pm 1$  SEM. (Right) Simulated proportion correct from our computational model; large points represent means over 50 simulations and individual points 30 individual simulations (simulations were highly consistent with the average). (B: Top left) Response time as a function of set size and (top right) in trials when the current trial stimulus had been presented in the immediately preceding trial (“no delay”) and was rewarded. Points represent means and shading  $\pm 1$  SEM. (Bottom left and right) Focusing in on the pattern of reaction time at two key points: the beginning of the first and second phases. Points represent means and error bars  $\pm 1$  SEM. (C) The correlation between proportion correct in the learning and test phases as a function of set size, empirically and in model simulations (correlations are shown from a single simulation of the same size as the empirical data set, although the same pattern was evident across simulations, as described in the text). SEM = standard error of the mean. See the online article for the color version of this figure.

errors and thus blunts learning by the RL system (Collins, 2018; Rac-Lubashevsky et al., 2023). Strikingly, Rac-Lubashevsky et al. (2023) found that this led to a full reversal of the faster learning for lower set sizes seen in the learning phase, so that, during the test

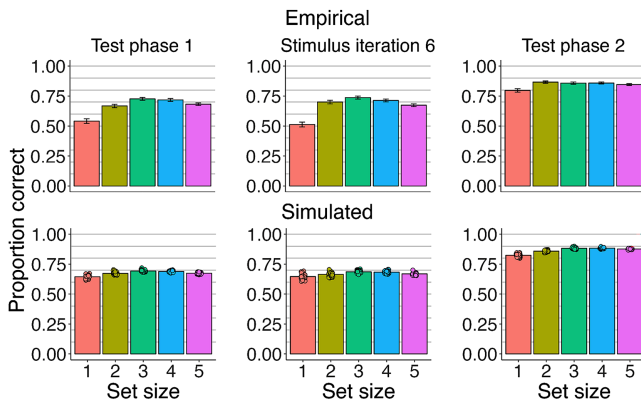
phase, performance actually parametrically improved as set size *increased*—the mirror image of the worse performance as a function of set size seen during learning. In that study, participants completed a 15-stimulus-iteration learning phase (by which point performance

was nearly perfect across set sizes), followed, after a pause, by a single test phase. Here, in contrast, participants only completed 10 learning stimulus iterations total and they underwent a test phase midway through learning as well as at its end (i.e., after 5 and 10 stimulus iterations), thereby enabling us to analyze retention at these earlier versus later points.

In contrast to the full parametric reversal in Rac-Lubashevsky et al. (2023; see Figures 3A and 4C on pp. 3136–3137 in that article), we found an inverted U pattern in Test Phase 1: While performance indeed parametrically improved through Set Size 3, it then modestly declined again in Set Sizes 4 and 5 (Test Phase 1 in Figure 3, top left). This reversal was also seen not only in the test phase but also at Stimulus Iteration 6 during learning—the first reenounter with the stimulus after the break, which is replotted in Figure 3 (top middle) to demonstrate that a similar inverted U pattern appears at this point as well. Indeed, quadratic as well as linear orthogonal polynomial terms for set size were statistically significant in logistic mixed-effects models at these time points—Test Phase 1: linear set size  $\beta$  ( $SE$ ) = 13.42 (3.24),  $p < 5e-5$ ; quadratic  $\beta$  ( $SE$ ) = -26.24 (3.43),  $p < 5e-14$ ; Stimulus Iteration 6: linear set size  $\beta$  ( $SE$ ) = 6.44 (2.44),  $p < 1e-2$ ; quadratic  $\beta$  ( $SE$ ) = -22.81 (2.49),  $p < 2e-16$ . In contrast, there were no statistically significant linear or quadratic effects in Test Phase 2 across all subjects ( $ps > .21$ ). Plots of the regression models' predictions confirm that they captured an inverted U pattern in Test Phase 1 and Stimulus Iteration 6, whereas there was no clear pattern as a function of set size evident in Test Phase 2 (Supplemental Figure S5).

**Figure 3**

*Test Phases—Proportion Selecting the Correct Action as a Function of Set Size*



**Note.** Empirical (top; bars are means and error bars  $\pm 1$  standard error of the mean) and simulated from our computational model (bottom), where in the latter the bars are means across 50 simulations and the points are individual simulations. As noted in the text, learning with feedback paused after Stimulus Iteration 5 and Test Phase 1 then occurred sometime later, at which point WM representations should have decayed. Stimulus Iteration 6 is then the first trial in Phase 2, before the participant has begun learning again (hence, the pattern of results is similar to Test Phase 1). Learning then continued for another five stimulus iterations followed by another pause; Test Phase 2 then again assessed what has been retained, now after there had been 10 learning iterations with each stimulus. WM = working memory. See the online article for the color version of this figure.

Our computational model qualitatively captured the inverted U pattern, and 96% of Test Phase 1 simulations and 60% of Stimulus Iteration 6 simulations showed a statistically significant quadratic term with a negative sign reflecting an inverted U. However, we note that the model simulations were unable to capture the full magnitude of the inverted U at Test Phase 1 and Stimulus Iteration 6, as the empirical data showed both a steeper increase from Set Sizes 1:3 and decrease from Set Sizes 3:5 than was predicted by the model (Figure 3, bottom).

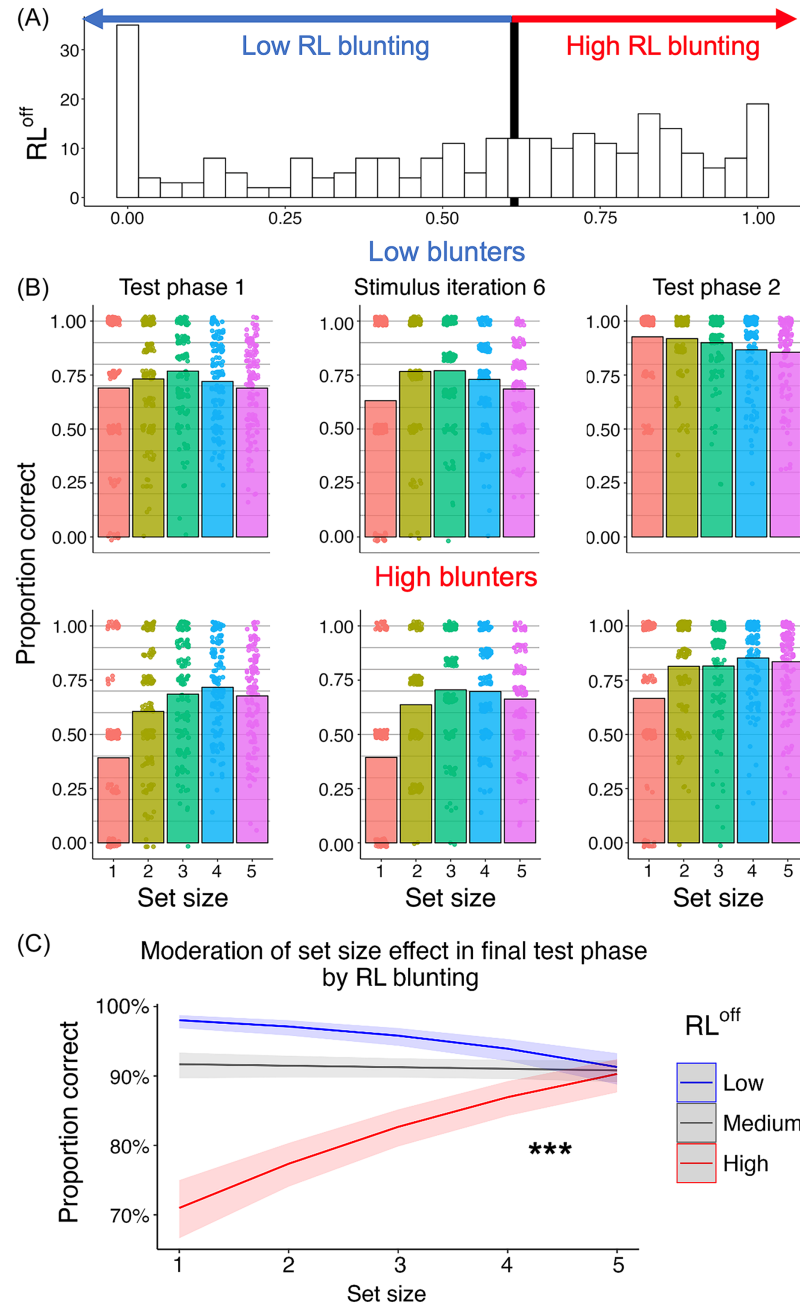
How is the model able to capture the inverted U pattern? Figure 4 demonstrates that the shape of the inverted U is modulated by the  $RL^{off}$  parameter, which captures individual differences in WM blunting of the RL system (see the Method section). In particular, low- versus high-blunting participants (defined for visualization by taking a median split on  $RL^{off}$ ) exhibited a qualitatively different pattern empirically. Low blunters showed, at Test Phase 1 and Stimulus Iteration 6, an inverted U pattern that peaked at Set Sizes 2 or 3 and then, in the final Test Phase 2, a parametric effect of declining performance with set size similar to the set size stratification at the end of learning (see Figure 2A). In contrast, high blunters showed a more severe decline in performance in the lower set sizes that persisted until Test Phase 2 (Figure 4B). Indeed, statistically,  $RL^{off}$  (entered as a continuous predictor) moderated the effect of set size on Test Phase 2 performance, producing a qualitatively different pattern—such that the lowest tercile of blunters ( $-1$  SD  $RL^{off}$ ) showed a parametric decline as a function of set size (i.e., the same direction as present at the end of learning), whereas the highest tercile of blunters ( $+1$  SD  $RL^{off}$ ) reversed that pattern—set size  $\beta$  ( $SE$ ) = -.035 (.041),  $p = .397$ ;  $RL^{off}$   $\beta$  ( $SE$ ) = -.541 (.092),  $p < 1e-8$ ; Set Size  $\times$   $RL^{off}$   $\beta$  ( $SE$ ) = .451 (.047),  $p < 2e-16$ . Thus, while Figure 3 had shown that, at the group level, there was no apparent effect of set size at Test Phase 2, the  $RL^{off}$  parameter reveals a notable source of heterogeneity, with subsets of participants exhibiting a qualitatively different pattern depending on this parameter that reflects individual differences in RL blunting.

### A Model With a Pure S-R, Rather Than RL, System Captures Some Key Effects but Misses Others

When we tested substituting an S-R learning module in place of value-based RL, as in Collins (2024), we found that this model provided a worse fit to the data than the value-based model ( $\Delta$  negative log likelihood = 500.08; as noted in the Method section, these models had the same number of parameters, so a penalization term was not needed to compare them). Notably, the retention phases were particularly diagnostic of the differences between the models. Indeed, while simulations from the S-R model were able to capture the learning curve similar to our primary model, they predicted largely the opposite pattern to the observed paradoxical effects in the retention phases in that they predicted parametrically decreasing performance as a function of set size (i.e., the same pattern as during learning) in Test Phase 1, Stimulus Iteration 6, and Test Phase 2 (Supplemental Figure S6A and S6C). That the S-R model is unable to capture the paradoxical effects is unsurprising: The S-R account predicts better retention of actions repeated most often during learning and thus parametrically better performance for correct S-Rs that had been repeated most consistently

**Figure 4**

Modulation of Test Phase Shape by a Model Parameter,  $RL^{off}$ , Controlling RL Blunting



*Note.* (A) The distribution of  $RL^{off}$  estimates across participants (the black bar shows the median value). (B) Empirical pattern of results for low- versus high-blunting participants, as per the median split on the  $RL^{off}$  parameter in A. Bars show the average across subjects, and points show the individual subjects (with width and height jitters of .17 and .02, respectively, added). (C) Logistic regression mixed-effects predictions in a model predicting proportion correct in the final test phase from the interaction between set size and  $RL^{off}$ . Predictions are shown at  $-1$  standard deviation (low), mean (medium), and  $+1$  standard deviation (high;  $RL^{off}$  was a continuous predictor in the model—these categories are only used for visualization). RL = reinforcement learning. See the online article for the color version of this figure.

\*\*\*  $p < .001$ .



(those learned under low WM demand).<sup>3</sup> The S-R model also predicted a lower rate of neutral preference during learning than was observed empirically (see the next section; Supplemental Figure S6D).

### Participants Exhibited a Short-Lived Preference for Neutral Actions Over Punishing Ones During Learning With Relatively Poor Retention Into a Test Phase

Our analyses have thus far focused on proportion correct (i.e., selection of the rewarding action), yet a key novel feature of our new task was that the other two actions led to different outcomes: punishment and neutral (which together constitute error trials; 25.03% of learning-phase and 22.99% of test-phase trials were error trials). As learning progressed in both phases, participants developed a tendency to select neutral over punishing actions (Figure 5A). We used Bayesian methods for statistical models of this effect due to convergence issues with frequentist ones; this analysis confirmed that the posterior density for stimulus iteration on neutral preference was above 0 on 100% and 92.75% of samples for Phases 1 and 2, respectively, reaching our criterion for significance. Our computational model captured this preference for neutral over punishment (Figure 5B), although with substantial variability across simulations (presumably because error trials were such a relatively small proportion of trials compared to those in which participants chose the rewarding action). Moreover, this preference was accounted for by a non-pun<sub>bonus</sub> parameter that assigns a bonus in WM to options that can be inferred to be nonpunishing (see the Method section); “lesioning” this parameter led to a reduction in the neutral preference inconsistent with the magnitude observed empirically (Figure 5C).

In the test phases, there was only marginal evidence for a neutral preference (Figure 6A, left), in that the posterior reflecting this preference (the intercept of the neutral preference during the test phase) had 10.27% of traces below 0 and thus did not meet our criterion for significance. It is also evident that the magnitude of the preference was lower than at its highest point during learning (compare Figure 6A, left, to Figure 5A) even in the second test phase, by which point all errors (choices of neutral or punishing stimuli) had been experienced (the 90% credibility interval of a test-phase regressor on neutral preference also overlapped 0, i.e., we did not find evidence that retention differed in the later vs. earlier phase:  $M = .04$ , 90% CI  $[-0.04, 0.12]$ ). Our computational model also predicted poor retention (Figure 6A, right). It did so, first, because the non-pun<sub>bonus</sub> parameter, which was key to capturing the learning-phase neutral preference (described above), operated only on WM and thus was not available during the test phases, and second, the RL system—that was available during the test phases—was fit with two different learning rates for learning from positive negative prediction errors,  $\alpha^-$  and  $\alpha^+$ , and for 93.82% of participants  $\alpha^+ > \alpha^-$  (Figure 6B, left; the parameters [on a log scale] also differed in a paired-samples  $t$  test;  $p < 2e-16$ ). Moreover, when we reran simulations with  $\alpha^-$  values set to those estimated for  $\alpha^+$ , the model’s predictions were substantially mis-specified—predicting a much greater neutral preference than was observed empirically (Figure 6B, right). Finally, model comparison supported dropping the  $\alpha^-$  parameter entirely ( $\Delta AIC = -279.57$  for a model with no RL updating at all after negative PEs). However, we have retained this

parameter in the model to allow for the direct comparisons between the different learning rates just described and given the importance of examining if individual differences in RL-based punishment learning relate to depression/anxiety symptoms in light of recent results implicating it therein (see the next section; Pike & Robinson, 2022).

### Little Evidence of Task Differences as a Function of Depression/Anxiety or Rumination

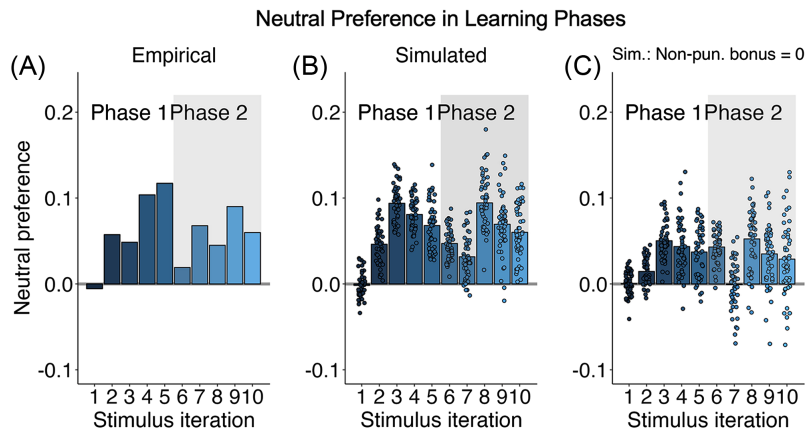
Contrary to our hypothesis, there were no consistent accuracy differences in the learning phase as a function of depression/anxiety or rumination symptoms (Supplemental Figures S7 and S8). In particular, the main effects of depression/anxiety and trait rumination, and their interaction with set size (which would reflect different performance as a function of WM demand), were mostly centered around 0 (Supplemental Figure 7B, left, top; Phase 1—depression/anxiety: posterior  $M < -.01$ , 90% CI  $[-0.04, 0.03]$ ; Depression/Anxiety  $\times$  Set Size:  $M < .01$ , 90% CI  $[-0.01, 0.02]$ ; rumination:  $M < -.01$ , 90% CI  $[-.04, .03]$ ; Rumination  $\times$  Set Size:  $M = .02$ , 90% CI  $[0.00, .03]$ ; Phase 2—depression/anxiety:  $M < -.01$ , 90% CI  $[-0.07, 0.08]$ ; Depression/Anxiety  $\times$  Set Size:  $M = .03$ , 90% CI  $[.00, .05]$ ; rumination:  $M < .01$ , 90% CI  $[-.08, .08]$ ; Rumination  $\times$  Set Size:  $M = .02$ , 90% CI  $[-0.01, .04]$ ). In fact, in the only two cases with some evidence, the effects indicated relatively *better* performance as a function of symptoms at higher set sizes (i.e., under higher WM demand)—the opposite of the predicted direction. These effects included marginal evidence for an interaction between rumination and set size in Phase 1 and between depression/anxiety and set size in Phase 2. However, the fact that these Symptom  $\times$  Set Size relationships were not repeated in the other phase is tantamount to a failed robustness check for this relationship. Altogether, given the weak evidence, multiple comparisons, opposite-of-predicted direction effects, and failed replication in the other phase, we interpret the overall pattern of results as providing no consistent evidence for a Symptom  $\times$  Set Size interaction. As expected, there were no accuracy differences as a function of symptoms in the test phase (Supplemental Figure S7B, left, bottom; depression/anxiety:  $M = .01$ , 90% CI  $[-0.07, 0.10]$ ; rumination:  $M < .01$ , 90% CI  $[-0.08, 0.09]$ ).

Contrary to our hypothesis, there was no greater punishment avoidance (operationalized as greater neutral [vs. punishment]

<sup>3</sup> As described in the Method section, in order to allow for the most direct comparison with our primary model, which included an RL blunting term that was key to capturing the test effects (as described above), we incorporated a blunting parameter serving the same function in the S-R model (although, from a theoretical perspective, we note that this parameter is not well motivated, in that it is inconsistent with the spirit of an S-R system that ingrains actions irrespective of contextual factors such as WM demand). In this implementation, a key reason that the SR-WM model was unable to capture test phase effects was that estimates for this parameter returned by optimization were very low ( $Mdn = .078$ ), whereas the RL blunting parameter fit at much higher values ( $Mdn = .614$ ); that is, the choice-kernel blunting parameter (unlike the RL-blunting parameter) “dropped out” or nearly dropped out of the model for many participants. In fact, an S-R model without the blunting parameter (CK<sup>off</sup>) improved fit relative to the more complex model ( $\Delta AIC = -107.28$ ), yet the reduced model still fit worse than the RL-based primary model ( $\Delta AIC = 892.88$ ). We nevertheless included CK<sup>off</sup> in the simulations in Supplemental Figure S6 to show that, even with this parameter included, the S-R model qualitatively mis-specified the predictions about the test phase.

**Figure 5**

*Neutral Preference (Within the Subset of Error Trials, Proportion Selecting the Neutral Action Minus Proportion Selecting the Punishing One) as Learning Progresses in Phases 1 and 2*



*Note.* (A) Empirical results, where bars are mean and error bars  $\pm 1$  standard error of the mean. (B) Model simulations, where the bars are average across 50 simulations and the points show the wide range across simulations (likely due to the relatively few error trials). (C) Reduction in neutral preference predicted by the computational model when non-pun<sub>bonus</sub> model parameter is set to 0. Sim. = Simulated. See the online article for the color version of this figure.

preference) as a function of depression/anxiety or rumination symptoms in the learning phase (Supplemental Figure S7B right, top row; depression/anxiety: posterior  $M = -.01$ , 90% CI  $[-.04, .01]$ ; rumination:  $M < .01$ , 90% CI  $[-.02, .03]$ ). There was also no difference in punishment avoidance retained into the test phase as a function of symptoms (Supplemental Figure S7B right, bottom row; depression/anxiety: posterior  $M < -.01$ , 90% CI  $[-.06, .05]$ ; rumination:  $M < -.01$ , 90% CI  $[-.06, .05]$ ).

When we entered computational-model parameters in Bayesian regressions predicting accuracy in the learning and test phase of the task (Figure 7, left), we found that various model parameters predicted performance in the expected direction: Worse performance was predicted by higher RL blunting ( $RL^{off}$ —especially at test, where RL is more dominant), higher WM decay ( $\phi$ ), higher lapse rate ( $\epsilon$ —especially at learning), and higher bonus assigned to items inferrable as nonpunishing (non-pun<sub>bonus</sub>—only outside of the 90% credibility interval during learning, where WM contributes). Conversely, higher WM usage and capacity ( $\rho$  and capacity  $\kappa$ —especially at learning) and higher RL rate from positive and negative prediction errors ( $\alpha^+$  and  $\alpha^-$ ; the former especially at test) all predicted better performance. These results confirm that the model parameters meaningfully corresponded to behavioral performance.

In contrast, symptoms showed much weaker relationships with model parameters (Figure 7, right; Supplemental Table S1). For rumination, the 90% highest density interval of all parameters overlapped 0, and the mean of many estimates was close to 0. For depression/anxiety, the 90% highest density interval of all parameters overlapped 0, although three parameters relating to our hypotheses had marginal evidence for them. Namely, the non-pun<sub>bonus</sub> parameter had a positive point estimate reflecting higher WM allocation to punishment as hypothesized ( $M = .06$ , 90% CI  $[-.04, .17]$ ). However, the point estimate for this parameter was lower when it was entered in a univariate model as a robustness check ( $M = .04$ , 90% CI

$[-.05, .14]$ ). Thus, especially when taken in combination with the lack of behavioral evidence for a neutral preference during learning as a function of depression/anxiety, we do not interpret the results as providing much evidence for our hypothesis. The other two parameters pertaining to our depression/anxiety hypotheses with some degree of support were WM capacity ( $\kappa$ ;  $M = -.07$ , 90% CI  $[-.19, .05]$ ) and decay ( $\phi$ ;  $M = -.05$ , 90% CI  $[-.17, .05]$ ). However, our prediction was of weaker WM contribution and these effects go in opposite directions: While the former does indeed reflect lower WM capacity, the latter indicates lower WM decay and thus reflects relatively better WM maintenance. Taken together, these findings provide little evidence of the hypothesized patterns due to internalizing-disorder symptoms—notwithstanding that our computational-model parameters strongly and meaningfully corresponded to task performance. Finally, in light of recent meta-analytic evidence of an elevated punishment learning rate in depression/anxiety (Pike & Robinson, 2022), it is noteworthy that the RL rate from negative prediction errors ( $\alpha^-$ ) was centered very close to 0 with a wide range of credible values ( $M = < .01$ , 90% CI  $[-.09, .11]$ ).

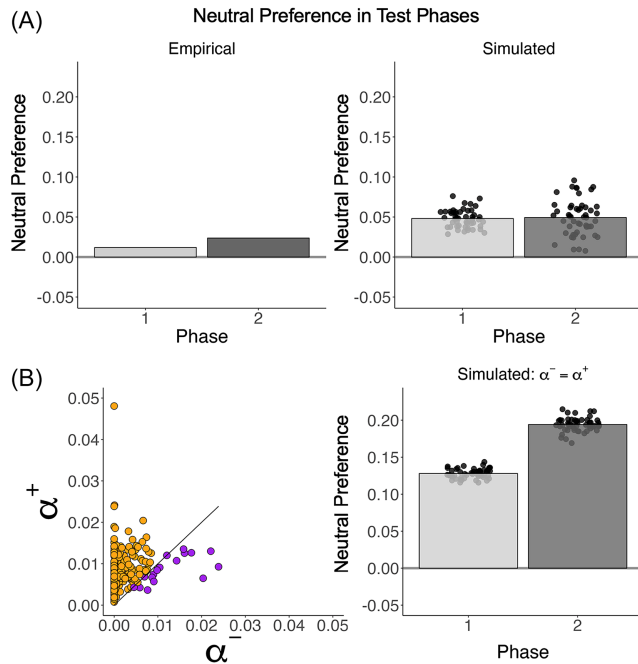
Findings were very similar when depression and anxiety symptoms were separately regressed on model parameters: All estimates had 90% highest density intervals that included 0 and thus did not meet our criterion for significance (Supplemental Figure S9).

## Discussion

Adaptive behavior requires taking rewarding actions while avoiding punishing ones. We studied how people learn these competing imperatives via a fast-but-fleeting WM in tandem with a slow-but-durable RL system. We did so via a novel approach/avoidance variant of an RL-WM task. In this variant, participants chose between three options on each trial that deterministically yielded reward, neutral, or punishment. Replicating a pattern found

**Figure 6**

*Poor Retention of Neutral Preference in the Test Phases Captured via Different RL Learning Rates*



*Note.* (A) The relatively poor retention of the neutral performance during the test phases (left; bars show the means) is captured by our computational model (right; points show the range across 50 simulations and bars show the mean across them). (B: Left) The relatively poor retention can be explained in the model in part by weak learning from negative prediction errors, reflected in the fact that (left) a higher learning rate from positive than negative prediction errors ( $\alpha^+ > \alpha^-$ ) was estimated for nearly all participants (the line is the identity line for  $\alpha^-$ ; purple points are the only participants for whom this parameter was higher than  $\alpha^+$ , whereas orange are the participants for whom  $\alpha^+$  was higher). (Right) When  $\alpha^-$  is set to the same value as was estimated for  $\alpha^+$ , the model predicts a higher rate of neutral preference in the test phase than was observed empirically (points again show the range across 50 simulations, and bars show means). RL = reinforcement learning. See the online article for the color version of this figure.

many times in the standard task, participants rapidly learned to pick the most rewarding option during acquisition phases, albeit parametrically less so as set size (and thus WM demand) increased.

To examine the effect of our addition of a punishment option, we next examined whether participants learned to prefer the neutral (and thus avoid the punishment) option within the subset of error trials. During learning, they indeed exhibited a preference for neutral relative to punishment options, but this preference was not robustly retained into test phases. Using computational modeling, we found that this pattern could be captured by allowing fleeting WM to ascribe a bonus to nonpunishing options, including those that were not directly experienced (see also Ben-Artzi et al., 2023; Biderman et al., 2023; Biderman & Shohamy, 2021), while the enduring RL system had differential learning from negative relative to positive PEs. Specifically, as in much past work (Collins, Ciullo, et al., 2017; Collins et al., 2014; Gershman, 2016; Master et al., 2020; Palminteri & Lebreton, 2022), optimization revealed a much lower RL rate for

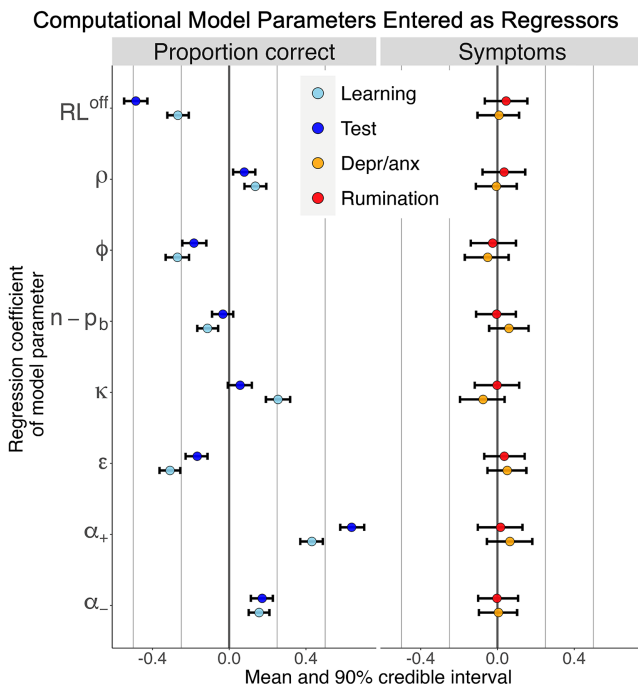
negative (than positive) PEs for the vast majority of participants, and a metric of penalized model fit supported dropping this parameter from the model entirely. Moreover, when we ran simulations with the RL rate for negative PEs set to the positive PE learning rate's value, the model incorrectly predicted much higher neutral preference retention into the test phase than was empirically observed. This demonstrates that poor neutral preference retention was not merely due to a paucity of error trials to enable sufficient learning, but rather due to a neglect of negative PEs by the RL system (but see also Collins, 2024; Palminteri, 2023; Sugawara & Katahira, 2021; Toyama et al., 2023, and further discussion below).

Our design also involved two surprise test phases interleaved with learning. These enabled a targeted test of findings from past RL-WM research that, when initial learning was within WM's capacity, the contribution of the RL system was blunted (Collins, 2018; Collins, Albrecht, et al., 2017; Collins, Ciullo, et al., 2017; Rac-Lubashevsky et al., 2023; see also Haile et al., 2024; Russin et al., 2024). In particular, Collins (2018) found that better performance in low versus high set sizes (3 vs. 6) during learning paradoxically reversed in the test phase. Hence, participants actually showed *enhanced* retention when they had learned under higher WM demand. Rac-Lubashevsky et al. (2023) found a full parametric reversal across five set sizes, such that the standard pattern of parametrically faster learning from Set Sizes 1 to 5 (low to high WM demand) showed a mirror-image pattern at later retention, with the best retention now at Set Size 5 and the worst at Set Size 1. Further, this study found that this test-phase retention was related to accelerated neural learning curves in the prior learning phase, due to large prediction errors when WM resources were degraded (Rac-Lubashevsky et al., 2023; see also Collins & Frank, 2018).

Our results are largely consistent with this interpretation, even though we found an inverted U rather than a linear decline as a function of set size in the first test phase—which we attribute to differences in our design. Specifically, whereas in Rac-Lubashevsky et al. (2023) the only testing took place after 15 learning experiences per stimulus (by which point learning was nearly perfect across set sizes), our first test phase was administered after just five learning experiences per stimulus. At this point, participants had not yet learned Set Sizes 4 and 5 enough to retain them well—an effect qualitatively captured by our model. Participants then resumed learning from feedback beginning at a much lower performance than where they had left off, leading to an increased rate of prediction errors. The model captured that the increased prediction errors help to consolidate RL-based learning, leading to a minimal effect of set size in a second test phase delivered at the end of learning (unlike in the standard paradigm, where the lack of a midway break leads to a virtual disappearance of prediction errors in lower set sizes in later parts of learning, eventually leading to worse retention in low set sizes during the sole test phase administered after learning; Collins, 2018; Rac-Lubashevsky et al., 2023). Of note, although our model qualitatively captured the inverted U pattern at the midway test, its replication in the first trial of the second learning phase, and its disappearance at the postlearning test, it did not capture the full magnitude of the inverted U (i.e., it predicted a less steep initial increase from Set Sizes 1 to 3 and a less sharp decline from Set Sizes 3 to 5), leaving room for improvement in future work.

However, we also found substantial individual differences that are not visible within the group pattern (see also Haile et al., 2024). Specifically, we found a broad range of estimated values for a

**Figure 7**  
*Model Parameters as Predictors of Performance and Symptoms*



*Note.* Bayesian regressions of computational-model z-scored parameter estimates on proportion correct during the learning and test phase (left) and psychiatric symptoms, specifically depression/anxiety and rumination (right).  $n - p_b$  = non-pun<sub>bonus</sub> parameter; RL = reinforcement learning. See the online article for the color version of this figure.

computational-model parameter that captured RL blunting,  $RL^{off}$ . High versus low blunters showed a qualitatively different pattern by the final test phase: Only the high blunters showed evidence of impeded retention in low set sizes wherein WM had dominated. In contrast, the low blunters appeared to have retained the benefits of faster acquisition under low set sizes (i.e., that it allows more experience picking the correct option; see also Collins, 2024; Miller et al., 2019) into the test phase—with parametrically higher retention in Set Sizes 1 to 5, mirroring the learning-phase pattern (however, even these relatively low blunters had poor Set Size 1 retention when it was tested midway through learning). These individual differences help to reconcile the RL-WM findings with other literature suggesting that WM and other top-down influences actually enhance RL (Cavanagh et al., 2010; Daniel et al., 2020; Doll et al., 2009, 2011; Farashahi et al., 2017; Frank & Claus, 2006; Geana et al., 2022; Gold et al., 2012; Hernaus et al., 2018, 2019; Hitchcock, Forman, et al., 2022; Hitchcock & Frank, 2024; Leong et al., 2017; Radulescu et al., 2016): They suggest that top-down influences that enhance initial performance help the RL system to ingrain the correct option—as long as concomitant RL blunting is not high enough that its negative effect outweighs this benefit.

In other results based on the test-phase data, we tested substituting an S-R module for the RL module in our model. Recent work by Collins (2024) found that such an S-R model could sufficiently account for the learning pattern in the RL-WM task—although the model was not fit to test-phase data, which should provide a purer

test of RL contribution (Collins, Ciullo, et al., 2017; Doll et al., 2011, 2016; Frank et al., 2007). We found that—although the WM and S-R model indeed adequately accounted for the accuracy pattern during the learning phase of the task and captured the poor retention of punishment avoidance into the test phase—it failed to capture the inverted U pattern in the first test phase, its repetition at the start of the next learning phase, or the pattern at the second test phase. Rather, it predicted parametrically declining performance as a function of set size (i.e., no paradoxical effect) at each of these points. This was true even though we allowed the S-R module to be blunted proportional to WM contribution, in the same way as the RL system (although we note that RL, rather than S-R, blunting is arguably more theoretically motivated and is consistent with neural findings; Collins, Ciullo, et al., 2017; Collins & Frank, 2018; Rac-Lubashevsky et al., 2023). The model with an S-R module also failed to capture the magnitude of neutral preference (punishment avoidance) during learning observed empirically.

In terms of psychiatric symptoms, contrary to our predictions that accuracy during learning would be impeded among depressed/anxious and trait ruminative individuals, respectively, due to worse executive function and WM preoccupation by off-task rumination, we in fact found that performance was remarkably similar as a function of symptoms (if anything, more depressed/anxious individuals showed slightly *less* deleterious effects of set size on performance; see also Frogner et al., 2025). Nor did we find any behavioral evidence for a greater neutral (over punishment) preference during learning as a function of depressed/anxious or ruminative symptoms or a paradoxical reversal of this effect in retention (as would be expected if WM had been more strongly allocated to punishment, paradoxically leading to worse test-phase retention due to RL blunting; see also Beltzer et al., 2023; Whitmer et al., 2012; however, given that neutral preference in the test phase was at floor at the group level [i.e., it did not statistically differ from 0], our design may not have been optimal to detect individual differences in the retention of punishment avoidance).

In computational modeling, we found modest evidence that depressed/anxious (but not ruminative) individuals prioritized punishment in WM, in terms of elevation within a model parameter that ascribed a bonus in WM to nonpunishing items. Although this pattern is consistent with our prediction that more depressed/anxious individuals would allocate greater WM to punishment avoidance, the 90% highest density interval for this effect overlapped 0, and a robustness check of it in a univariate model found a lower point estimate for this relationship. We did not find any evidence for weakened durable RL-based learning from punishment as a function of depression/anxiety or rumination (which would be predicted if there were a paradoxical effect on punishment retention due to heightened WM punishment allocation during learning); indeed, the credibility intervals for the effect of symptoms on the RL rate for negative PEs were centered near 0. Although there was modest evidence for two model parameters related to overall WM function differing as a function of depression/anxiety—namely, the WM capacity and the WM decay rate were both lower as a function of symptoms, although neither fell outside the 90% credibility interval—these parameters have opposite effects on task performance: The former reflects lower WM capacity and thus harms performance, whereas the latter reflects better retention of it and thus improves performance. Thus, the overall pattern was not consistent with our hypothesis of impaired WM as a function of depression/



anxiety. The effects of trait rumination on WM parameters were centered close to 0, again contrary to our predictions.

Overall, our findings suggest noteworthy sparing of learning in this task as a function of internalizing-disorder symptoms. Interestingly, Cheng et al. (2024) recently found that adolescents and young adults with a history of unipolar major depression (vs. no history of psychopathology) showed a reduced RL rate in the standard RL-WM task. It is plausible that these discrepant findings come from differences in the population (dimensionally oversampled adults from Prolific here vs. adolescents/young adults assessed for lifetime history via structured interview), task (the reward–punishment variant including two test phases used here vs. the reward-only classic task without a test phase in their study), and/or computational model (Cheng et al.’s, 2024, model had worse parameter recovery for the RL rate than ours and did not use separate learning rates for positive vs. negative PEs).

A strength of our new task variant is that, as far as we are aware, it is the first that is capable of disentangling WM and RL contributions in the ubiquitous everyday situation where actions can lead either to reward or to punishment. We developed a computational model that captured all the key features of this new task, including the magnitude of the learning curve stratified by set size, its precipitous decline and partial reversal after a retention test midway through learning, and the reestablishment of the set size stratification by the end of the second learning phase; the inverted U retention pattern midway through learning, its replication at Stimulus Iteration 6 (the start of the second learning phase), and its disappearance in the final retention test; and the development of a fleeting neutral preference during learning that was poorly retained into test phases. Specific parameters within our model offered insight into these effects: An  $RL^{off}$  parameter implementing RL blunting controlled the shape of the inverted U, a non-pun<sub>bonus</sub> WM parameter allowed for the short-lived neutral preference during learning, and allowing the RL system to have different learning rates for negative versus positive PEs (with optimization returning a much lower estimate for the former for the vast majority of participants) was needed to capture the poor ultimate retention of the neutral preference.

A further strength is that we used hierarchical Bayesian regression models to quantify the precise strength of evidence for the effects of psychiatric symptoms on task behavior and model parameters, allowing us to go beyond merely documenting null findings and instead to quantify the strength (or lack of strength) for various hypothesized effects. We found that learning was, for the most part, markedly spared under these task conditions. Results have been quite mixed regarding RL differences under internalizing-disorder symptoms (reviewed in Bishop & Gagne, 2018; Chen et al., 2015; Hitchcock, Fried, & Frank, 2022; Pike & Robinson, 2022; Yamamori et al., 2023). Our finding of spared learning came in the context of a design that had many features whose absence could have plausibly led to null results in prior studies (such as a task that requires executive function and RL cooperation, wherein it is possible to disentangle WM from RL, where there is a competing demand to learn to approach reward vs. avoid loss, with relatively high  $N$ , with depression and anxiety symptoms assessed dimensionally, with an operant design, and with identifiable RL rates from positive and negative PEs). That notable learning alterations were still not evident in a task with all of these features (and rather, learning was instead notably spared) should help the field of computational psychiatry to home in on whether there are other settings under which differences

reliably are present, such as if they are only present among participants meeting a diagnostic threshold or in designs with other features (e.g., probabilistic contingencies; nonstationary probabilities; independent bandit arms; monetary or intrinsic incentives for task performance; gamification; losses from endowments or with primary aversive value; elicitation of disorder-relevant states, such as rumination; young adult/adolescent samples; distinguishing cognitive and physiological anxiety; elicitation of an especially rich task context and/or personal goals; see, e.g., Banker et al., 2025; Beltzer et al., 2023; Blain et al., 2023; Brown et al., 2021; Cheng et al., 2024; Gagne et al., 2020; Hitchcock, Forman, et al., 2022; Karvelis et al., 2023; Pike & Robinson, 2022; Rutledge et al., 2017; Senta et al., 2025; Suddell et al., 2024; Wise & Dolan, 2020; Yamamori et al., 2023).

Our study also had important limitations. First, although our design choice to add a test phase midway through performance had the desired effect of decreasing performance at this point, and thereby increasing errors, this deterministic task nonetheless had few overall error trials. That we had relatively few errors to analyze led to more noise while investigating punishment avoidance, as was evident, for instance, in the range of simulated results with the same model and parameter values in Figure 5B and Figure 5C. Future studies may need to include task features—such as independent bandit arms (see Pike & Robinson, 2022) and probabilistic designs where WM and RL can to some extent still be disentangled (e.g., McDougale & Collins, 2021, Experiment 3)—to increase the error rate further. Further, although our task/computational model could disentangle WM from RL and moreover demonstrated some distinct predictions from RL and S-R modules, our experiment was not suited to disentangle the contributions of another system that might well be at play in this task: the episodic memory system (for instance, we did not use trial-unique stimuli; Bornstein et al., 2017; Bornstein & Norman, 2017; Gershman & Daw, 2017; Lengyel & Dayan, 2007). Hierarchical Bayesian computational modeling was also beyond the scope of this article, given the complexity of our model. Hence, we were unable to propagate uncertainty in parameter estimates into the symptom ~ parameter relationships shown on the right side of Figure 7, and parameter recovery may have been negatively affected (although we note that recovery was still adequate to high for all parameters and that we tested an approximation of full hierarchical Bayesian modeling—normalizing estimates by the group-level statistics, which can sometimes improve recovery, e.g., Frey et al., 2021; Hitchcock & Frank, 2024—although it did not do so for this task/model). Finally, although we embedded attentional checks into questionnaires, the type of check that we used (e.g., “Mark this item ‘Not at all’”) may be less effective in flagging careless/inattentive responding than checks requiring more careful attention, which could in turn have led to inflation within psychiatric symptoms (Zorowitz et al., 2023; although this concern is mitigated by the fact that such inflation increases the risk of false positives, whereas we found learning as a function of psychiatric symptoms was largely spared).

In sum, we developed a novel variant of an RL-WM task to understand how these systems interact while concurrently learning to gain reward and avoid punishment. We found that participants showed temporary punishment avoidance via WM, but with little retention of this preference; individual differences in the extent to which RL is blunted by WM; and spared learning as a function of internalizing-disorder symptoms. Our findings pave the way for

mechanistic insights into how WM and RL are allocated to approach reward and concurrently avoid punishment, and help to reveal boundary conditions under which learning is—and is not—altered as a function of psychiatric symptoms.

## Constraints on Generality

We investigated reward pursuit and punishment avoidance in an online American adult sample, about 3/4 of whom identified as White, collected from Prolific. We oversampled for participants who self-identified as experiencing depression or anxiety, but found that learning was notably similar among these individuals and an unselected subset of participants. As described in the Method section, participants were required to demonstrate basic capacity to understand and complete the task via a practice phase before moving onto the main task, and we removed a small subset of participants whose performance was indistinguishable from chance prior to analyses. Hence, we assume our participants had basic computer literacy and did not have starkly impaired memory. It is unclear to what extent our findings would generalize to non-English-speaking or memory-impaired participants, those without computer literacy, children or older adults, or non-Western, Educated, Industrialized, Rich, and Democratic participants.

## References

- Abramovitch, A., Short, T., & Schweiger, A. (2021). The C Factor: Cognitive dysfunction as a transdiagnostic dimension in psychopathology. *Clinical Psychology Review*, 86, Article 102007. <https://doi.org/10.1016/j.cpr.2021.102007>
- Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected papers of Hirotugu Akaike* (pp. 199–213). Springer. [https://doi.org/10.1007/978-1-4612-1694-0\\_15](https://doi.org/10.1007/978-1-4612-1694-0_15)
- Banker, S. M., Harrington, M., Schafer, M., Na, S., Heflin, M., Barkley, S., Trayvick, J., Peters, A. W., Thinakaran, A. A., Schiller, D., Foss-Feig, J. H., & Gu, X. (2025). Phenotypic divergence between individuals with self-reported autistic traits and clinically ascertained autism. *Nature Mental Health*, 3, 286–297. <https://doi.org/10.1038/s44220-025-00385-8>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. arXiv. <https://doi.org/10.48550/arXiv.1406.5823>
- Beck, A. T., Steer, R. A., & Brown, G. (1996). Beck Depression Inventory—II. *San Antonio*, 78(2), 490–498.
- Beltzer, M. L., Daniel, K. E., Daros, A. R., & Teachman, B. A. (2023). Examining social reinforcement learning in social anxiety. *Journal of Behavior Therapy and Experimental Psychiatry*, 80, Article 101810. <https://doi.org/10.1016/j.jbtep.2022.101810>
- Ben-Artzi, I., Kessler, Y., Nicenboim, B., & Shahar, N. (2023). Computational mechanisms underlying latent value updating of unchosen actions. *Science Advances*, 9(42), Article eadi2704. <https://doi.org/10.1126/sciadv.adi2704>
- Bideman, N., Gershman, S. J., & Shohamy, D. (2023). The role of memory in counterfactual valuation. *Journal of Experimental Psychology: General*, 152(6), 1754–1767. <https://doi.org/10.1037/xge0001364>
- Bideman, N., & Shohamy, D. (2021). Memory and decision making interact to shape the value of unchosen options. *Nature Communications*, 12(1), Article 4648. <https://doi.org/10.1038/s41467-021-24907-x>
- Bishop, S. J., & Gagne, C. (2018). Anxiety, depression, and decision making: A computational perspective. *Annual Review of Neuroscience*, 41(1), 371–388. <https://doi.org/10.1146/annurev-neuro-080317-062007>
- Blain, B., Pinhorn, I., & Sharot, T. (2023). Sensitivity to intrinsic rewards is domain general and related to mental health. *Nature Mental Health*, 1(9), 679–691. <https://doi.org/10.1038/s44220-023-00116-x>
- Bomstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). Reminders of past choices bias decisions for reward in humans. *Nature Communications*, 8(1), Article 15958. <https://doi.org/10.1038/ncomms15958>
- Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience*, 20(7), 997–1003. <https://doi.org/10.1038/nn.4573>
- Brown, V. M., Zhu, L., Solway, A., Wang, J. M., McCurry, K. L., King-Casas, B., & Chiu, P. H. (2021). Reinforcement learning disruptions in individuals with depression and sensitivity to symptom change following cognitive behavioral therapy. *JAMA Psychiatry*, 78(10), 1113–1122. <https://doi.org/10.1001/jamapsychiatry.2021.1844>
- Bürkner, P. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>
- Cavanagh, J. F., Frank, M. J., Klein, T. J., & Allen, J. J. B. (2010). Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. *NeuroImage*, 49(4), 3198–3209. <https://doi.org/10.1016/j.neuroimage.2009.11.080>
- Chen, C., Takahashi, T., Nakagawa, S., Inoue, T., & Kusumi, I. (2015). Reinforcement learning in depression: A review of computational research. *Neuroscience and Biobehavioral Reviews*, 55, 247–267. <https://doi.org/10.1016/j.neubiorev.2015.05.005>
- Cheng, Z., Moser, A. D., Jones, M., & Kaiser, R. H. (2024). Reinforcement learning and working memory in mood disorders: A computational analysis in a developmental transdiagnostic sample. *Journal of Affective Disorders*, 344, 423–431. <https://doi.org/10.1016/j.jad.2023.10.084>
- Collins, A. G. E. (2018). The tortoise and the hare: Interactions between reinforcement learning and working memory. *Journal of Cognitive Neuroscience*, 30(10), 1422–1432. [https://doi.org/10.1162/jocn\\_a.01238](https://doi.org/10.1162/jocn_a.01238)
- Collins, A. G. E. (2024). *RL or not RL? Parsing the processes that support human reward-based learning*. PsyArXiv. <https://doi.org/10.31234/osf.io/he3pm>
- Collins, A. G. E., Albrecht, M. A., Waltz, J. A., Gold, J. M., & Frank, M. J. (2017). Interactions among working memory, reinforcement learning, and effort in value-based choice: A new paradigm and selective deficits in schizophrenia. *Biological Psychiatry*, 82(6), 431–439. <https://doi.org/10.1016/j.biopsych.2017.05.017>
- Collins, A. G. E., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working memory contributions to reinforcement learning impairments in schizophrenia. *The Journal of Neuroscience*, 34(41), 13747–13756. <https://doi.org/10.1523/JNEUROSCI.0989-14.2014>
- Collins, A. G. E., Ciullo, B., Frank, M. J., & Badre, D. (2017). Working memory load strengthens reward prediction errors. *The Journal of Neuroscience*, 37(16), 4332–4342. <https://doi.org/10.1523/JNEUROSCI.2700-16.2017>
- 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), 1024–1035. <https://doi.org/10.1111/j.1460-9568.2011.07980.x>
- Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, 120(1), 190–229. <https://doi.org/10.1037/a0030852>
- Collins, A. G. E., & Frank, M. J. (2018). Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning

- and working memory. *Proceedings of the National Academy of Sciences of the United States of America*, 115(10), 2502–2507. <https://doi.org/10.1073/pnas.1720963115>
- Collins, A. G. E., & Shenhav, A. (2022). Advances in modeling learning and decision-making in neuroscience. *Neuropsychopharmacology*, 47(1), 104–118. <https://doi.org/10.1038/s41386-021-01126-y>
- Conway, C. C., Hammen, C., & Brennan, P. A. (2012). Expanding stress generation theory: Test of a transdiagnostic model. *Journal of Abnormal Psychology*, 121(3), 754–766. <https://doi.org/10.1037/a0027457>
- Daniel, R., Radulescu, A., & Niv, Y. (2020). Intact reinforcement learning but impaired attentional control during multidimensional probabilistic learning in older adults. *The Journal of Neuroscience*, 40(5), 1084–1096. <https://doi.org/10.1523/JNEUROSCI.0254-19.2019>
- Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6), 1204–1215. <https://doi.org/10.1016/j.neuron.2011.02.027>
- Doll, B. B., Bath, K. G., Daw, N. D., & Frank, M. J. (2016). Variability in dopamine genes dissociates model-based and model-free reinforcement learning. *The Journal of Neuroscience*, 36(4), 1211–1222. <https://doi.org/10.1523/JNEUROSCI.1901-15.2016>
- Doll, B. B., Hutchison, K. E., & Frank, M. J. (2011). Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *The Journal of Neuroscience*, 31(16), 6188–6198. <https://doi.org/10.1523/JNEUROSCI.6486-10.2011>
- Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research*, 1299, 74–94. <https://doi.org/10.1016/j.brainres.2009.07.007>
- Farashahi, S., Rowe, K., Aslami, Z., Lee, D., & Soltani, A. (2017). Feature-based learning improves adaptability without compromising precision. *Nature Communications*, 8(1), Article 1768. <https://doi.org/10.1038/s41467-017-01874-w>
- Frank, M. J., & Claus, E. D. (2006). Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, 113(2), 300–326. <https://doi.org/10.1037/0033-295X.113.2.300>
- Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences of the United States of America*, 104(41), 16311–16316. <https://doi.org/10.1073/pnas.0706111104>
- Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science*, 306(5703), 1940–1943. <https://doi.org/10.1126/science.1102941>
- Frey, A.-L., Frank, M. J., & McCabe, C. (2021). Social reinforcement learning as a predictor of real-life experiences in individuals with high and low depressive symptomatology. *Psychological Medicine*, 51(3), 408–415. <https://doi.org/10.1017/S0033291719003222>
- Frogner, E. R., Dahl, A., Kjelkenes, R., Moberget, T., Collins, A., Westlye, L. T., & Pedersen, M. L. (2025, March 17). *Linking cognitive mechanisms of instrumental learning to age and symptoms of anxiety and depression in adolescence*. [https://doi.org/10.31219/osf.io/b826r\\_v1](https://doi.org/10.31219/osf.io/b826r_v1)
- Gagne, C., Zika, O., Dayan, P., & Bishop, S. J. (2020). Impaired adaptation of learning to contingency volatility in internalizing psychopathology. *eLife*, 9, Article e61387. <https://doi.org/10.7554/eLife.61387>
- Geana, A., Barch, D. M., Gold, J. M., Carter, C. S., MacDonald, A. W., III, Ragland, J. D., Silverstein, S. M., & Frank, M. J. (2022). Using computational modeling to capture schizophrenia-specific reinforcement learning differences and their implications on patient classification. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 7(10), 1035–1046. <https://doi.org/10.1016/j.bpsc.2021.03.017>
- Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, 71, 1–6. <https://doi.org/10.1016/j.jmp.2016.01.006>
- Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, 68(1), 101–128. <https://doi.org/10.1146/annurev-psych-122414-033625>
- Ghalanos, A., & Theussl, S. (2011). *Package "Rsolnp."* <https://cran.r-project.org/web/packages/Rsolnp/index.html>
- Gold, J. M., Waltz, J. A., Matveeva, T. M., Kasanova, Z., Strauss, G. P., Herbener, E. S., Collins, A. G., & Frank, M. J. (2012). Negative symptoms and the failure to represent the expected reward value of actions: Behavioral and computational modeling evidence. *Archives of General Psychiatry*, 69(2), 129–138. <https://doi.org/10.1001/archgenpsychiatry.2011.1269>
- Grahek, I., Shenhav, A., Musslick, S., Krebs, R. M., & Koster, E. H. W. (2019). Motivation and cognitive control in depression. *Neuroscience and Biobehavioral Reviews*, 102, 371–381. <https://doi.org/10.1016/j.neubiorev.2019.04.011>
- Griffith, J. W., Zinbarg, R. E., Craske, M. G., Mineka, S., Rose, R. D., Waters, A. M., & Sutton, J. M. (2010). Neuroticism as a common dimension in the internalizing disorders. *Psychological Medicine*, 40(7), 1125–1136. <https://doi.org/10.1017/S0033291709991449>
- Haile, T. M., Prat, C. S., & Stocco, A. (2024). One size does not fit all: Idiographic computational models reveal individual differences in learning and meta-learning strategies. *Topics in Cognitive Science*. Advance online publication. <https://doi.org/10.1111/tops.12730>
- Hernaus, D., Frank, M. J., Brown, E. C., Brown, J. K., Gold, J. M., & Waltz, J. A. (2019). Impaired expected value computations in schizophrenia are associated with a reduced ability to integrate reward probability and magnitude of recent outcomes. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 4(3), 280–290. <https://doi.org/10.1016/j.bpsc.2018.11.011>
- Hernaus, D., Gold, J. M., Waltz, J. A., & Frank, M. J. (2018). Impaired expected value computations coupled with overreliance on stimulus-response learning in schizophrenia. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(11), 916–926. <https://doi.org/10.1016/j.bpsc.2018.03.014>
- Hitchcock, P. F., Forman, E., Rothstein, N., Zhang, F., Kounios, J., Niv, Y., & Sims, C. (2022). Rumination derails reinforcement learning with possible implications for ineffective behavior. *Clinical Psychological Science*, 10(4), 714–733. <https://doi.org/10.1177/21677026211051324>
- Hitchcock, P. F., & Frank, M. J. (2024). The challenge of learning adaptive mental behavior. *Journal of Psychopathology and Clinical Science*, 133(5), 413–426. <https://doi.org/10.1037/abn0000924>
- Hitchcock, P. F., Fried, E. I., & Frank, M. J. (2022). Computational psychiatry needs time and context. *Annual Review of Psychology*, 73(1), 243–270. <https://doi.org/10.1146/annurev-psych-021621-124910>
- Karvelis, P., Paulus, M. P., & Diaconescu, A. O. (2023). Individual differences in computational psychiatry: A review of current challenges. *Neuroscience and Biobehavioral Reviews*, 148, Article 105137. <https://doi.org/10.1016/j.neubiorev.2023.105137>
- Lahey, B. B. (2009). Public health significance of neuroticism. *American Psychologist*, 64(4), 241–256. <https://doi.org/10.1037/a0015309>
- Lengyel, M., & Dayan, P. (2007). Hippocampal contributions to control: The right way. *Neural Information Processing Systems*, 20, 889–896. <https://proceedings.neurips.cc/paper/2007/hash/1f4477bad7af3616c1f933a02bfa4be4e-Abstract.html>
- Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, 93(2), 451–463. <https://doi.org/10.1016/j.neuron.2016.12.040>
- Masset, P., & Gershman, S. J. (in press). Reinforcement learning with dopamine: A convergence of natural and artificial intelligence. In S. Cragg & M. Walton (Eds.), *The handbook of dopamine*. Elsevier.
- Master, S. L., Eckstein, M. K., Gotlieb, N., Dahl, R., Wilbrecht, L., & Collins, A. G. E. (2020). Corrigendum to “Disentangling the systems contributing to changes in learning during adolescence” [Dev. Cogn.



- Neurosci. 41, 2020, 100732]. *Developmental Cognitive Neuroscience*, 45, Article 100854. <https://doi.org/10.1016/j.dcn.2020.100854>
- McDoughle, S. D., & Collins, A. G. E. (2021). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychonomic Bulletin & Review*, 28(1), 20–39. <https://doi.org/10.3758/s13423-020-01774-z>
- Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological Review*, 126(2), 292–311. <https://doi.org/10.1037/rev0000120>
- Mirman, D. (2014). *Growth curve analysis and visualization using R*. CRC Press.
- Molinero, G., & Collins, A. G. E. (2023). A goal-centric outlook on learning. *Trends in Cognitive Sciences*, 27(12), 1150–1164. <https://doi.org/10.1016/j.tics.2023.08.011>
- Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3), 139–154. <https://doi.org/10.1016/j.jmp.2008.12.005>
- Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, 22(10), 1544–1553. <https://doi.org/10.1038/s41593-019-0470-8>
- Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *The Journal of Neuroscience*, 35(21), 8145–8157. <https://doi.org/10.1523/JNEUROSCI.2978-14.2015>
- Nolen-Hoeksema, S., & Morrow, J. (1991). A prospective study of depression and posttraumatic stress symptoms after a natural disaster: The 1989 Loma Prieta Earthquake. *Journal of Personality and Social Psychology*, 61(1), 115–121. <https://doi.org/10.1037/0022-3514.61.1.115>
- O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, reward, and decision making. *Annual Review of Psychology*, 68(1), 73–100. <https://doi.org/10.1146/annurev-psych-010416-044216>
- Palminteri, S. (2023). Choice-confirmation bias and gradual perseveration in human reinforcement learning. *Behavioral Neuroscience*, 137(1), 78–88. <https://doi.org/10.1037/bne0000541>
- Palminteri, S., & Lebreton, M. (2022). The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences*, 26(7), 607–621. <https://doi.org/10.1016/j.tics.2022.04.005>
- Pike, A. C., & Robinson, O. J. (2022). Reinforcement learning in patients with mood and anxiety disorders vs control individuals: A systematic review and meta-analysis. *JAMA Psychiatry*, 79(4), 313–322. <https://doi.org/10.1001/jamapsychiatry.2022.0051>
- Piray, P., Dezfouli, A., Heskes, T., Frank, M. J., & Daw, N. D. (2019). Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies. *PLOS Computational Biology*, 15(6), Article e1007043. <https://doi.org/10.1371/journal.pcbi.1007043>
- Provenza, N. R., Gelin, L. F. F., Mahaphanit, W., McGrath, M. C., Dastin-van Rijn, E. M., Fan, Y., Dhar, R., Frank, M. J., Restrepo, M. I., Goodman, W. K., & Borton, D. A. (2022). Honeycomb: A template for reproducible psychophysiological tasks for clinic, laboratory, and home use. *Revista Brasileira de Psiquiatria*, 44(2), 147–155. <https://doi.org/10.1590/1516-4446-2020-1675>
- R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Rac-Lubashevsky, R., Cremer, A., Collins, A. G. E., Frank, M. J., & Schwabe, L. (2023). Neural index of reinforcement learning predicts improved stimulus–response retention under high working memory load. *The Journal of Neuroscience*, 43(17), 3131–3143. <https://doi.org/10.1523/JNEUROSCI.1274-22.2023>
- Radulescu, A., Daniel, R., & Niv, Y. (2016). The effects of aging on the interaction between reinforcement learning and attention. *Psychology and Aging*, 31(7), 747–757. <https://doi.org/10.1037/pag0000112>
- Reilly, E. E., Lavender, J. M., Berner, L. A., Brown, T. A., Wierenga, C. E., & Kaye, W. H. (2019). Could repetitive negative thinking interfere with corrective learning? The example of anorexia nervosa. *International Journal of Eating Disorders*, 52(1), 36–41. <https://doi.org/10.1002/eat.22997>
- Rmus, M., McDoughle, S. D., & Collins, A. G. E. (2021). The Role of Executive Function in Shaping Reinforcement Learning. *Current Opinion in Behavioral Sciences*, 38, 66–73. <https://doi.org/10.1016/j.cobeha.2020.10.003>
- Russin, J., Pavlick, E., & Frank, M. J. (2024). *Curriculum effects and compositionality emerge with in-context learning in neural networks*. arXiv. <https://doi.org/10.48550/arXiv.2402.08674>
- Rutherford, A. V., McDoughle, S. D., & Joermann, J. (2023). “Don’t [ruminate], be happy”: A cognitive perspective linking depression and anhedonia. *Clinical Psychology Review*, 101, Article 102255. <https://doi.org/10.1016/j.cpr.2023.102255>
- Rutledge, R. B., Moutoussis, M., Smittenaar, P., Zeidman, P., Taylor, T., Hryniewicz, L., Lam, J., Skandali, N., Siegel, J. Z., Ousdal, O. T., Prabhu, G., Dayan, P., Fonagy, P., & Dolan, R. J. (2017). Association of neural and emotional impacts of reward prediction errors with major depression. *JAMA Psychiatry*, 74(8), 790–797. <https://doi.org/10.1001/jamapsychiatry.2017.1713>
- Senta, J., Bishop, S., & Collins, A. G. (2025). *Dual process impairments in reinforcement learning and working memory systems underlie learning deficits in physiological anxiety*. bioRxiv. <https://doi.org/10.1101/2025.02.14.638024>
- Sharp, P. B., Russek, E. M., Huys, Q. J. M., Dolan, R. J., & Eldar, E. (2022). Correction: Humans perseverate on punishment avoidance goals in multigoal reinforcement learning. *eLife*, 11, Article e83998. <https://doi.org/10.7554/eLife.83998>
- Snyder, H. R. (2013). Major depressive disorder is associated with broad impairments on neuropsychological measures of executive function: A meta-analysis and review. *Psychological Bulletin*, 139(1), 81–132. <https://doi.org/10.1037/a0028727>
- Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, 166(10), 1092–1097. <https://doi.org/10.1001/archinte.166.10.1092>
- Suddell, S., Zhang, L., Lee, C., Senta, J., O’Keane, V., Ward, T., Stephan, K. E., Fox, C. A., Hanlon, A., Lynch, K., Harty, S., Gillan, C., & Richards, D. (2024, September 17). *Limited evidence for reduced learning rate adaptation in anxious-depression, before or after treatment*. <https://doi.org/10.31234/osf.io/hm46n>
- Sugawara, M., & Katahira, K. (2021). Dissociation between asymmetric value updating and perseverance in human reinforcement learning. *Scientific Reports*, 11(1), Article 3574. <https://doi.org/10.1038/s41598-020-80593-7>
- Toyama, A., Katahira, K., & Kunisato, Y. (2023). Examinations of biases by model misspecification and parameter reliability of reinforcement learning models. *Computational Brain & Behavior*, 6(4), 651–670. <https://doi.org/10.1007/s42113-023-00175-4>
- Westbrook, A., van den Bosch, R., Hofmans, L., Papadopetraki, D., Määttä, J. I., Collins, A. G. E., Frank, M. J., & Cools, R. (2024). *Striatal dopamine can enhance learning, both fast and slow, and also make it cheaper*. bioRxiv. <https://doi.org/10.1101/2024.02.14.580392>
- Whitmer, A. J., Frank, M. J., & Gotlib, I. H. (2012). Sensitivity to reward and punishment in major depressive disorder: Effects of rumination and of single versus multiple experiences. *Cognition and Emotion*, 26(8), 1475–1485. <https://doi.org/10.1080/02699931.2012.682973>
- Wierenga, C. E., Reilly, E., Bischoff-Grethe, A., Kaye, W. H., & Brown, G. G. (2022). Altered reinforcement learning from reward and punishment in anorexia nervosa: Evidence from computational modeling. *Journal of the International Neuropsychological Society*, 28(10), 1003–1015. <https://doi.org/10.1017/S1355617721001326>
- Wise, T., & Dolan, R. J. (2020). Associations between aversive learning processes and transdiagnostic psychiatric symptoms in a general population



- sample. *Nature Communications*, 11(1), Article 4179. <https://doi.org/10.1038/s41467-020-17977-w>
- Yamamori, Y., Robinson, O. J., & Roiser, J. P. (2023). Approach-avoidance reinforcement learning as a translational and computational model of anxiety-related avoidance. *eLife*, 12, Article RP87720. <https://doi.org/10.7554/eLife.87720.4>
- Yoo, A. H., & Collins, A. G. E. (2022). How working memory and reinforcement learning are intertwined: A cognitive, neural, and computational perspective. *Journal of Cognitive Neuroscience*, 34(4), 551–568. [https://doi.org/10.1162/jocn\\_a\\_01808](https://doi.org/10.1162/jocn_a_01808)
- Zorowitz, S., Solis, J., Niv, Y., & Bennett, D. (2023). Inattentive responding can induce spurious associations between task behaviour and symptom measures. *Nature Human Behaviour*, 7(10), 1667–1681. <https://doi.org/10.1038/s41562-023-01640-7>
- Received October 14, 2024  
Revision received May 7, 2025  
Accepted June 7, 2025 ■