RL or not RL? Parsing the processes that support human reward-based learning.

Anne GE Collins (ORCID: 0000-0003-3751-3662)
Department of Psychology
Helen Wills Neuroscience Institute
University of California
Berkeley
annecollins@berkeley.edu

Abstract

Reinforcement Learning (RL) algorithms have had tremendous success accounting for reward-based learning across species, in both behavior and brain. In particular, simple model-free RL models, such as delta-rule or Q-learning, are routinely used to model instrumental learning in bandit tasks, and they capture variance in brain signals. However, reward-based learning in humans recruits multiple processes, including high-level processes such as memory and low-level ones such as choice perseveration; their contributions can easily be mistakenly attributed to RL computations. Here, we investigate how much of RL-like behavior is supported by RL computations in a context where other processes can be factored out. Re-analysis and computational modeling of seven data sets spanning hundreds of participants show that in this instrumental context, reward-based learning is best explained by a combination of working memory and a habit-like associative process, with no RL-like value-based incremental learning. Simulations show that this combination nevertheless approximates the adaptive policy of a value-based RL agent, explaining why RL computations are mistakenly inferred when working memory is not parsed out. Our results raise important questions for the interpretation of RL as a meaningful process across brain and behavior, and call for a reconsideration of how we interpret findings in reinforcement learning across levels of analysis.

Introduction

The reinforcement learning (RL) framework in computational cognitive neuroscience has been tremendously successful, largely because RL purportedly bridges between behavior and brain levels of analysis (Niv, 2009; Niv & Langdon, 2016). Model-free RL algorithms track the expected value of a state and update it in proportion to a reward prediction error (RPE) (Sutton & Barto, 2018); this interpretable computation also accounts for important aspects of dopaminergic signaling and striatal activity (Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997). Indeed, extensive research has supported the

theory that cortico-striatal networks support RL-like computations for reward-based learning, and that disruption of this network causes predicted deficits in behavior (Frank, Seeberger, & O'reilly, 2004; Tai, Lee, Benavidez, Bonci, & Wilbrecht, 2012). In parallel, similar model-free RL algorithms have been broadly and successfully used to explain and capture many aspects of reward-based learning behavior across species, from simple classical conditioning (Wagner & Rescorla, 1972) to more complex multi-armed contextual bandit tasks (Daw & Tobler, 2014; Palminteri & Lebreton, 2022).

However, there is also strong evidence that other cognitive processes, supported by separable brain networks, also contribute to reward-based learning (Rmus, McDougle, & Collins, 2021; Yoo & Collins, 2022). Early research in rodents showed a double dissociation between so-called "habits" (thought to relate to the RL process) and more "goal-directed" processes, which are more sensitive to knowledge about the task environment, and thus support more flexible behavior (Tolman, 1948; Yin, Knowlton, & Balleine, 2004; Yin, Ostlund, Knowlton, & Balleine, 2005). Widely accepted dual process theories of learning typically capture the slow/inflexible processes with model-free RL algorithms (Daw, Gershman, Seymour, Dayan, & Dolan, 2011). However this apparent consensus hides broad ambiguity and disagreement about what the fast/flexible vs. slow/inflexible processes are (Collins & Cockburn, 2020). Indeed, recent literature has highlighted multiple processes that strongly contribute to learning. In more complex environments with navigation-like properties this may entail the use of a map of the environment for forward planning (Daw et al., 2011). Even in simple environments typically modeled with model-free RL, additional processes such as working memory (Yoo & Collins, 2022), episodic memory (Bornstein & Norman, 2017; Gershman & Daw, 2017), and choice perseveration strategies (Sugawara & Katahira, 2021) have been found to play an important role. In particular, instrumental learning tasks such as contextual multi-armed bandits rely mostly on working memory, with contributions of a slow RL-like process when load overcomes working memory resources (Collins & Frank, 2012; McDougle & Collins, 2021).

Because the RL family of models is highly flexible (Sutton & Barto, 2018), RL models have nonetheless successfully captured behavior that is likely more driven by other processes such as working memory. Indeed, in most simple laboratory tasks, non-RL processes make very similar predictions to RL ones - for example, perseveration strategies might be mistaken for a learning rate asymmetry in RL (Sugawara & Katahira, 2021), and working memory contributions might be mistaken for high learning rates (Collins & Frank, 2012). Non-RL processes become identifiable only in environments explicitly designed to attempt to disentangle them (Bornstein & Norman, 2017; Collins & Frank, 2012). Thus, non-RL processes' contributions to learning are often attributed to RL computations, and this misattribution of various processes' to "RL" may lead to confusion in the literature, when findings relying on RL modeling are mistakenly attributed to RL brain processes (Collins, Brown, Gold, Waltz, & Frank, 2014; Eckstein, Wilbrecht, & Collins, 2021).

Here, we investigate how much of reward-based instrumental learning actually reflects a model-free RL process, as typically formulated in the literature. Because of the well characterized and major contributions of working memory in instrumental learning, we focus on a task context where working memory's contribution can be adequately parsed out, the RLWM paradigm (Collins & Frank, 2012). We reason that a key characteristic of model-free RL is that it integrates reward outcomes over time to build a cached value esti-

mate that drives policy; more generally, negative prediction error in model-free RL should make an agent less likely to repeat the corresponding choice. We thus focus here on how positive ("correct, +1") and, more importantly, negative ("incorrect, 0") outcomes affect later choices.

Behavioral analysis and computational modeling of 7 datasets across two experimental paradigm versions (5 previously published and one novel for the deterministic version, RLWM; 1 previously published for the probabilistic version, RLWMP) show that, when parsing out working memory, we cannot detect evidence of RL in reward-based learning. Indeed, predictions including an RL process are falsified Palminteri, Wyart, and Koechlin (2017). Instead, all behavior can be explained by a mixture of a fast, flexible, and capacity limited process (working memory) and a slower, broader process that tracks stimulus-action associations, irrespective of outcomes, and is thus not RL. These findings call for a reconsideration of how we interpret findings in reinforcement learning across levels of analysis.

Results

The RLWM task was designed to disentangle the contributions of working memory (WM) learning from those of slower RL processes to reward-based learning by manipulating information load. Across independent blocks, participants learned stable stimulus-action associations between a novel set of stimuli (set size ranging from 2 to 6 items within participants) and 3 actions. The correct action for each stimulus was deterministically signaled by a correct (or +1) feedback, while the other two incorrect actions were signaled with incorrect (or 0) (Fig. 1A). Participants' behavior in low set sizes appeared was close to optimal, but increasing set size led to increasingly incremental learning curves(Fig. 1B), a pattern replicated across multiple previous studies in diverse populations (Collins, 2018a, 2018b; Collins, Albrecht, Waltz, Gold, & Frank, 2017; Collins et al., 2014; Collins, Ciullo, Frank, & Badre, 2017; Collins & Frank, 2012, 2018; Master et al., 2020; McDougle & Collins, 2021; Rmus et al., 2023; Yoo, Keglovits, & Collins, 2023; Zou, Muñoz Lopez, Johnson, & Collins, 2022). This pattern could only be captured by the RLWM model, a mixture model of two processes representing WM and RL. In this model, the RL process is a standard delta-rule learner; the WM module has a learning rate of 1 to capture immediate perfect learning, but also decay to capture WM's short time scale of maintenance; the mixture reflects WM resource limitations, such that behavior is mostly driven by fast and forgetful WM when the load is within WM resources, but supplemented by RL with increasing load (see Methods). This model included a bias weight parameterizing asymmetric updating of positive and negative feedback. This bias was shared between WM and RL and modulated learning rates for incorrect vs. correct outcomes. Previous model fitting of the bias parameter (shared between WM and RL) revealed that incorrect outcomes had a weaker impact on subsequent choices than correct outcomes (see e.g. (Master et al., 2020)).

Value and reward integration

To better identify the slower (RL) learning component in this task, we first sought to understand how positive and negative outcomes were integrated to impact policy. Specifically, we reasoned that a process learning from reward prediction errors in an RL-like way should use negative feedback in error trials to make participants less likely to repeat mis-

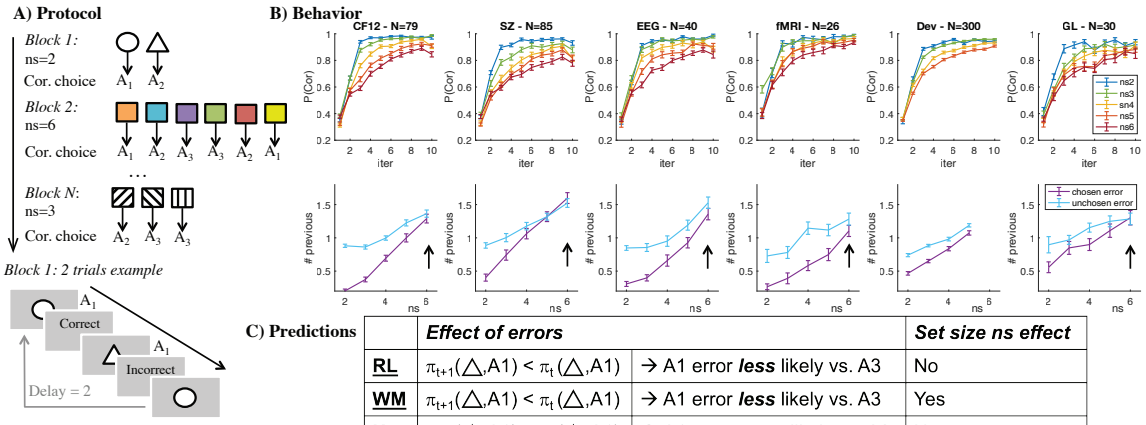
takes, and more so the more they made the same mistakes (see methods, Fig. 1C). We thus computed, within error trials, whether the specific error participants made (out of 2 possible errors for a given stimulus) was indeed the one that had been made less frequently than the other error.

Across all 6 datasets in the RLWM task, the number of previous errors was overall lower for the chosen vs. unchosen error (all $ts > 4$, $ps < 10^{-4}$, see table 1), showing that participants did use negative feedback overall in the task. As expected if participants' ability to use WM to guide choices decreased with set size, higher set sizes led to an increase in the number of previous errors for both chosen and unchosen. The difference between error type numbers, indicating participants' ability to avoid previously unrewarded choices, decreased with set size, as expected if higher set sizes reflected a higher portion of responsibility from a slower learning process (all $ts > 2.28$, $ps < 0.05$; see supplementary table 2). However, we observed in all data sets that the difference decreased strongly (see blue vs. purple curves in Fig. 1B, arrows at $ns = 6$), such that participants' policy appeared to become insensitive to negative outcomes selectively in set size $ns = 6$ in 4 out of 5 data sets that included set size 6 (see supplementary table 1). The effect even appeared to reverse in late learning in two datasets (Dev and SZ), such that errors committed late in learning in large set sizes had been repeated more often than the other error ($t's > 4.4$,$p < 10^{-4}$, see table 3), showing error perseveration effects.

We compared participants' patterns of errors to the predictions from four variants of the RLWM model - one treating gains and losses equally in both WM and RL models, one with a shared bias (Master et al., 2020), and the two best fitting RLWM models with no or weak bias against errors in WM and full bias in RL, indicating complete neglect of negative outcomes in the RL module. All models captured the set-size effect of performance in the qualitative pattern of the learning curves (Fig. 2B), the main effect of the chosen vs. unchosen error, and the increase in number of previous errors for both chosen and unchosen. The models also predicted that the difference between error type numbers (indicating participants' ability to avoid previously unrewarded choices) decreased with set size. However, all models predicted that the difference should remain large even in large set sizes (see blue vs. purple curves in Fig. 2B, arrows at $ns = 6$), contrary to what we observed empirically. In all 6 data sets, the magnitude of the difference decrease between the past number of chosen vs. unchosen errors could not be accounted for by any RLWM model, particularly late in learning (Fig. 2B bottom, grey curves).

New WMH model explains behavior

The behavioral and modeling results so far showed efficient integration of negative outcomes in low set sizes but not high set sizes, supporting the idea that working memory uses negative outcomes to guide avoidance in policy, but the slower, less resource limited process that supports instrumental learning under higher load does not. However, even with an RL negative learning rate $\alpha_- = 0$, RLWM models could not capture the pattern, because WM contributes to the choices even in high set sizes. Further variants of the RLWM family model, including with policy-compression mechanisms, could not reproduce the qualitative pattern (see supplementary figure 9). We reasoned that the slow process should, to a degree, counteract WM's ability to learn to avoid errors from negative outcomes. We thus explored a family of models where the slow module association weights (Q-values for RL)



C) Predictions

	Effect of errors		*Set size ns effect*
RL	$\pi_{t+1}(\triangle,A1) < \pi_t(\triangle,A1)$	→ A1 error ***less*** likely vs. A3	No
WM	$\pi_{t+1}(\triangle,A1) < \pi_t(\triangle,A1)$	→ A1 error ***less*** likely vs. A3	Yes
H	$\pi_{t+1}(\triangle,A1) > \pi_t(\triangle,A1)$	→ A1 error ***more*** likely vs. A3	No

Figure 1

A - RLWM Experimental paradigm. Participants performed multiple independent blocks of reinforcement learning task, using deterministic binary feedback to identify which of three actions is correct for each of ns stimuli. Varying set size ns targets working memory load and allows us to isolate its contribution (Collins & Frank, 2012). B - Behavior (mean, standard error) across 6 data sets on the RLWM task: CF12 (Collins & Frank, 2012); SZ (Collins et al., 2014); EEG (Collins, 2018a); fMRI (Collins, Ciullo, et al., 2017); Dev (Master et al., 2020); GL (unpublished). Top: learning curves show the probability of a correct action choice as a function of stimulus iteration number, plotted per set size, illustrating a strong set-size effect that highlights working memory contributions to behavior. Bottom: Error trial analysis: number of previous errors that are the same as the chosen error (purple) or the other possible error (unchosen; teal) as a function of set size. The large gap in low set sizes indicates that participants avoid errors they made previously more often than other errors; the absence of a gap in high set sizes indicates that participants are unable to learn to avoid their past errors (black arrows). C - qualitative predictions for the RL, WM and novel H modules, based on trial example in panel A). Only WM modules predict a set size effect (Collins & Frank, 2012). Only H modules predicts that participants are more likely to repeat a previous error (e.g. selecting action A_1 for the triangle) than more likely to avoid it.

were updated with a subjective outcome r_0 for negative outcomes of $r = 0$. Surprisingly, the best fitting model across 6 data sets (see Fig. 2 left) was a model with fixed $r_0 = 1$, such that receiving incorrect feedback led to the same positive prediction error as correct feedback would. Negative learning rates still included a bias term shared across both modules. Note that this slow module cannot be interpreted as an RL module anymore, as the association weights track a relative frequency of stimulus-action choice, irrespective of outcomes, rather than an estimated value. This module can be thought of as an associative "Hebbian", or "habit-like", thus we label it H-agent, with the mixture model WMH. While it is similar to a choice perseveration kernelToyama, Katahira, and Kunisato (2023), note that it is not purely motor, but stimulus-dependent - indeed, all models also include a motor choice perseveration mechanism capturing participants tendency to repeat actions across trials.

The WMH model fit quantitatively better than models with RL. It was also successful at producing the qualitative pattern of errors observed in real participants, such that errors in high set size appeared to fully neglect negative outcomes in a way RLWM models could not (Fig. 2C bottom, supplementary figures 9 for full validation of all models in panel A in lal data sets). We further verified that this pattern of error changed dynamically over the course of learning in participants in a way that the model could capture (Fig. 2C bottom).

WMH also explains behavior in probabilistic tasks

While using the RLWM task was useful to adequately factor out WM contributions to reward-based learning, it has the downside that the task does not necessitate integration of reward in the same way probabilistic tasks do (Frank et al., 2004). We thus sought to confirm whether our finding would hold in a probabilistic version of the task RLWMP; to that effect, we re-analyzed a previously published data-set (see (McDougle & Collins, 2021), experiment 3). As previously reported, behavior in this task was sensitive to set size ($F(1, 33) = 55.99$, $p < 0.001$), indicating that WM contributes to learning even in probabilistic environments thought to be more suited to eliciting RL-like computations. Similar to the deterministic task, we modeled behavior with a mixture of a fast, forgetful learning process (related to WM), and a slower, non-forgetful one, with the constraint that the first process contributed relatively more to mixture behavior in low than high set size (see methods). We compared models where the slow process was either RL like (i.e., integrating negative outcomes differently from positive ones; RLWM) or association like (i.e., integrating negative outcomes similarly to positive outcomes). Supporting previous results, the best model was a WMH model including a fast, WM-like process that integrated negative outcomes, but an outcome insensitive slower learning component (Fig. 4 right, supplementary figure 13). This WMH model also fit better than the best single process model, and captured the qualitative pattern of learning curves (Fig. 4B), bottom left).

RL-like policy with a simpler H algorithm

Our results show that behavior that is typically modeled with RL algorithms appears to instead be generated with non-RL processes, including a fast, forgetful, and capacity-limited process that integrates outcome valence, and a slow and resource-unlimited "H" process that encodes association strengths between stimuli and actions, irrespective of out-

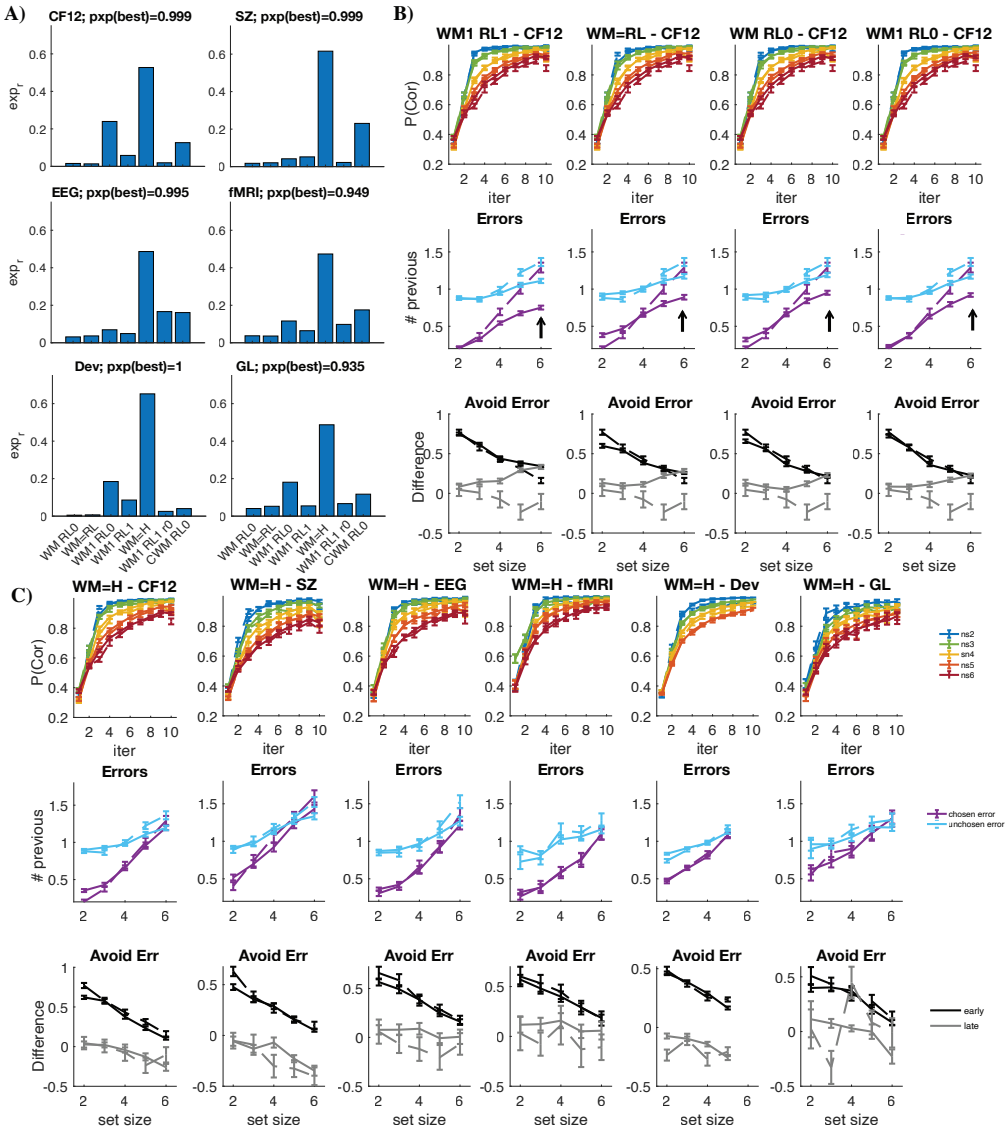


Figure 2

A) Model comparison yields similar results in all 6 RLWM data sets, with the novel association model WM=H winning over the previous best model. Bars represent model frequency in the populations; all corrected exceedance probabilities are > 0.93 *for WM=H. B) Varying bias parameterization within the RL-WM family of models improves fit compared to previous models, by capturing spread in learning curves better (top); however, the models cannot capture the pattern of errors (middle row). Bottom: Difference in past number of chosen vs. unchosen error in error trials for early (iteration 1-5, black) vs. late (iterations 6 and above) is not captured by any model. Models are illustrated on dataset CF12; see supplement for other data sets. Dashed lines are empirical data; full lines model simulations. C) Winning model WM=H captures patterns of behavior better in all 6 data-sets. Top: The spread in learning curves across set sizes is better captured. Middle: The new model captures the qualitative pattern of errors, such that in large set sizes, participants' errors are not dependent on their history or negative outcomes. Bottom: Neglect of negative feedback pattern differs in early (iterations 1-5) and late (iterations 6 and above) parts of learning; WM=H model captures this dynamic. Models are indexed by their modules (WM, RL or H; see methods); the bias term within their module (0 indicates* $\alpha- = 0$*; 1 indicates* $\alpha_- = \alpha_+$*, no number indicates a free parameter; = indicates a shared free parameter);* r_0 *indicates a free parameter for the 0 outcome in RL; C indicates use of policy compression.*

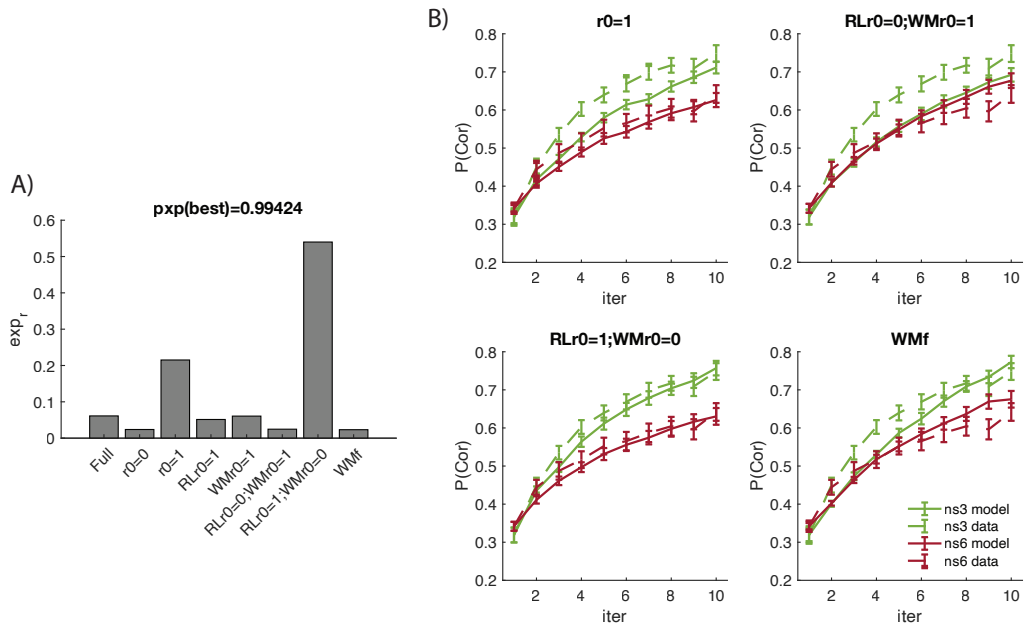


Figure 3

A) Model comparison. Family of models manipulating subjective outcome value of outcome 0 r_0 for RL, WM or both - with r_0 a free parameter unless labeled to its fixed value. $r_0 = 0$ corresponds to standard RL or WM computations; $r_0 = 1$ corresponds to an H agent that handles both outcomes similarly. The winning model $RLr_0 = 1; WMr_0 = 1$ assumes RL $r_0 = 1$, WM $r_0 = 0$, and is thus a WMH agent, replicating findings in the deterministic version of the task. We further verified that the winning model was better than the best single process model WMf (see methods). B): A set size effect is also observed in a probabilistic version of the task; the winning model (bottom left) captures the learning curve pattern better than the competing models.

come valences. This leaves two questions open: A) What is the computational function of this slow process, and B) why is it mistaken for value-based RL, for example in previous RLWM modeling Collins and Frank (2012); Zou et al. (2022)? Indeed, on its own, the slow "H" process cannot learn a good policy, but only tends to repeat previous actions, and thus seems functionally maladaptive. To investigate this question, we simulated both RLWM and WMH models in a standard probabilistic two-armed bandit task, varying the probability p of a reward for the correct choice (see Fig. 4 left, methods). RL policies track this value, and thus convert to a graded policy where the agent is more likely to select the correct bandit the higher p (green curve in fig. 4 right). By contrast, an H agent on its own performs at chance, no matter p (blue curve in fig. 4 right, $\rho_{WM} = 0$). However, when the agents' choices invoke a mixture of policies, including a WM policy that tracks outcomes, the policy learned by the H-agent does resemble a standard RL policy (dark blue curves). Indeed, even with low WM weights (e.g., $\rho_{WM} = 0.5$), WM's contribution is enough to bootstrap choice selection of the good option, which leads the H agent to select

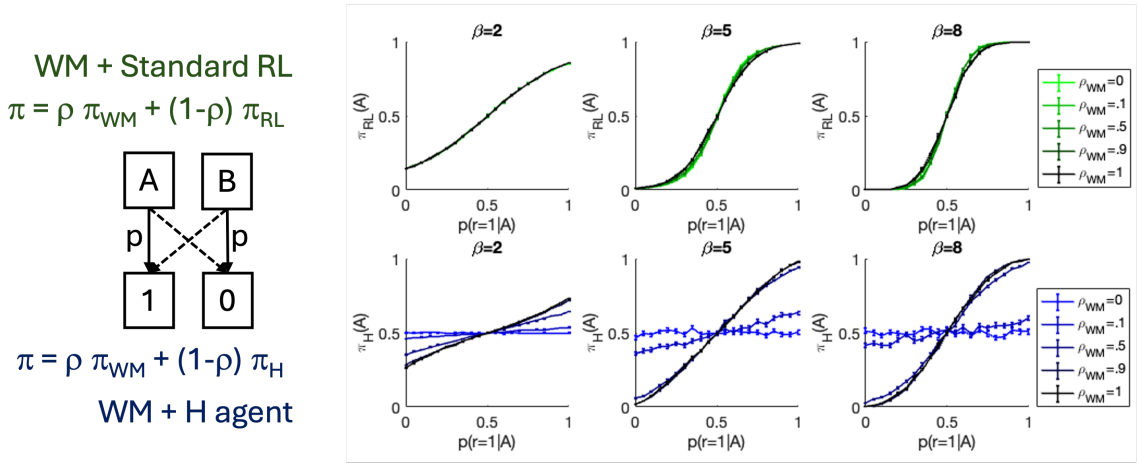


Figure 4

Left: We simulate RLWM (top) or WMH (bottom) mixture agents on a simple probabilistic 2-armed bandit task. Right: the policy learned by the H-agent (bottom) resembles an RL policy (top) when there is enough WM contribution to choices, in a probabilistic 2-armed bandit task. We vary parameter ρ indicating the contributions of the WM module, and β indicating the noise in the softmax policy.

this action more often and thus develop a good policy. This simulation shows that in the absence of specific task features decorrelating contributions of rewards from contributions of errors to behavior (such as the ability to consider multiple errors, something not feasible in most binary choice tasks), contributions of an H agent might be mistaken for an RL policy. Furthermore, in this mixture context, which likely corresponds to most human learning, we observe that the H agent does implement an adaptive policy with a simpler learning rule than the RL process.

Discussion

We analyzed 6 previously published data sets and one novel data set to investigate how different processes contribute to reward-based learning in humans. Such learning had previously been explained with model-free RL algorithms, which use a cached value estimate integrated past reward outcomes for given stimuli and actions to guide decisions. Behavioral analyses gave strong evidence across 6 datasets that integration of outcomes to guide future decisions is dependent on load, and becomes weak or absent in higher set sizes. Computational modeling revealed that this pattern could only be explained by a mixture model, with two very distinct processes. The first, a working-memory like process that learns fast but is limited both in the amount of information and duration it can be held, appeared to successfully integrate reward outcomes into its policy. By contrast, a second, slower but less limited process, appeared to fully neglect outcomes, updating in the same direction for wins and losses, and thus only tracked association strengths, in what could be likened to a hebbian or habit-like process (H-agent).

Notably, although reward-based learning is, at first glance, well approximated by

RL algorithms, neither of these processes correspond to what is typically thought of as an RL cognitive process. The fast (WM) process integrates outcome values into a policy as an RL algorithm should, but has properties not typically associated with RL, such as capacity limitations and rapid forgetting. By contrast, the slow, unlimited H-process is more in line with what is typically thought of as RL along those dimensions, but does not rely on reward-prediction errors – and indeed does not approximate values – as is typically expected from model-free RL algorithms in the context of cognitive neuroscience (Niv, 2019; Sutton & Barto, 2018).

We showed with simulations that the H-agent, despite its learning rule that is on its own unsuited to learning from rewards, is nonetheless able to develop appropriate policies within a mixture model context. Indeed, using WM to bootstrap adaptive choice selection leads the agent to more frequently select actions avoiding bad outcomes, which further enables it to select good actions and reinforce them. This agent, which is mathematically equivalent to a stimulus-dependent choice perseveration kernel (Toyama et al., 2023), is reminiscent of the "habits without value" model (Miller, Shenhav, & Ludvig, 2019), which showed similar properties of developing good policies without value tracking. Here, our model extends the same theoretical approach to a stimulus-dependent learning context, and we validated experimentally the usefulness of this approach across 7 datasets. The H-agent uses a simpler learning rule to learn a similar policy to an R-agent in a mixture context, which might be a more resource-rational way to lead to adaptive behavior.

An important question concerns the generalizability of this finding to other learning tasks. Is it possible that the RLWM task, with deterministic feedback, incites participants to de-activate RL-like processes? While it is a possible explanation, we think it is unlikely. First, RL is not typically thought to be under explicit meta-control, but rather to occur implicitly in the background (Cortese, Lau, & Kawato, 2020; Pessiglione et al., 2008), thus it is unclear why this would not be the case here. Second, computational modeling supports similar conclusions in the probabilistic version of RLWM-P, where integrating reward outcomes over multiple trials is useful. We limit our investigation here to the RLWM experimental framework because it offers a solid grounding for factoring out explicit working-memory processes, and analyzing what remains. It remains an important future research direction to find experimental and modeling approaches that will better allow us to parse out different processes, including WM, from learning behavior, and to probe the generalizability of this finding to other tasks typically modeled with RL.

Another important question concerns the brain mechanisms underlying the processes identified here through modeling. A reason for the success of RL frameworks is their ability to map on to brain mechanisms in striato-cortical loops with dopaminergic signaling, including for example RL reward-prediction errors in striatal bold signal (Daniel and Pollmann (2014), fig. 5). If learning from reward in humans appears RL-like at a first approximation, but actually reflects two non-RL processes, how can we reconcile this with a wealth of RL model-based neuroscience findings? One possibility is that most of human reward-based learning tasks tap on WM processes that are at first an approximation well described by RL (as here in the RLWM-P dataset), such that the striatal circuits support a more cognitive, explicit version of RL than typically assumed; in parallel, the H-agent might reflect hebbian cortico-cortical associations (Frank et al., 2004). Another possibility is that value learning in striatal-based networks does occur, but does not strongly con-

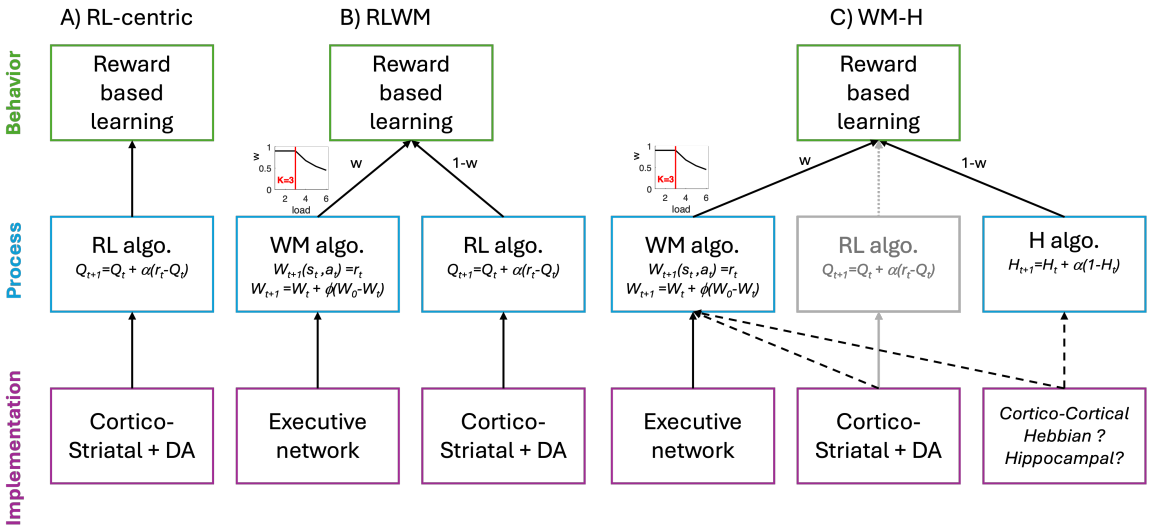


Figure 5

A)Standard RL-centric approaches to reward-based learning assume a close correspondance between the equations that govern behavior and their implementation in a cortico-striatal, dopamine-dependent (DA) network Eckstein et al. (2021) B) The RLWM framework factors out contributions of working memory, which are very high under low load, and decrease when load is under high capacity. C) Our results reveal no influence of RL-like computations on behavior (dotted line), but instead highlight an important H-agent contributing to reward-based learning. Underlying neural substrates are speculative (dashed lines). See methods for model equations and parameters.

tribute to behavior in many human experiments. Further research will necessitate careful task design, modeling, and concurrent imaging to unconfound possible RL processes from other learning processes such as WM and H, and further our understanding of their neural correlates. Patient studies in the RLWM domain, including with lesion patients or patients with dopaminergic medications targeting striatum should help shed light on these questions.

Our findings have important implications. First, they strengthen mounting evidence that RL modeling in reward-based learning tasks is useful but fraught (Eckstein et al., 2022, 2021). While RL models capture much variance of learning behavior, our findings hint that it does so often without actually capturing the dynamical cognitive processes that support behavior. In addition to blurring our theoretical understanding, this may in practice lead to misinterpretations when RL models are used for model-based analysis of neural signals (Cohen et al., 2017; Collins, Ciullo, et al., 2017; Rac-Lubashevsky, Cremer, Collins, Frank, & Schwabe, 2023), or when fit RL parameters are used as mechanistically interpretable proxys for individual differences, for example in developmental and aging research (Eckstein et al., 2022; Nussenbaum & Hartley, 2019; Rmus et al., 2023) or computational psychiatry (Collins, Albrecht, et al., 2017; Collins et al., 2014; Montague, Dolan, Friston, & Dayan, 2012; Zou et al., 2022).

Second, our findings further highlight the fact that, beyond elegant parsimonious

single process accounts of behavior or broad dual process ones, cognitive research has established a vast knowledge of multiple separable processes that support decision making, including explicit memory processes such as working memory. Even simple tasks designed to elicit a target process (such as bandit tasks for RL) recruit multiple other processes, but those processes may be unidentifiable in such tasks. Disentangling multiple processes requires considering more complex tasks to elicit differentiable behavior. Future research in learning and behavior should consider the parsimony/complexity trade-off carefully within the context of our knowledge of the complexity of human behavior.

In conclusion, our findings reveal that when learning from rewards, humans use effortful active maintenance of information to guide good choices in the short term, and rely on iteration of choices over time to build a good policy, bootstrapped by limited memory. We find no evidence here of standard value-based RL contribution to learning, and falsify the predictions of models that do includ RL. These findings call for care in interpreting any RL-based findings with important implications for behavioral, clinical, developmental, and neuro-cognitive scientists.

References

Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature neuroscience*, *20*(7), 997–1003.

Cohen, J. D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., ... others (2017). Computational approaches to fmri analysis. *Nature neuroscience*, *20*(3), 304–313.

Collins, A. G. (2018a). Learning structures through reinforcement. In *Goal-directed decision making* (pp. 105–123). Elsevier.

Collins, A. G. (2018b). The tortoise and the hare: Interactions between reinforcement learning and working memory. *Journal of cognitive neuroscience*, *30*(10), 1422–1432.

Collins, A. G., 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.

Collins, A. G., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working memory contributions to reinforcement learning impairments in schizophrenia. *Journal of Neuroscience*, *34*(41), 13747–13756.

Collins, A. G., Ciullo, B., Frank, M. J., & Badre, D. (2017). Working memory load strengthens reward prediction errors. *Journal of Neuroscience*, *37*(16), 4332–4342.

Collins, A. G., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, *21*(10), 576–586.

Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035.

Collins, A. G., & 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*, *115*(10), 2502–2507.

Cortese, A., Lau, H., & Kawato, M. (2020). Unconscious reinforcement learning of hidden brain states supported by confidence. *Nature communications*, *11*(1), 4429.

Daniel, R., & Pollmann, S. (2014). A universal role of the ventral striatum in reward-based learning: evidence from human studies. *Neurobiology of learning and memory*, *114*, 90–100.

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.

Daw, N. D., & Tobler, P. N. (2014). Value learning through reinforcement: the basics of dopamine and reinforcement learning. In *Neuroeconomics* (pp. 283–298). Elsevier.

Eckstein, M. K., Master, S. L., Xia, L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. (2022). The interpretation of computational model parameters depends on the context. *Elife*, *11*, e75474.

Eckstein, M. K., Wilbrecht, L., & Collins, A. G. (2021). What do reinforcement learning models measure? interpreting model parameters in cognition and neuroscience. *Current opinion in behavioral sciences*, *41*, 128–137.

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.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annual review of psychology*, *68*, 101–128.

Lai, L., & Gershman, S. J. (2021). Policy compression: An information bottleneck in action selection. In *Psychology of learning and motivation* (Vol. 74, pp. 195–232). Elsevier.

Master, S. L., Eckstein, M. K., Gotlieb, N., Dahl, R., Wilbrecht, L., & Collins, A. G. (2020). Disentangling the systems contributing to changes in learning during adolescence. *Developmental cognitive neuroscience*, *41*, 100732.

McDougle, S. D., & Collins, A. G. (2021). Modeling the influence of working memory, reinforcement, and action uncertainty on reaction time and choice during instrumental learning. *Psychonomic bulletin & review*, *28*, 20–39.

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological review*, *126*(2), 292.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of neuroscience*, *16*(5), 1936–1947.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences*, *16*(1), 72–80.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

Niv, Y. (2019). Learning task-state representations. *Nature neuroscience*, *22*(10), 1544–1553.

Niv, Y., & Langdon, A. (2016). Reinforcement learning with marr. *Current opinion in behavioral sciences*, *11*, 67–73.

Nussenbaum, K., & Hartley, C. A. (2019). Reinforcement learning across development: What insights can we draw from a decade of research? *Developmental cognitive neuroscience*, *40*, 100733.

Palminteri, S., & Lebreton, M. (2022). The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in cognitive sciences*, *26*(7), 607–621.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in cognitive sciences*, *21*(6), 425–433.

Pessiglione, M., Petrovic, P., Daunizeau, J., Palminteri, S., Dolan, R. J., & Frith, C. D. (2008). Subliminal instrumental conditioning demonstrated in the human brain. *Neuron*, *59*(4), 561–567.

Rac-Lubashevsky, R., Cremer, A., Collins, A. G., Frank, M. J., & Schwabe, L. (2023). Neural index of reinforcement learning predicts improved stimulus–response retention under high working memory load. *Journal of Neuroscience*, *43*(17), 3131–3143.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *Neuroimage*, *84*, 971–985.

Rmus, M., He, M., Baribault, B., Walsh, E. G., Festa, E. K., Collins, A. G., & Nassar, M. R. (2023). Age-related differences in prefrontal glutamate are associated with increased working memory decay that gives the appearance of learning deficits. *Elife*, *12*, e85243.

Rmus, M., McDougle, S. D., & Collins, A. G. (2021). The role of executive function in shaping reinforcement learning. *Current Opinion in Behavioral Sciences*, *38*, 66–73.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Sugawara, M., & Katahira, K. (2021). Dissociation between asymmetric value updating and perseverance in human reinforcement learning. *Scientific reports*, *11*(1), 3574.

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

Tai, L.-H., Lee, A. M., Benavidez, N., Bonci, A., & Wilbrecht, L. (2012). Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nature neuroscience*, *15*(9), 1281–1289.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, *55*(4), 189.

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.

Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in pavlovian conditioning: Application of a theory. *Inhibition and learning*, 301–336.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *Elife*, *8*, e49547.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European journal of neuroscience*, *19*(1), 181–189.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial

striatum in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 513–523.

Yoo, A. H., & Collins, A. G. (2022). How working memory and reinforcement learning are intertwined: A cognitive, neural, and computational perspective. *Journal of cognitive neuroscience*, *34*(4), 551–568.

Yoo, A. H., Keglovits, H., & Collins, A. G. (2023). Lowered inter-stimulus discriminability hurts incremental contributions to learning. *Cognitive, Affective, & Behavioral Neuroscience*, *23*(5), 1346–1364.

Zou, A. R., Muñoz Lopez, D. E., Johnson, S. L., & Collins, A. G. (2022). Impulsivity relates to multi-trial choice strategy in probabilistic reversal learning. *Frontiers in Psychiatry*, *13*, 800290.

Methods

Experimental design

All datasets were previously published (Collins, 2018a; Collins et al., 2014; Collins, Ciullo, et al., 2017; Collins & Frank, 2012; Master et al., 2020), except data-set #6 *GL*. All experiments relied on the RLWM protocol developed in (Collins & Frank, 2012), with minor variations to the protocol across data-sets. We first describe the shared components of the RLWM task, then describe specific details.

Shared

In all experiments participants' goal was to learn stimulus-action associations using truthful, binary feedback (correct/incorrect or +1/0). Actions corresponded to one of three adjacent key presses (or play console button presses). Each experiment included multiple independent blocks requiring learning for a novel set of easily identifiable stimuli.

Within each block, stimuli were presented for 10-15 iterations depending on the specific experiment, in an interleaved fashion. The number of stimuli (or set-size ns) was manipulated across blocks, and varied between 2 and 6; this key manipulation enabled us to affect load and thus identify working memory contributions. The stimulus presentation order was pseudo-randomized to control for the delay between two successive iterations of the same stimuli, with a close to uniform distribution between 1 and $2 * ns - 1$. This was important to identify the forgetting component of WM. The number of blocks ranged from 10 to 22 depending on the experiment.

Stimuli were presented for a short period (typically 1.5s, depending on the specific experiment), during which the participant made a key press; this was followed by a short feedback (.5-1s), then a short ITI (typically .5s, but see details of each published dataset). Stimuli within one block consisted of highly discriminable and familiar exemplars of a category (e.g., a cat, a cow, an elephant and a tiger in the animal category).

Participants' instructions fully described the structure of the task, including the fact that feedback was truthful and correct stimulus-action associations did not change within a block.

Published data-sets 1-5

- Dataset CF12 (Collins & Frank, 2012) included $N = 79$ participants who performed the RLWM task in person. Set sizes ranged 2-6, for a total of 18 blocks.

- Dataset SZ (Collins et al., 2014) included $N = 85$ participants who performed the RLWM task in person, including patients with schizophrenia and matched controls. To accommodate patients, trial dynamics were slower; to keep the task within a shorter duration, the number of blocks was decreased to 13. See published methods for details.

- Dataset EEG (Collins & Frank, 2018) included $N = 40$ participants who performed the RLWM task in person while wearing an EEG measurement cap. There were 22 blocks.

- Dataset fMRI (Collins & Frank, 2012) included $N = 26$ participants who performed the task in the scanner. To accommodate fMRI timing constraints, the ITI durations were jittered, resulting in fewer blocks (18).

- dataset DEV (Master et al., 2020) included $N = 300$ participants age $8 - 30$ who performed the task in person. To accommodate younger participants, the maximum set size was $ns = 5$, and the number of blocks was reduced to 10. Participants used a game console with three buttons instead of a keyboard.

Data-set 6 GL

In this unpublished data-set, $N = 30$ participants performed a variant of the experiment where half of the blocks were "gain" blocks and half were "loss" blocks. In G blocks, participants tried to gain points, using feedback "+1 vs. 0". In L blocks, participants tried to avoid losing points, using feedback "0" vs. "-1" respectively for the correct choice vs. the two incorrect ones for each stimulus. There were 18 blocks, and 15 iterations per stimuli. We observed no difference in behavior between the G and L blocks, and computational modeling did not uncover any differences either (i.e., making any parameter from the winning model dependent on block condition did not improve fit). For the purpose of behavioral and modeling analyses in this paper, outcomes $0/-1$ in the loss blocks were treated as correct/incorrect in the same way as outcomes $1/0$ in gain blocks.

RLWMP experiment

The RLWM-P experiment was a variant of the RLWM experiment with probabilistic feedback. Previous analysis confirmed a set size effect, showing WM involvement even when learning in a probabilistic context (experiment 3 in (McDougle & Collins, 2021)). In this experiment, selecting the correct action led to positive feedback with probability $p = 0.92$ or $p = 0.77$ across blocks, while selecting the incorrect action led to negative feedback with the same probability. Participants ($N = 34$) were informed of the probabilistic nature of the task. Participants only experienced two set sizes across 14 blocks (8 vs. 6 for $ns = 3$ and $ns = 6$) with 12 iterations per stimuli.

Participants

All procedures were approved by institutional review boards where data was collected (including the Committee for the Protection of Human subjects at the University of California, Berkeley, for unpublished dataset #6 GL). Participants provided informed consent and were free to stop participation at any time of their choosing. Please refer to corresponding publications for further participant and procedures details.

Behavioral analysis

Set size effects on accuracy

We visualize data for each dataset using the same learning curve as in previously published analyzes, where the average choice accuracy is plotted as a function of specific stimulus iteration number, separately for each set size.

Error analysis

To investigate the effect of negative outcomes on behavior, we designed a novel error trial analysis. We reasoned that if participants integrated negative feedback into their policy, they should be less likely to repeat a previous error. There were two possible errors for each stimulus (e.g., A2 is correct for the triangle stimulus in fig. 1 A), then A1 or A3 are possible errors; the A3 error should be more likely after A1 is tried and receives incorrect feedback). Thus, if participants performed an error E_t for stimulus S_t, we counted how many times the participant had made the same error for stimulus S_t up to trial $t-1$ ("chosen error"), and how many times they had made the other possible error ("unchosen error"); this corresponds to the blue and purple curves in figures 1 and 2. To measure success at avoiding error, we also compute the average error avoidance success by substracting the number of previous unchosen errors from the number of previous chosen errors (black and grey curves in figures 2).

Computational modeling

Model fitting

We used Matlab 2020B with `fmincon` optimization to fit all computational models, with 10 random starting points per participant and capacity (for discrete capacity models). We sought parameters that optimized the log-likelihood of the data under the model assumptions (Wilson & Collins, 2019).

Model comparison

Exceedance probability. For model comparison, we computed the protected exceedance probability using the `spm_bmf` function (Rigoux, Stephan, Friston, & Daunizeau, 2014), separately within each dataset, and also report model responsibility `exp_r`. Where comparable (datests 1-6), we observed highly convergent best results (see Fig. 6).

Model space exploration. Because of the breadths of potential model space, we limited model space exploration to sequential families as described below. We performed model comparison within a model family; selected the best model out of each family; then performed model comparison again between winning models.

Model validation. To validate the winning model vs. competing models, we simulated winning models with fit parameters, with 20 simulated agents per participant. Summary statistics of interest (e.g., learning curves, error analysis) were averaged over agents within participants first, to average out stochasticity in simulations. We then plotted the resulting synthetic data-set behavior across participants in the same way we plotted participants' behavior (including mean and SEM across synthetic participants).

Checks

Model identifiability. We performed model identifiability analyses within the key models of interests that represent theoretically interesting contrast (Wilson & Collins, 2019). We ensured that competing models were identifiable with confusion matrices (see supplementary fig. 7).

Computational models - RLWM

Mixture model

Previous work (Collins, 2018a, 2018b; Collins et al., 2014; Collins, Ciullo, et al., 2017; Collins & Frank, 2012, 2018; Master et al., 2020; Rac-Lubashevsky et al., 2023) showed that behavior in the RLWM task cannot be adequately captured with a single process model. We used the RLWM modeling framework as a baseline, which assumes that policy is the mixture of a working memory policy, designed to capture fast but forgetful information integration, and a non-forgetful integrative process, typically RL.

$$\pi_{mixture}(a|s) = \rho_{WM}\pi_{WM}(a|s) + \rho_{WM}\pi_{other}(a|s)$$

where "other" is typically RL. The mixture weight $\rho_{WM}(ns)$ is set-size dependent and serves to capture resource or capacity limitations of the WM process. In the context where set size is $\in \{2, ..., 6\}$, the mixture weight is set to $\rho_{WM} = \rho \min(1, ns/K)$ where $K \in \{2, ..., 5\}$ is a capacity parameter, and $\rho \in [0, 1]$ regulates the overall balance of WM vs. non WM in the policy. If there are only two set sizes, the mixture weight is parameterized per set size ($\rho_{WM} = \rho_3, \rho_6$).

This full policy is typically mixed with a uniform random policy to capture random lapses in choices to produce the final full policy, with noise parameter $\epsilon \in [0, 1]$:

$$\pi(a|s) = (1 - \epsilon)\pi_{mixture} + \epsilon\frac{1}{n_A}$$

WM module

The WM module tracks information in an association weight matrix initialized at the beginning of each block at $W_0 = 1/n_A$ reflecting the initial expectation that one out of $n_A = 3$ actions leads to reward 1 (vs. 0). After observing stimuli, actions and rewards (S_t, a_t, r_t) at trial t, the update is

$$W_{t+1}(S_t, a_t) = W_t(S_t, a_t) + \alpha_{WM}(r_t)(r_t - W_t(S_t, a_t))$$

To capture one-shot encoding of information, we set $\alpha_{WM}(1) = 1$. To capture potential neglect of negative outcomes, we set $\alpha_{WM}(0) = bias_{WM}$ as a parameter, which is either free ($bias_{WM} \in [0, 1]$) or fixed depending on the model considered. To capture short-term maintenance in WM, WM weights are decayed at each trial toward initial values for all (s, a) not observed at t:

$$\forall(s, a), W_{t+1}(s, a) = W_t(s, a) + \phi_{WM}(W_0 - W_t(S, a))$$

The working memory policy transforms the WM weights through a standard softmax:

$$\pi_{WM}(a|s) = \frac{\exp \beta W(s, a)}{\sum_i \exp \beta W(s, a_i)}$$

the temperature parameter β is typically fixed to a high value (here $\beta = 25$) for theoretical reasons (this ensures that WM policy of a repeated trial is perfect) and identifiability reasons (this ensures that the RL learning rate is identifiable and RL an WM modules are separable).

In the absence of a free β parameter, noise in the choice policy is instead parameterized as lapses in the overall policy via parameter ϵ, which is highly recoverable (see supplementary figure 8).

RL module

The RL module is a standard delta-rule agent which tracks Q-values for each stimulus and action pair. Q is initialized at $Q_0 = 1/n_A$, reflecting the initial expectation that one out of $n_A = 3$ actions leads to reward 1 (vs. 0). The delta-rule update is:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_{RL}(r_t)(r_t - Q_t(S_t, a_t))$$

The positive learning rate parameter $\alpha_{RL}(1) \in [0, 1]$ is free, and then negative learning rate $\alpha_{RL}(0) = bias_{RL} \times \alpha_{RL}(1)$ is also parameterized by a bias parameter ($bias_{RL} \in [0, 1]$) which is free or fixed depending on the specific model.

The RL policy transforms the Q-values through a standard softmax:

$$\pi_{RL}(a|s) = \frac{\exp \beta Q(s, a)}{\sum_i \exp \beta Q(s, a_i)}$$

The temperature parameter β is fixed and shared with the WM module (see above).

RL-like module extension (H-agent)

We extend the RL module to novel versions of the algorithm to capture the novel observed error effects which standard RLWM models cannot capture.

Specifically, the H module tracks association weights in a way very similar to an RL module, and is initialized also at $H_0 = 1/nA$. The update is:

$$H_{t+1}(s_t, a_t) = H_t(s_t, a_t) + \alpha_H(r_t)(SR(r_t) - H_t(S_t, a_t))$$

The only difference is the subjective outcome SR, which is fixed at $SR(1) = 1$ for correct outcomes, and parameterized at $SR(0) = r_0$ for incorrect outcomes, with parameter $r_0 \in [0, 1]$, free or fixed depending on the model. With $r_0 = 0$, the H agent reduces to an RL agent. With $r_0 = 1$, the H-agent treats correct and incorrect outcomes exactly identically and increases weights of the selected action no matter the outcome, thus only tracking a function of stimulus-action associations. The learning rate α_H is parameterized in the same way as α_{RL}. The H policy transforms the H-values through a standard softmax:

$$\pi_H(a|s) = \frac{\exp \beta H(s, a)}{\sum_i \exp \beta H(s, a_i)}$$

The temperature parameter β is fixed and shared with the WM module.

H-agents replace RL agents in the standard RLWM mixture policy to form WMH mixtures:

$$\pi_{mixture}(a|s) = \rho_{WM}\pi_{WM}(a|s) + \rho_{WM}\pi_H(a|s)$$

Choice kernels

We explore including different choice kernels in the policy to A) investigate whether it improves model fit B) ensure that such choice kernels cannot account for observed effects. We incorporate the choice kernels in both policies in the mixture.

Sticky choice. This captures stimulus-independent choice perseveration, i.e. the tendency to repeat the same key-press in consecutive trials. Specifically, we implement it within the softmax policy as:

$$\pi_{WM}(a|s) = \frac{\exp\left(\beta W(s,a) + \kappa I(a,a_{t-1})\right)}{\sum_i \exp\left(\beta W(s,a_i)\kappa I(a,a_{t-1})\right)}$$

where $I(a_i, a_j) = 1$ if $i = j$ and 0 otherwise, and $\kappa \in [-1, 1]$ captures a tendency to repeat or switch away from the previous key press. We apply the same approach to $Q-$ and $H-$ agents, with shared parameters.

Regularization. Policy compression adds a choice kernel that favors default actions, e.g. actions that are valid across more stimuli than others (Lai & Gershman, 2021). Specifically, we implement it within the softmax policy as

$$\pi_{RL}(a|s) = \frac{\exp\left(\beta Q(s,a) + \tau\tilde{Q}(a)\right)}{\sum_i \exp\left(\beta Q(s,a_i)\tau\tilde{Q}(a_i)\right)}$$

where $\tilde{Q}(a) = \text{mean}_i(Q(s_i, a))$. We apply the same approach to $WM-$ and $H-$ agents with shared parameters.

Model space

The model space resulting from the factorial combination of all considered mechanisms is too large too explore. We first considered mechanisms that may absorb variance of no current theoretical interest, and asked whether adding them to the starting, best-so-far RLWM model (based on (Master et al., 2020)) could improve fit. Specifically, we validated that sticky-choice and ϵ-noise in the policy systematically improved fit across data sets, but policy compression didn't (and could not capture qualitative patterns of behavior, see fig. 2).

We thus explored two families of models systematically:

- we first systematically explored the RLWM model (including free κ and ϵ parameters) with bias parameters $bias_{RL}$ and $bias_{WM}$ either free, fixed to 0, fixed to 1, or shared, for a total of 10 models (see supplementary figure 6 for model comparison). The best two models of this family (WM RL0 and WM1 RL0) both have fixed $bias_{RL} = 0$ (thus no update in RL after negative outcomes) and $bias_{WM}$ either free or fixed to 1, thus limited learning bias in WM. In particular, they outperform the published baseline RL=WM model where a single bias parameter is shared (Master et al., 2020).

- we then systematically explored the WMH model with r_0 free or fixed to 0 (same as RL) or 1, and free or fixed bias parameters. The winning model has fixed $r_0 = 1$ (pure H-agent with subjective outcome $SR(0) = SR(1)$), and shared free parameter $bias_{WM} = bias_H$.

- we additionally explored adding a *policy compression* mechanism to all models; the winning model from the corresponding family is labeled with "C". This did not improve fit and could not explain error patterns.

Models included in the model comparison figure 2 are listed below. All include at least 6 free parameters for WM capacity K, WM weight ρ, WM decay ϕ, noise ϵ, sticky choice κ, H or RL learning rates α_H or $\alpha_R L$:

1. WM RL0: RLWM model with free $bias_{WM}$, fixed $bias_{RL} = 0$. Total 7 free parameters.
2. WM=RL: RLWM model with free $bias_{WM} = bias_{RL}$. Total 7 free parameters. This model corresponds to (Master et al., 2020) with an additional sticky choice mechanism which improved fit.
3. WM1 RL0: RLWM model with fixed $bias_{WM} = 1$, $bias_{RL} = 0$. Total 6 free parameters.
4. WM1 RL1: RLWM model with fixed $bias_{WM} = 1$, $bias_{RL} = 1$. Total 6 free parameters. This model is the "no bias" model.
5. WM=H: Overall winnning WMH model with free $bias_{WM} = bias_H$. Total 7 free parameters.

WM1 RL1 r0 : RLWM model with fixed $bias_{WM} = bias_{RL} = 1$, free RL $SR(0) = r_0$. Total 7 free parameters. This model captures qualitative behavior similarly to the WM=H model, because the r_0 parameter is fit to a high value, similar to an H agent

CWMRL0 : best model in the policy compression RLWM family, with 7 parameters including free $bias_{WM}$ and τ parameters.

Computational models - RLWM-P

The computational model for the RLWM-P model was also a mixture model, with a slightly different WM module, and identical RL/H module. In the deterministic experiment, the WM module approximates encoding of the trial information in WM by maintaining relative state-action association weights. In a probabilistic context, by contrast, it is possible that participants hold in mind a hypothesis about the best action, rather than specifically the last trial information. We seek to incorporate this into an extended version of the WM module.

To approximate WM and contrast it to either an RL agent or an H agent, we include the following assumptions:

- we constrain $\rho_{WM}(ns = 6) < \rho_{WM}(ns = 3)$, as a theoretical interpretability constraint ensuring that the WM-labeled module is more expressed under lower load.
- we include forgetting only in the WM module, to associate variance captured with rapid forgetting to the WM-labeled process. This is not a theoretical commitment to RL/H agents not potentially also experiencing decay, but rather that any decay should be stronger in WM, and thus a pragmatic choice to enable identification of the modules.

With these constraints, we use the same formulation as above for WM weights, but we let α_{WM} be a free parameter, such that the WM module might remember only the last trial for a given stim-action-reward (if $\alpha_{WM} = 1$), but might integrate over a few trials otherwise, capturing hypothesis maintenance. In this sense, the WM module is approximated by an RL-like computation with decay, and is forced to contribute more to $ns = 3$ than $ns = 6$.

The full model includes 11 parameters: one per module each of positive and negative learning rates $\alpha(r)$ (4); 2 mixture weight parameters (ρ), one decay parameter ϕ, one noise parameter ϵ, one perseveration parameter κ, and one $SR(0) = r_0$ parameter each (similar to the H module above).

To explore the model space, we first fit the full model, then fixed the r_0 parameter to 0 (standard) or 1 (H-agent) in either the WM or the RL module, or both. The winning model had fixed $r_0(RL) = 1$ (pure H agent) and $r_0(WM) = 0$ (standard WM agent). We next verified that fixing any other parameter (including ρ, κ,ϕ or ϵ or biases) to fixed values did not improve fit over the winning model. Last, we verified that the winning model fit better than a single module model that included all mechanisms and differential noise per set sizes (WMf, fig. 3). We performed model recovery and parameter recovery checks as previously described for RLWM, see supplementary figures 13,12.

Simulations

Environment

To investigate the computational role of an H-like agent, we ran simulations of two mixture agents representing RLWM (mixture of WM and standard RL), and WMH (mixture of WM and no-outcome associative H agent) on a simple probabilistic 2-arm bandit task. Agents chose between two options (A, B) for T trials, and received reward $r = 0/1$ with $P(r = 1|A) = p$ and $P(r = 1|B) = 1-p$. We varied $p \in [0 : .05 : 1]$ and $T \in [20, 50]$. Results were similar for the two learning durations, so we only plotted $T = 50$. We investigated three values of the exploration softmax parameter $\beta \in \{2, 5, 8\}$.

Model

The agents made choices based on the mixture model policy $\pi = \rho_{WM}\pi_{WM} + (1 - \rho_{WM}\pi_{H/RL})$. However, we are interested in the policy learned by the non-WM model in the presence of WM to guide choices, and thus plot π_{RL} and π_H, rather than π.

We approximated a WM process with a simplistic 1-back memory process, such that after each choice $C_t \in \{A, B\}$ and outcome r_t, we updated a working memory associations buffer with $W_{t+1}(C_t) = r_t$. This captures the last reward obtained for each choice, and crudely captures a no-integration, resource limited, short term memory process. WM policy was derived through a softmax transform: $\pi_{WM}(C) \propto \exp(\beta W(C))$.

The standard RL agent tracked the value Q(C) of each choice by updating with a standard delta rule $Q_{t+1}(C) = Q_t(C) + \alpha(r_t - Q_t(C))$. Learning rate parameter was fixed to $\alpha = 0.1$. RL policy was derived through a softmax transform: $\pi_{RL}(C) \propto \exp(\beta Q(C))$.

The associative H agent tracked the association strenght H(C) of each choice by updating with an outcome neglect learning rule $H_{t+1}(C_t) = H_t(C_t) + \alpha(1 - H_t(C_t))$. Learning

rate parameter was fixed to $\alpha = 0.1$; results were similar with other α values. H policy was derived through a softmax transform: $\pi_H(C) \propto \exp(\beta Q H(C))$.

Open science

All code and data will be made available on OSF at the time of publication.

Supplementary Information

Behavior statistics

Table 1

Effect of number of previous chosen vs. unchosen - across all set/sizes; NS6 only

Data Set	Main p-value	Main t-statistic	NS6 p-value	NS6 t-statistic
CF12	8.8073×10^{-24}	$t(78) = -14.4631$	0.10414	$t(78) = -1.6443$
SZ	0.00010146	$t(80) = -4.0918$	0.13665	$t(84) = 1.5028$
EEG	5.3815×10^{-10}	$t(39) = -8.1843$	0.027081	$t(39) = -2.2969$
fMRI	4.6131×10^{-7}	$t(25) = -6.7389$	0.051384	$t(25) = -2.0463$
Dev	5.9833×10^{-25}	$t(279) = -11.3941$	NA	NA
GL	1.4585×10^{-5}	$t(22) = -5.5347$	0.19309	$t(25) = -1.3375$

Table 2

Effect of set-size on chosen-unchosen errors, with mixed effect regression model `DeltaError ~ SetSize + (1|ParticipantID)`.

Data Set	Estimate	(95% CI)	p-value	t-value
CF12	0.14709	$(0.12314, 0.17105)$	9.1625×10^{-29}	12.074
SZ	0.13888	$(0.11189, 0.16587)$	1.1698×10^{-21}	10.113
EEG	0.093142	$(0.062595, 0.12369)$	8.6434×10^{-09}	6.0128
fMRI	0.059083	$(0.0081222, 0.11004)$	0.023418	2.294
Dev	0.057864	$(0.036687, 0.079042)$	9.963×10^{-08}	5.3608
GL	0.055854	$(0.0075775, 0.10413)$	0.02369	2.2884

Table 3

Effect of number of previous chosen vs. unchosen in highest set size - late learning only (iteration>5).

Data Set	p-value	t-value
CF12	0.29737	$t(75) = -1.0494$
SZ	2.9628×10^{-5}	$t(78) = -4.4366$
EEG	0.99329	$t(39) = -0.0084678$
fMRI	0.83304	$t(24) = -0.21311$
Dev	3.6127×10^{-7}	$t(274) = -5.2159$
GL	0.93336	$t(25) = -0.084464$

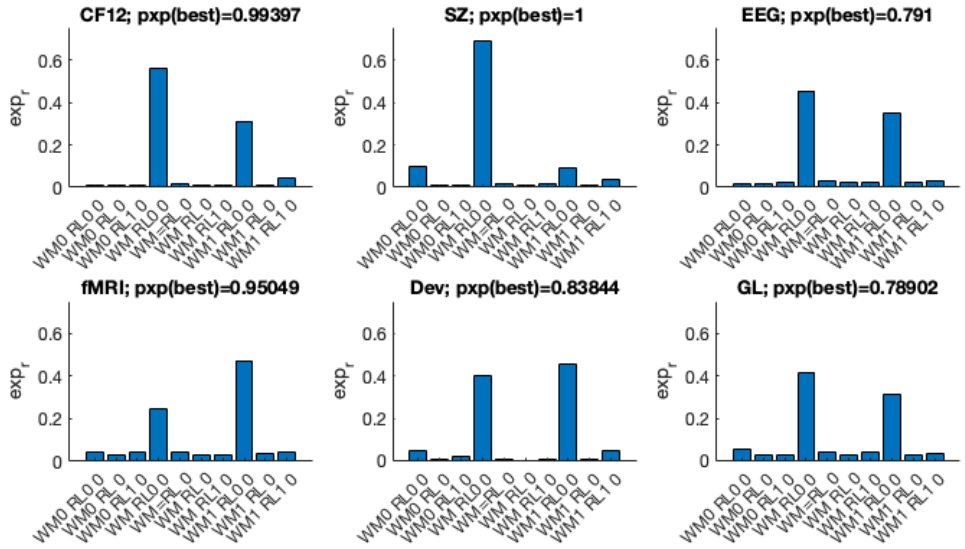


Figure 6

Model comparison within standard RLWM model family. Models are indexed by their modules (WM or RL; see methods); the bias term within their module (0 indicates $\alpha- = 0$; 1 indicates $\alpha_- = \alpha_+$, no number indicates a free parameter; = indicates a shared free parameter). Here, the "0" label at the end indicates fixed $r_0 = 0$.

Model comparison

Model identifiability

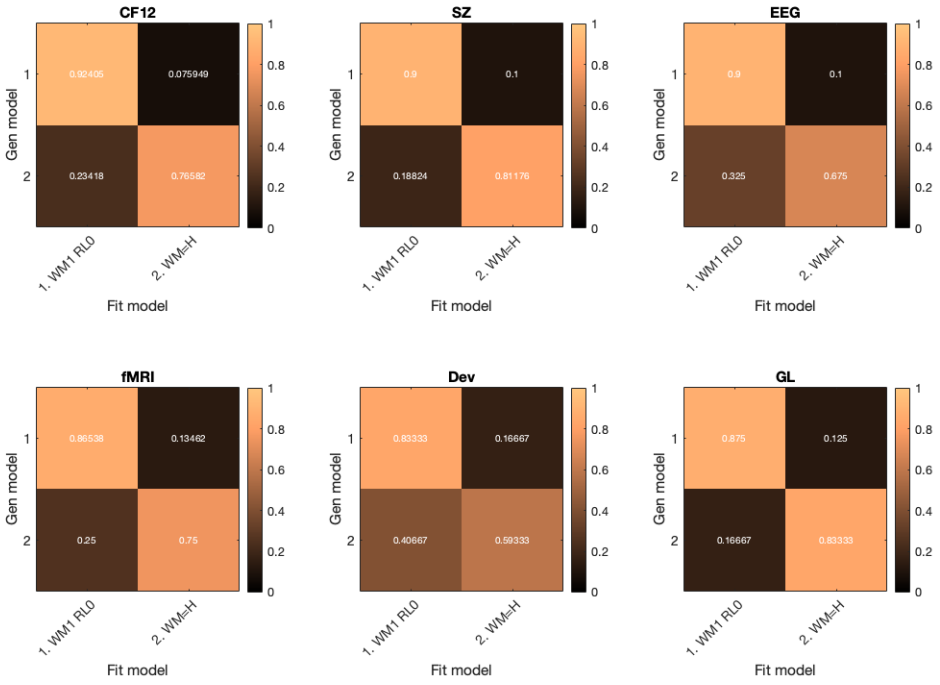


Figure 7

Winning WMH model is identifiable vs. best RLWM model. We simulated artificial datasets using both models and parameters fit on individual participants. For each participants, we simulated enough times to ensure we included at least 100 simulations (e.g., twice per participant in CF12, once in Dev, 4 times in fMRI). We then fit artificial datasets with both models using the same procedure as for real participants. For each dataset, we assign a winning model as the model with the lowest AIC. We verified that BIC overpenalized complexity and lead to worse confusion matrices. WMH is highly recoverable, with the lowest recoverability in the Dev data set where the number of blocks and lack of set-size 6 decreases the difference between H and RL agents.

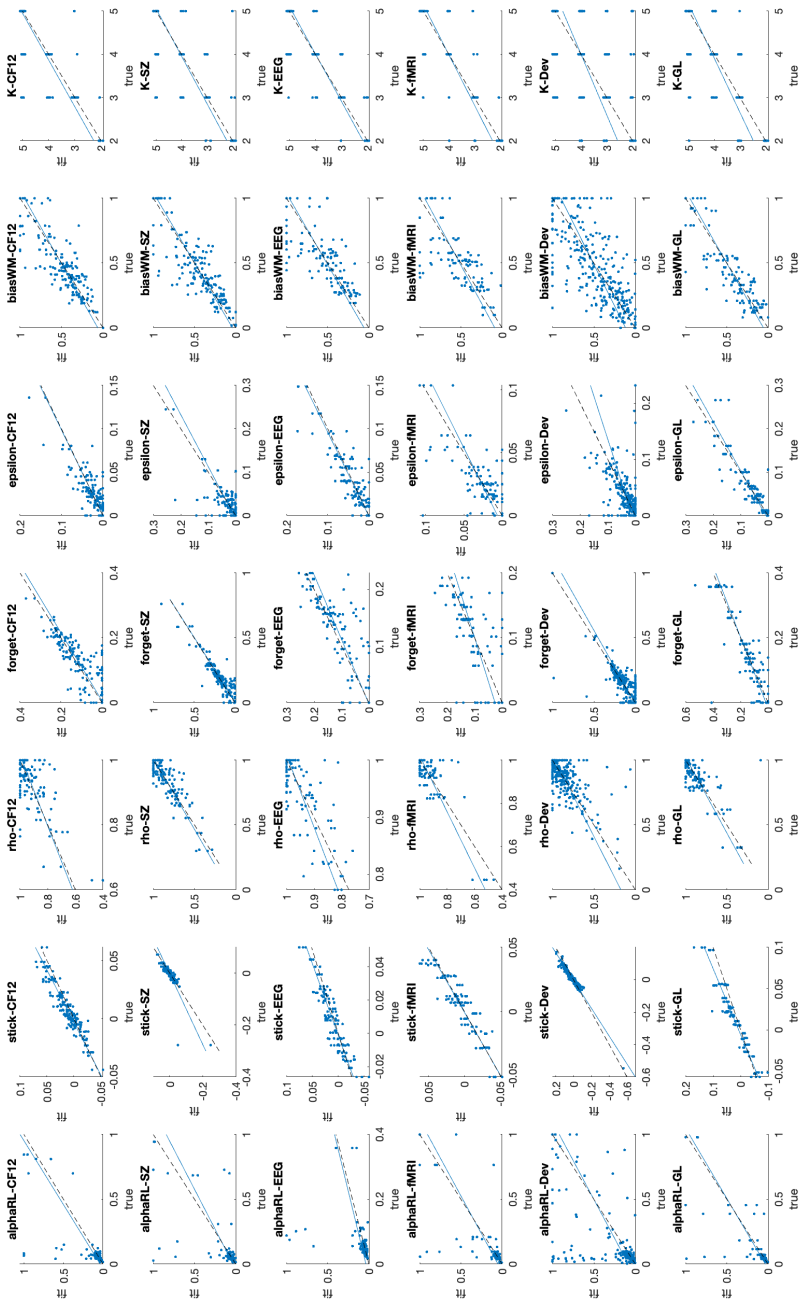
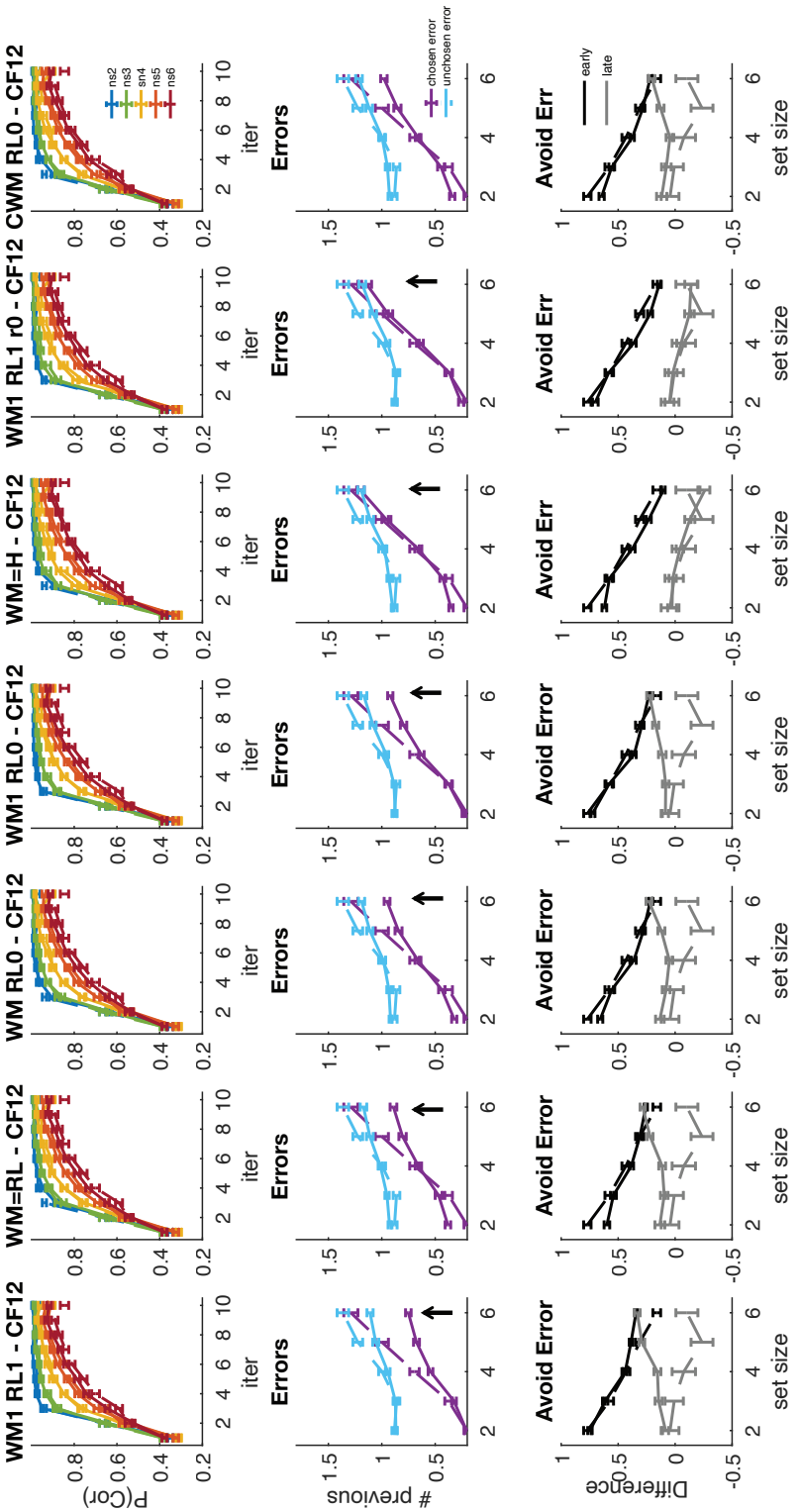

alphaRL-CF12
alphaRL-SZ
alphaRL-EEG
alphaRL-fMRI
alphaRL-Dev
alphaRL-GL
stick-CF12
stick-SZ
stick-EEG
stick-fMRI
stick-Dev
stick-GL
rho-CF12
rho-SZ
rho-EEG
rho-fMRI
rho-Dev
rho-GL
forget-CF12
forget-SZ
forget-EEG
forget-fMRI
forget-Dev
forget-GL
epsilon-CF12
epsilon-SZ
epsilon-EEG
epsilon-fMRI
epsilon-Dev
epsilon-GL
biasWM-CF12
biasWM-SZ
biasWM-EEG
biasWM-fMRI
biasWM-Dev
biasWM-GL
K-CF12
K-SZ
K-EEG
K-fMRI
K-Dev
K-GL
true
est

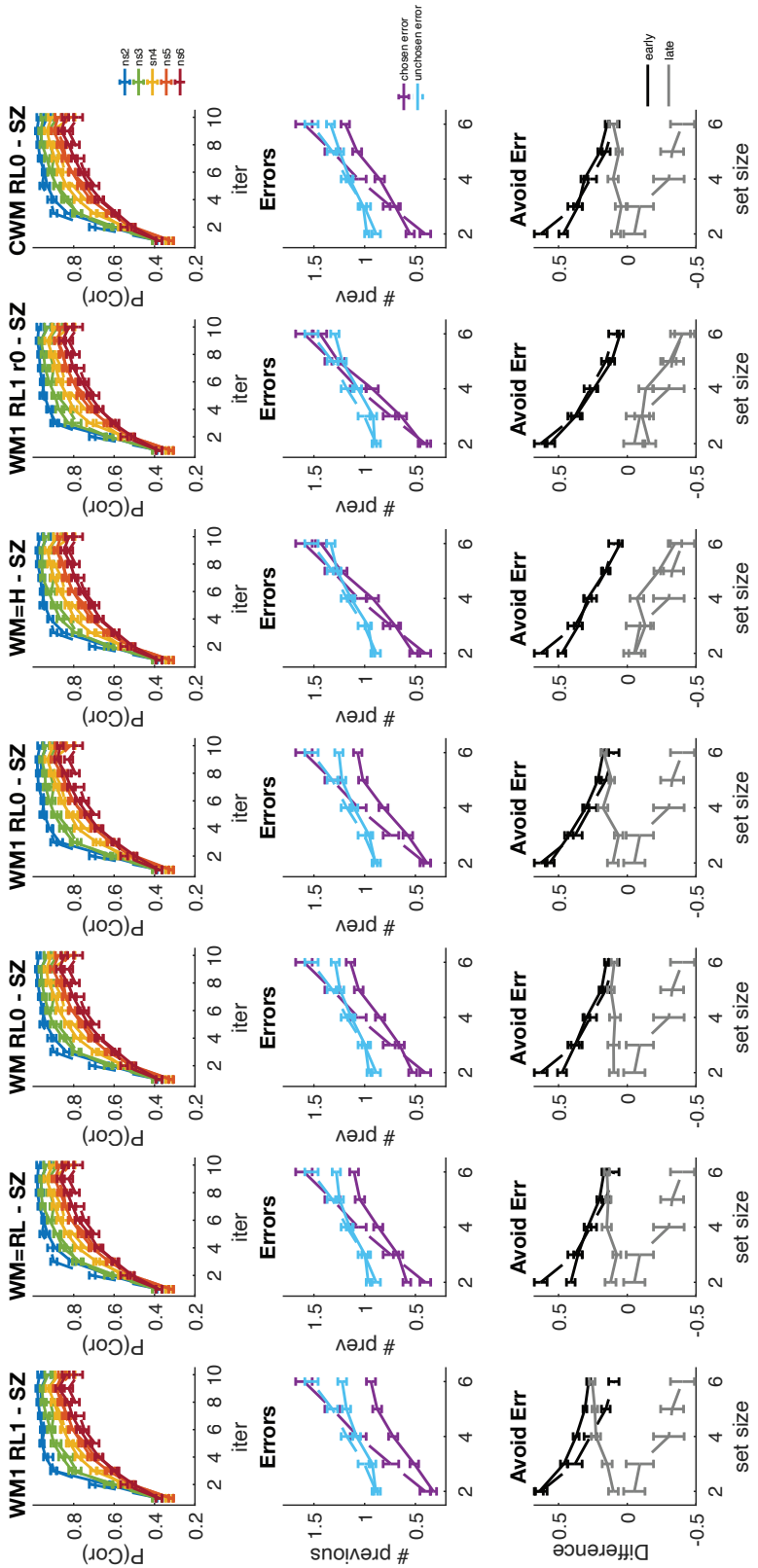
Figure 8

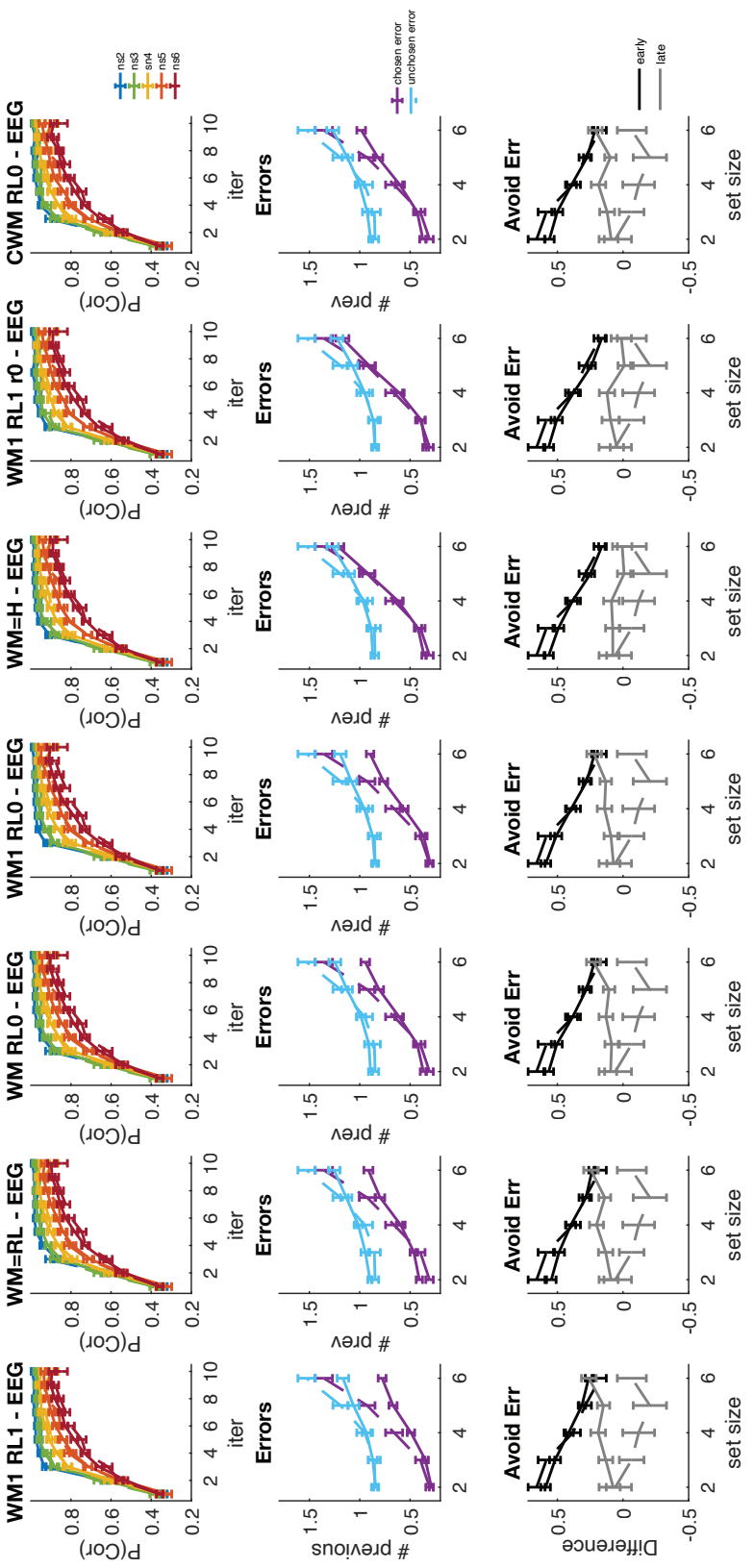
*Parameter recovery analysis. Parameters in the winning WMH model are highly recoverable. We simulated and fit the same number of agents as in Supplementary Fig. 7, and compared true generating parameters with fit recovered parameters, obtaining a high correlation for all parameters in all datasets. Dashed black line is unity line; blue line is least squared regression line. Note that this figure also provides the distribution of best fit parameters across the group in all datasets. For visualization purposes, the discrete capacity parameter was slightly jittered with .05*normal noise. All Spearman rho* $> .56$, $p < 10^{-8}$, *uncorrected.*

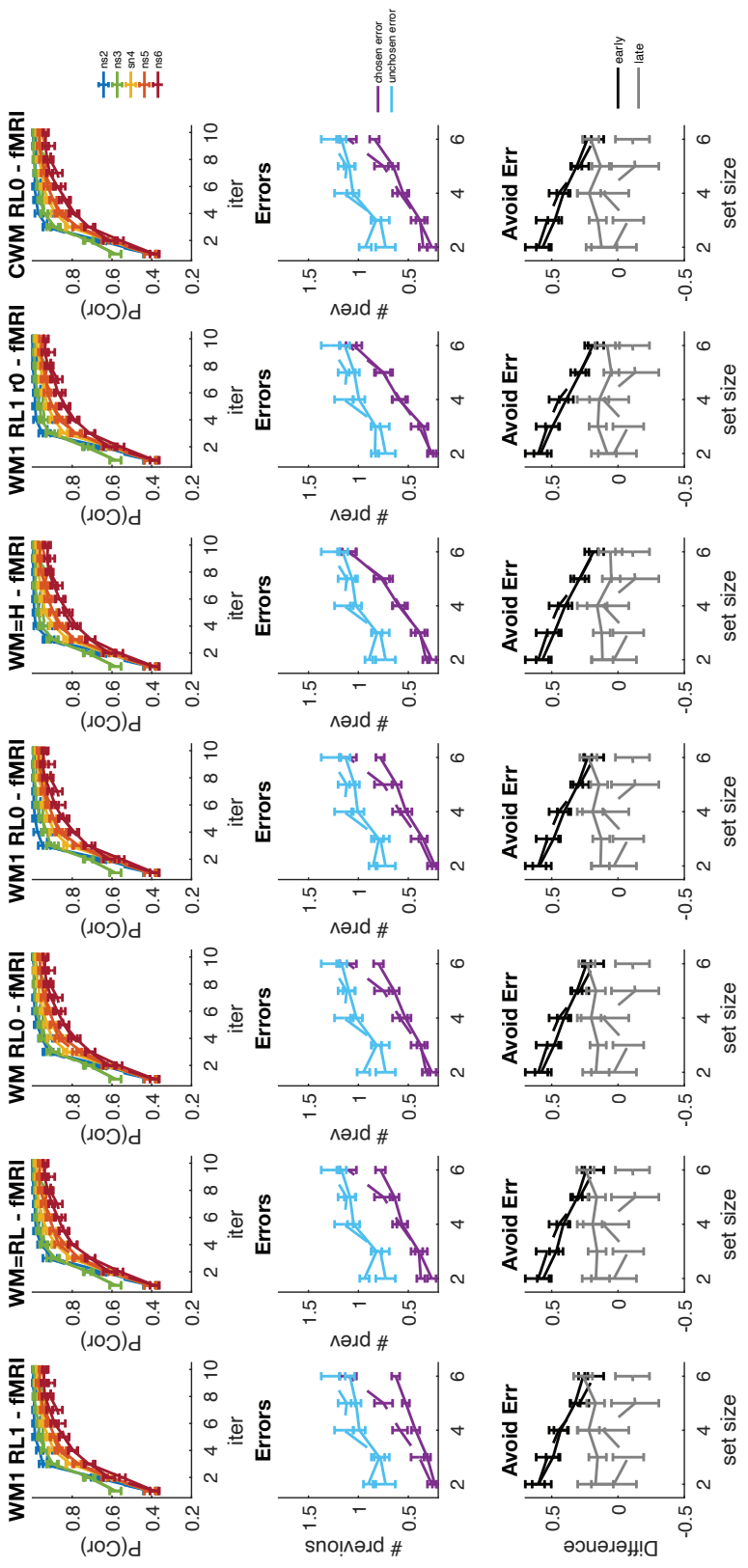
Model validation figures

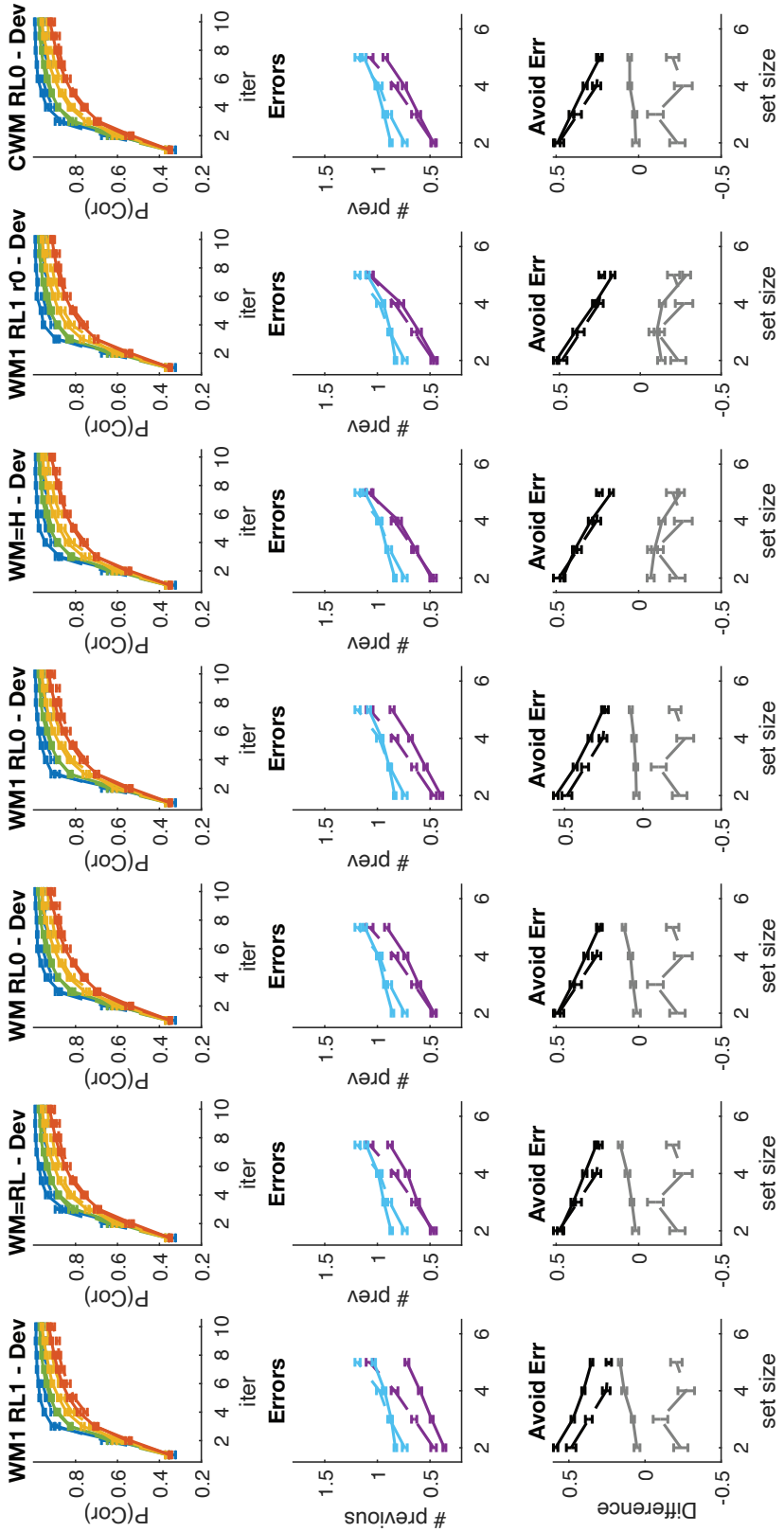
Following figures include model validation figures for all 6 data sets with the same models as in 2.


WM1 RL1 - CF12
WM=RL - CF12
WM RL0 - CF12
WM1 RL0 - CF12
WM=H - CF12
WM1 RL1 r0 - CF12
CWM RL0 - CF12
P(Cor)
iter
ns2
ns3
sn4
ns5
ns6
previous
Errors
chosen error
unchosen error
Avoid Error
Avoid Err
Difference
set size
early
late


WM1 RL1 - SZ
WM=RL - SZ
WM RL0 - SZ
WM1 RL0 - SZ
WM=H - SZ
WM1 RL1 r0 - SZ
CWM RL0 - SZ
P(Cor)
iter
ns2
ns3
ns4
ns5
ns6
Errors
prev
previous
chosen error
unchosen error
Avoid Err
Difference
set size
early
late


WM1 RL1 - EEG
WM=RL - EEG
WM RL0 - EEG
WM1 RL0 - EEG
WM=H - EEG
WM1 RL1 r0 - EEG
CWM RL0 - EEG
P(Cor)
iter
ns2
ns3
ns4
ns5
ns6
previous
prev
Errors
chosen error
unchosen error
Difference
Avoid Err
set size
early
late


WM1 RL1 - fMRI
WM=RL - fMRI
WM RL0 - fMRI
WM1 RL0 - fMRI
WM=H - fMRI
WM1 RL1 r0 - fMRI
CWM RL0 - fMRI
ns2
ns3
sn4
ns5
ns6
P(Cor)
iter
Errors
prev
previous
chosen error
unchosen error
Avoid Err
Difference
set size
early
late


WM1 RL1 - Dev
WM=RL - Dev
WM RL0 - Dev
WM1 RL0 - Dev
WM=H - Dev
WM1 RL1 r0 - Dev
CWM RL0 - Dev
P(Cor)
iter
Errors
previous
prev
Avoid Err
Difference
set size

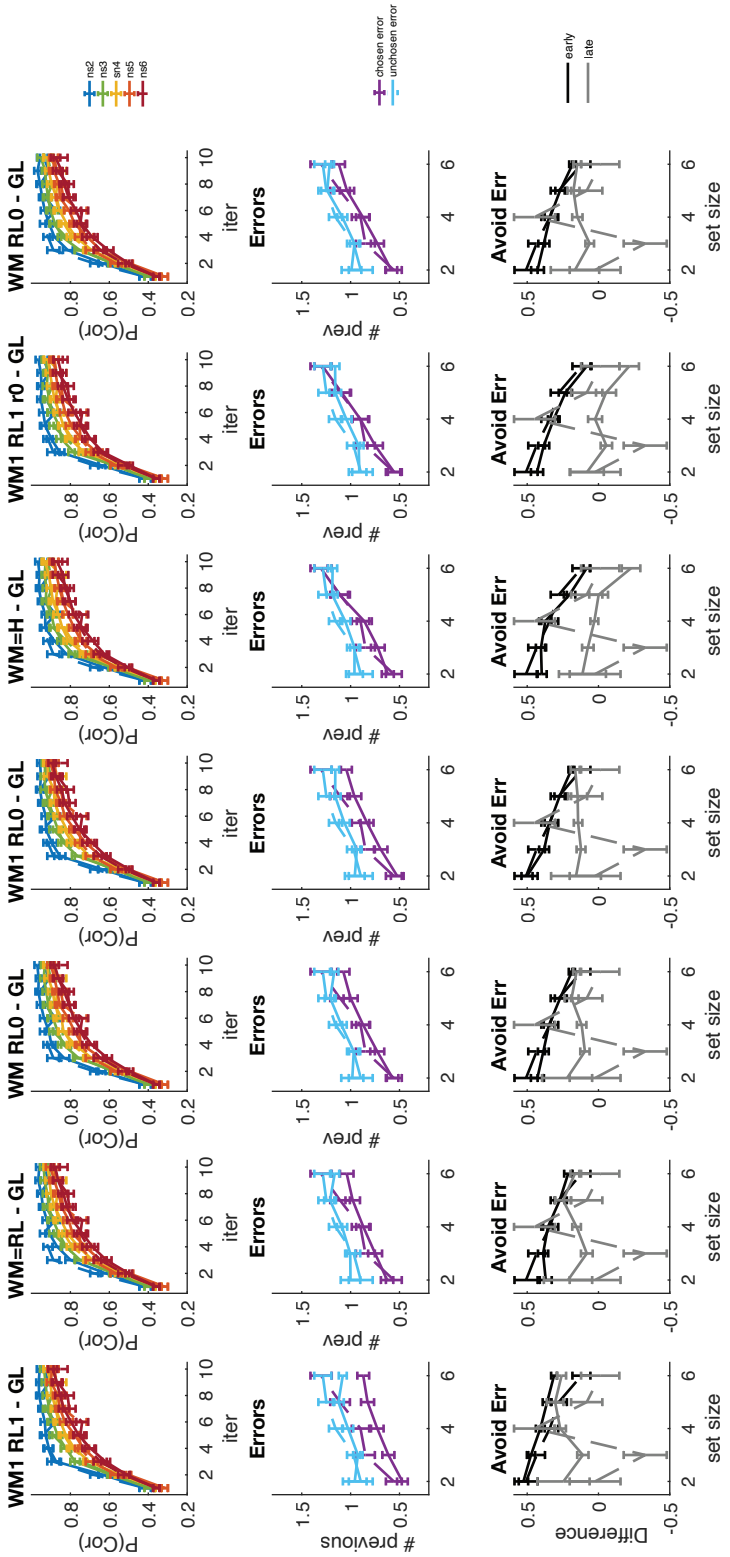

WM RL0 - GL
WM1 RL1 r0 - GL
WM=H - GL
WM1 RL0 - GL
WM=RL - GL
WM1 RL1 - GL
P(Cor)
iter
Errors
prev
previous
chosen error
unchosen error
Avoid Err
Difference
set size
early
late
ns2
ns3
sn4
ns5
ns6

Figure 9

Validation figures for all 6 RLWM data sets. See legend for fig. 2. Model WM1RL1r0 captured behavior as well as WM=H, but was quantitativley penalized for the extra parameter (see fig. 2A). Figure 10 shows that the r_0 parameter was typically fitted at a high value that would lead to positive reinforcement of negative outcomes, making this model similar to WMH.

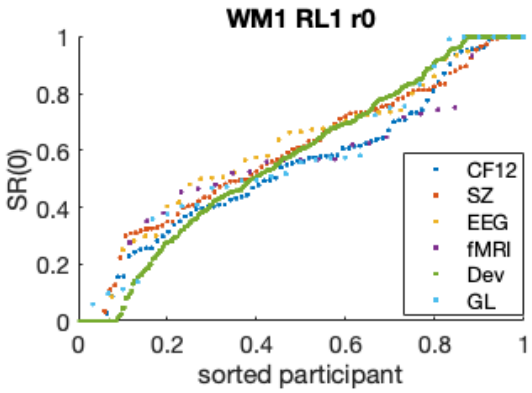


Figure 10

WM1 RL1 r0 model's fit r_0 parameter distribution across data sets shows that a majority of participants are fit with a value higher than the initialization of $Q_0 = 1/3$, such that this model falls into the H family of the spectrum (making an error makes the error more likely to be repeated).

RLWMP validation

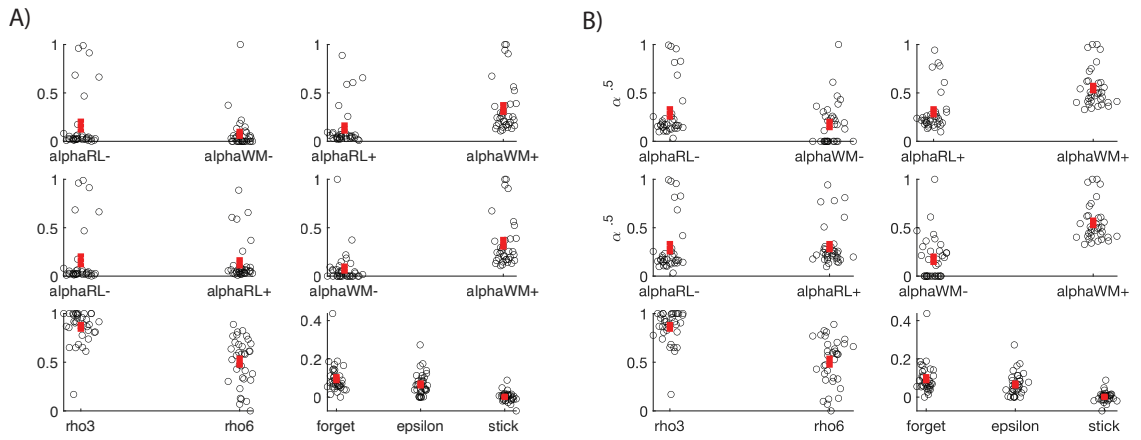


Figure 11

Fit parameters for the winning model in RLWM-P data set. A) Absolute parameters. B) for better visualization of the distribution of learning rates, we plot transformed parameters with $\sqrt{\alpha}$*. The WM positive learning rate is significantly higher than the RL one, highlighting the faster learning dynamic, as expected for a WM-based process. The weight to the WM process in ns3 is close to 1, as expected for WM use under low load. Red error bars indicate mean, SEM.*

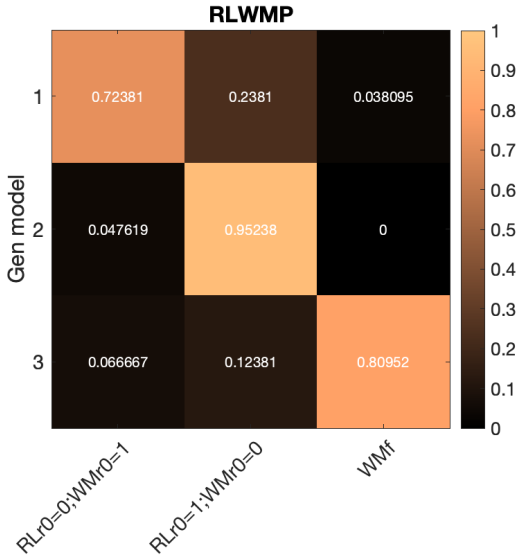


Figure 12

We use the same procedure as above to confirm model identifiability between the best model and competing ones.

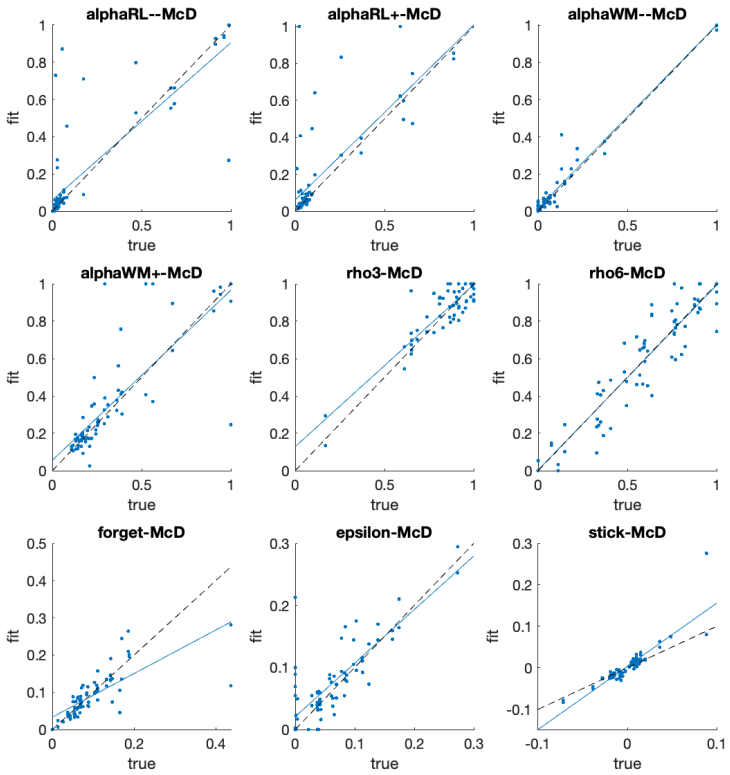


Figure 13

Parameters for the winning model in RLWM-P data set are identifiable, as shown by a generate and recover procedure.