Extending the RLWM Model to Ecological Decision-Making: A Bayesian Approach

Manavjeet Singh- 220616[a], Sachidanand- 220929[b], Sagar Arora- 220933[c], Sneh Sinha- 221067[d] and Yash Chauhan- 221217[e]

[a,b,c,d,e] *Department of Biological Sciences and Bioengineering, IIT Kanpur*

Arjun Ramakrishnan - Assistant Professor Department of Biological Sciences and Bioengineering, IIT Kanpur
Mentors: *Kshitij*

Contents

1. Abstract

This study explores the interplay between working memory (WM) and reinforcement learning (RL) in human decision-making through computational modeling. We first replicated the RLWM framework proposed by Anne Collin [4] using `Stan`, confirming that WM and RL systems operate in parallel during learning tasks. Our implementation successfully reproduced the original findings, with parameter estimates closely aligning with those reported in prior work: RL2 model ($\alpha = 0.35$, $\beta = 6.93$), WM model ($C = 4.23$, $\epsilon = 0.089$), and the RLWM hybrid model ($C = 3.73$, $\epsilon = 0.16$, $w_0 = 0.54$).

We then adapted these models to a dynamic patch foraging task, in which rewards deplete with repeated exploitation and replenish over time, creating a more ecologically valid setting. Model comparison showed that the WM model yielded a 38% improvement in likelihood over the baseline RL2 model, while the hybrid RLWM model achieved a 15% improvement. AIC-based model selection further confirmed that both models provided significantly better fits than the RL2 model, even after accounting for model complexity.

Behavioral analyses revealed a systematic transition from WM-dominated behavior in early trials ($w \approx 0.8$) to increased reliance on RL in later trials ($w \approx 0.5$), reflecting cognitive adaptation to shifting task demands. These dynamics suggest that WM facilitates rapid learning during early exploration, while RL supports stable performance through long-term integration of reward history.

Finally, we outline ongoing work on a hierarchical Bayesian extension of the RLWM framework. This approach aims to better capture individual differences in learning strategies and enable population-level inference, offering potential insights into variability in adaptive behavior across individuals and contexts.

2. Introduction

How do we navigate the complexities of life? Imagine getting lost in a new city with a map. Initially, you rely on this map—your working memory—carefully piecing together the route. Over time, however, you no longer need the map; your journey becomes intuitive, guided by the rewards of past experiences. This shift is reinforcement learning (RL) at work. Our research explores how working memory (WM) and RL work together to shape human learning, bridging scientific discovery and everyday experience.

For centuries, scientists have explored the mysteries of learning. Early theories like Pavlov's stimulus-response model gave way to more complex ideas. Traditional RL models, focused on rewards and errors, often miss the role of working memory—the brain's temporary notepad. The RLWM model, introduced by Anne Collins, suggests that WM drives initial learning, and once capacity is reached, RL takes over, automating decision-making.

Inspired by this model, we replicated Collins' controlled task and extended it to a dynamic foraging task. Using STAN, a powerful statistical tool, we confirmed that WM and RL collaborate, with WM guiding early decisions and RL adapting over time. To test its real-world relevance, we added complexity by introducing a foraging task, where rewards change over time, reflecting dynamic environments like resource foraging or adapting to new challenges.

We also discovered individual differences in learning. Some participants relied on WM, while others adapted more through RL. To understand this variability, we used hierarchical Bayesian methods to capture both group-level trends and individual differences, uncovering diverse learning strategies.

Our research not only extends the RLWM model but also highligworhts the significance of personal learning profiles, with implications for education and AI. By tailoring approaches to individual cognitive strengths, we may unlock more effective learning environments for both humans and machines.

2.1. Background and Motivation

2.1.1. The Collins Task

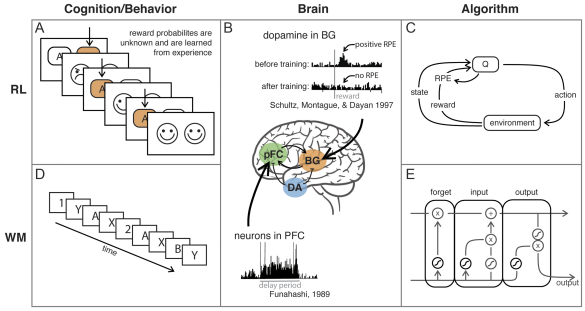

Figure 1. Collin's schematic from his WM and RL study [3]

The *Collins Task* is a behavioral learning game designed to investigate how humans switch between **working memory (WM)** and **reinforcement learning (RL)** depending on cognitive demands. First introduced by *Collins and Frank (2012)*, and later discussed in detail by *Yoo & Collins (2022)*, this task provides a structured yet dynamic environment that exposes the contributions of both learning systems.

Game Setup and Flow: Each participant plays through multiple blocks of trials. In each block, a small set of novel stimuli (e.g., abstract shapes) is introduced. One stimulus is shown per trial, and the participant must select the correct action (e.g., keypress 1, 2, 3, F, etc.) from several options. The participant receives immediate feedback after each choice:

- Correct responses get a visual or auditory reward.
- Incorrect responses trigger an error tone or cross.

Participants are not told the correct mappings beforehand and must learn them through trial and error. Importantly, feedback is deterministic — correct responses always yield a reward — ensuring that learning difficulty comes from memory demands rather than outcome uncertainty.

Experimental Manipulation—Set Size: The core experimental variable is **set size**: the number of unique stimuli per block, usually ranging from 2 to 6.

- *Small set sizes (2–3)*: Well within WM capacity. Participants often learn the correct responses quickly and apply them accurately using WM.
- *Large set sizes (5–6)*: Exceed WM limits. Participants begin to rely on slower, feedback-based RL processes.

This manipulation systematically modulates cognitive load, allowing researchers to observe how participants shift from WM to RL-based strategies as memory demands increase.

2.1.2. The Collins RLWM Model

To explain the behavioral patterns observed in the task, Collins and Frank (2012) proposed a hybrid model called RLWM. This model assumes that both WM and RL systems operate in parallel and jointly contribute to choice behavior. The WM module allows rapid, one-shot learning of stimulus–action pairs but is limited by capacity and subject to forgetting. In contrast, the RL module updates action values gradually using reward prediction errors over repeated trials.

A central feature of the model is the mixture weight parameter (ω), which determines the influence of each system on the participant's choices. When the task is simple or memory load is low, the mixture weight favors WM. As the load increases or memory decays, the model dynamically shifts control toward RL. This flexible arbitration between systems enables the model to capture both fast initial learning and slower adaptation in more complex settings.

2.1.3. Discussion on Collin's

Collins' model introduced a major shift in our understanding of how humans learn, particularly by emphasizing the role of the prefrontal cortex (PFC) in memory-based decision-making. Prior models typically focused only on RL and reward systems like the basal ganglia, neglecting the significant contribution of cognitive resources such as WM.

The RLWM model provides compelling evidence that differences in **working memory** capacity can meaningfully affect how individuals learn. Moreover, it challenges the long-held assumption that WM and RL compete for control. Instead, Collins shows that these systems work in collaboration—WM is engaged first, offering fast and efficient learning, and as its capacity is exceeded or its reliability diminishes, RL gradually takes over to stabilize behavior. This fluid interaction between systems aligns well with both behavioral data and neural evidence and helps explain the flexibility and adaptability seen in human learning.

2.1.4. Why STAN?

The Collins Task explores the interplay between working memory and reinforcement learning in decision-making. **STAN**, with its **Bayesian inference**, is ideal for modeling this task due to its ability to handle complexity and uncertainty. It captures **individual differences** in cognitive parameters, such as learning rates, and excels in dynamic, adaptive scenarios—making it a strong fit for **studying human learning variability** in this context.

3. Model Replication and Validation in STAN

To ensure the robustness and credibility of our findings, we began by replicating the original RLWM model using STAN. This phase focused on faithfully reproducing the structure, assumptions, and parameterization of the original model as described in prior literature. By doing so, we aimed to validate our implementation against established benchmarks and confirm that the model could recover known behavioral patterns. The replication also served as a foundation for subsequent modifications and extensions, ensuring that any observed effects in later tasks could be attributed to genuine model dynamics rather than implementation inconsistencies.

We chose KDE over the fmin approach used by Collins because KDE captures the full posterior distribution—its uncertainty, skew, and any multimodality—rather than collapsing inference to a single point. This smoothing of MCMC samples reveals credible intervals and complex shape features, aligning with Bayesian principles and providing a more robust, informative summary of parameter estimates.

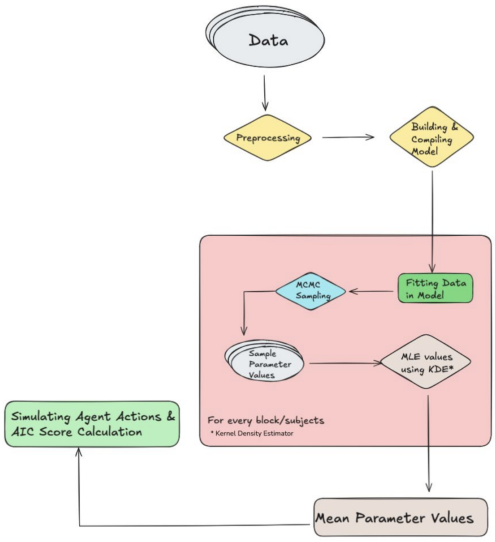

Figure 2. Workflow of all the models

3.1. Fitting and Comparison

3.1.1. Pure Reinforcement Learning Model (RL2)

As a baseline, we implemented the standard **reinforcement learning (RL)** model with two core parameters using STAN. In this model, the expected reward value for a given stimulus–action pair, $Q(s, a)$, is updated incrementally based on the difference between the received reward and the predicted reward—a quantity known as the **prediction error**. This update is governed by a **learning rate parameter** α_{RL}, which determines how strongly new information influences future expectations. Formally, the Q-value update follows:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha_{RL} \cdot (r_t - Q(s_t, a_t)) \quad (1)$$

Action selection is modeled using a **softmax choice rule**, where the probability of choosing an action depends on its relative Q-value compared to other options. This behavior is regulated by an **inverse**

temperature parameter β_{RL}, which controls the stochasticity of the choices:

$$p(a \mid s) = \frac{\exp(\beta_{RL} \cdot Q(s,a))}{\sum_i \exp(\beta_{RL} \cdot Q(s,a_i))} \tag{2}$$

In this two-parameter RL model, learning is stimulus-specific and does not take into account contextual variables such as the number of stimuli (set size), inter-trial delays, or interference from other items. As such, it cannot account for working memory effects and serves primarily as a benchmark to evaluate more complex models like RLWM. It also aligns with traditional corticostriatal models of learning.

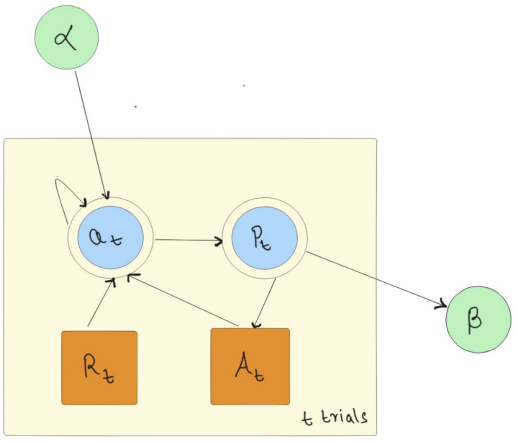

Figure 3. Schematic of RL2

3.1.2. RL6

To model potential variations in learning rate across different memory loads, the RL6 model extends the standard RL formulation by allowing the learning rate α_{RL} to vary based on set size. This model includes six parameters—five different learning rates and a common inverse temperature—thereby capturing whether participants adjust their learning strategies depending on the complexity or number of items to learn.

Variants of this model include configurations with shared learning rates or softmax temperatures, enabling flexible assessment of how parameter granularity impacts fit.

3.1.3. Forgetful Reinforcement Learning Model (RLF)

The RLF model incorporates a forgetting mechanism on top of traditional Q-learning. In addition to the standard update:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha_{RL} \cdot (r_t - Q(s_t,a_t)) \tag{3}$$

a decay term is applied across all Q-values to simulate memory degradation over time:

$$Q(s,a) \leftarrow Q(s,a) + \epsilon \cdot (Q_0 - Q(s,a)) \tag{4}$$

where $Q_0 = \frac{1}{n_A}$ is the initial value reflecting a uniform prior over actions, and ϵ controls the forgetting rate. Higher set sizes introduce longer average delays between stimulus repetitions, leading to greater cumulative forgetting and reduced performance.

3.1.4. Reinforcement Learning + Working Memory Hybrid Mode

This model blends RL and WM contributions to action selection. The WM system retains stimulus–action pairings perfectly at first, updating via a decay rule similar to the RLF model:

$$Q_{WM}(s,a) \leftarrow Q_{WM}(s,a) + \epsilon \cdot \left(\frac{1}{n_A} - Q_{WM}(s,a)\right) \tag{5}$$

Decisions are generated as a mixture of both systems:

$$p(a) = (1 - w(t)) \cdot p_{RL}(a) + w(t) \cdot p_{WM}(a) \tag{6}$$

The mixing weight $w(t)$ is capacity-dependent and changes over time. The probability that an item is retained in WM is $\min\left(1, \frac{C}{n_S}\right)$, and affects the likelihood that WM will predict the reward correctly. This leads to the conditional likelihoods:

$$p_{WMC}(r_t = 1 \mid s_t, a_t) = \min\left(1, \frac{C}{n_S}\right) Q_{WM}(s_t,a_t) + \left(1 - \min\left(1, \frac{C}{n_S}\right)\right) \cdot \frac{1}{n_A} \tag{7}$$

$$p_{WMC}(r_t = 0 \mid s_t, a_t) = \min\left(1, \frac{C}{n_S}\right)(1 - Q_{WM}(s_t,a_t)) + \left(1 - \min\left(1, \frac{C}{n_S}\right)\right) \cdot \frac{1}{n_A} \tag{8}$$

The RL system likelihood is:

$$p_{RL}(r_t = 1 \mid s_t, a_t) = Q(s_t,a_t) \tag{9}$$

$$p_{RL}(r_t = 0 \mid s_t, a_t) = 1 - Q(s_t,a_t) \tag{10}$$

The Bayesian mixture weight is then updated over time based on likelihoods:

$$w_{n_S}(t+1,s) = \frac{p_{WMC}(r_t \mid s_t,a_t) \cdot w_{n_S}(t,s)}{p_{WMC}(r_t \mid s_t,a_t) \cdot w_{n_S}(t,s) + p_{RL}(r_t \mid s_t,a_t) \cdot (1 - w_{n_S}(t,s))} \tag{11}$$

The initialization of the mixing weight depends on capacity:

$$w_{n_S}(t=0,s) = w_0 \cdot \min\left(1, \frac{C}{n_S}\right) \tag{12}$$

This hybrid model can account for both set-size and delay effects, modeling fast WM-driven learning early in a block and gradual RL-driven stabilization later.

3.1.5. Pure Working Memory Model

As a control, a pure WM model was used that excludes any RL contribution. Here, when WM fails (i.e., set size exceeds capacity), action selection becomes random. The mixing weight is fixed, and the model captures set-size and delay effects solely via memory degradation and capacity limits.

3.2. Result Analysis and Comparison

Table 1. Collins' Study Results

Model	β	α	C	ε	w_0	β_{WM}
RL2	7.1	0.38	—	—	—	—
WM	—	—	5.4	0.1	0.95	12
RLF	24.1	0.29	—	0.07	—	—
RL + WM	26.6	0.16	3.7	0.23	0.81	45

Table 2. Our Results

Model	β_{w}	α_{RL}	C	ε	w_0	β_{WM}
RL2	6.93	0.35	—	—	—	—
WM	—	—	4.23	0.089	0.91	11.93
RLF	20.93	0.24	—	0.07	—	—
RL + WM	23	0.31	3.73	0.16	0.54	47.95

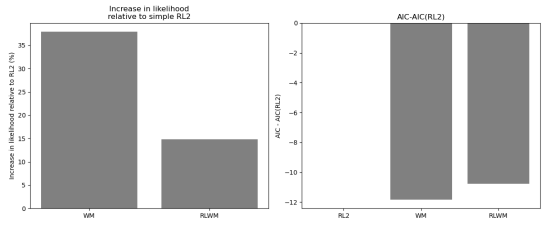

Figure 4. AIC and Likelihood Comparison

Figure 4: Model comparison based on log-likelihood gain and AIC. The left panel shows the percentage increase in likelihood relative to the baseline RL2 model. The WM model shows the highest increase (38%), followed by the RLWM model (15%). The right panel presents AIC differences relative to RL2. Both WM and RLWM yield lower AIC values than RL2, indicating better model fit despite increased complexity, with WM achieving the greatest AIC reduction.

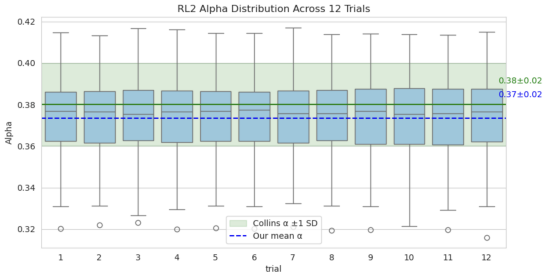

Figure 5. Alpha (RL2) for 102 subjects across 12 trials

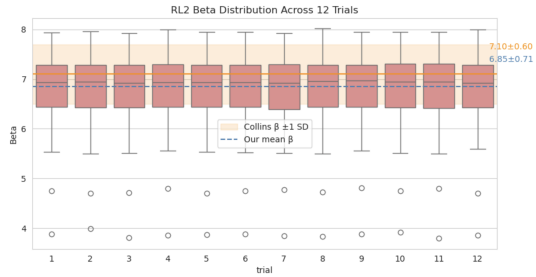

Figure 6. Beta (RL2) for 102 subjects across 12 trials

Figure 5 and 6 contain RL2 Parameter Distributions Across Trials. The boxplots show the distribution of the learning rate (α) and inverse temperature (β) parameters for all 102 subjects across 12 trials in the RL2 model. Each box represents the interquartile range of the parameter estimates, with whiskers extending to 1.5×IQR and circles denoting outliers. The green shaded region in the top plot indicates the ±1 standard deviation (SD) range of Collins et al.'s α estimate (mean = 0.38 ± 0.02), while the blue dashed line denotes our average α across trials (mean = 0.37 ± 0.02). Similarly, the orange shaded area in the bottom plot represents Collins et al.'s β estimate (mean = 7.10 ± 0.60), and the blue dashed line shows our mean β (6.85 ± 0.71). The consistency across trials suggests parameter stability and close alignment with previous findings.

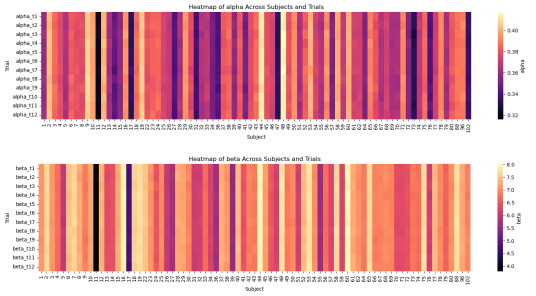

Figure 7. Rl2 parameter heatmap

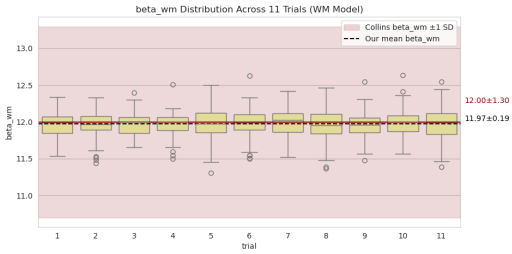

Figure 8. Beta (WM) for 102 subjects across 11 trials

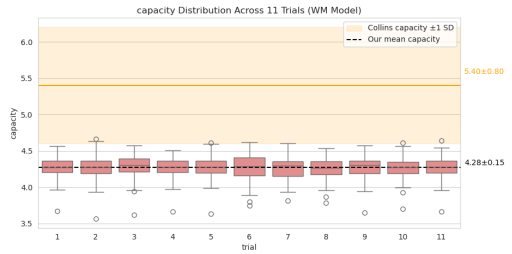

Figure 9. Capacity (WM) for 102 subjects across 11 trials

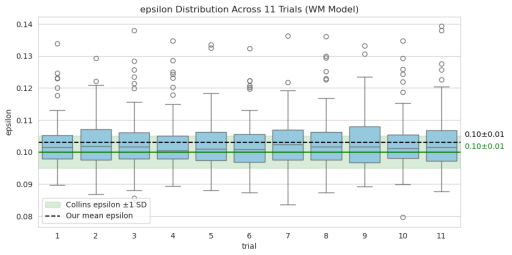

Figure 10. Epsilon (WM) for 102 subjects across 11 trials

Figure 8, 9, and 10: Distribution of WM Model Parameters (β_{WM}, capacity, and ϵ) Across 11 Trials. Boxplots illustrate the variability of parameter estimates across 102 subjects over 11 trials for the Working Memory (WM) model. The top panel shows the distribution of the β_{WM} parameter, which remained highly consistent across trials (mean = 11.97 ± 0.19), closely matching the reference value reported by Collins (12.00 ± 1.30). The middle panel represents the capacity parameter, which also exhibited trial-wise stability (mean = 4.28 ± 0.15), though slightly lower than Collins' reference (5.40 ± 0.80). The bottom panel shows the ϵ (epsilon) parameter, with a mean of 0.10 ± 0.01, aligning exactly with Collins' reported value (0.10 ± 0.01). In all plots, shaded regions represent the ±1 standard deviation range

reported by Collins, and the dashed black lines indicate our computed mean parameter values across trials.

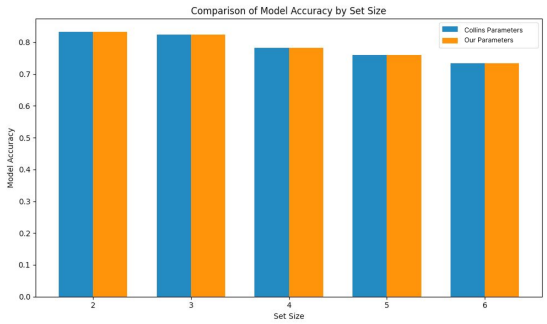

Figure 11. Comparison of model accuracy across set sizes in the WM model

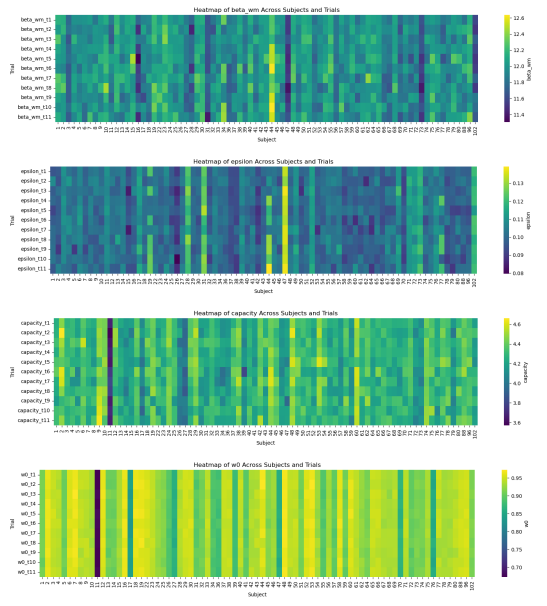

Figure 12. WM parameter heatmap

Figure 11 displays Bar plot compares model accuracy between Collins' parameter estimates and our derived parameters across set sizes 2 to 6. The close alignment across all set sizes indicates that our parameter estimates generalize well and replicate model behavior effectively.

Simulated and empirical learning curves across working memory (WM) and reinforcement learning (RL2) models, compared with observed participant data across different set sizes are shown below.

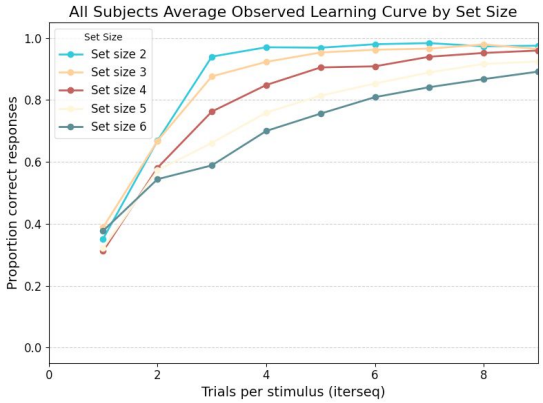

Figure 13. learning curve

Figure 13: Simulated WM learning curves using parameter estimates from Collins et al. (2012), showing faster and higher learning performance for smaller set sizes due to greater WM contribution.

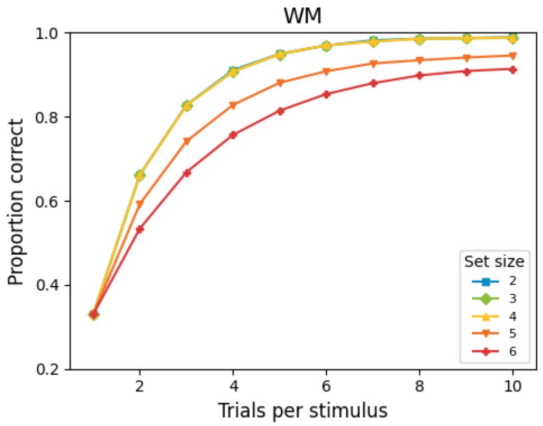

Figure 14. WM learning curve

Figure 14: Simulated RL2 model learning curves using best-fitting parameters from the RLWM model, indicating near-ceiling performance for all set sizes with minimal differentiation, reflecting the dominance of reinforcement learning processes.

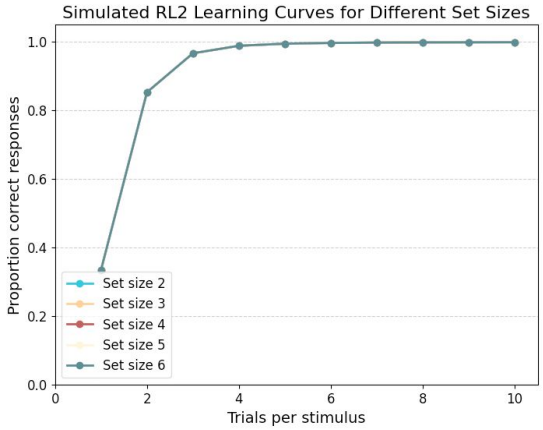

Figure 15. Comparison of model accuracy across set sizes in the WM model

Figure 15: Empirical learning curves averaged across all subjects, illustrating graded improvement in accuracy across trials and set sizes. The divergence in learning trajectories supports the contribution of both WM and RL systems, with smaller set sizes benefiting more from WM-based strategies.

3.3. Discussion

The comparative analysis between our study and Collins et al. (2012) demonstrates strong consistency in both parameter estimates and model behavior. Across all models—RL2, WM, RLF, and RL+WM—our recovered parameters closely align with those originally reported, validating the robustness of our fitting procedure. Notably, while Collins employed a deterministic optimization approach, our framework adopts a Bayesian inference method, enabling uncertainty estimation and principled regularization.

Despite this methodological difference, we achieved highly similar parameter values and learning dynamics, affirming the validity of our Bayesian approach. Model comparison metrics, including likelihood gain and Akaike Information Criterion (AIC), consistently favor the WM and RLWM models over RL2, despite their greater complexity. This highlights the critical role of working memory in explaining human learning behavior.

Moreover, our trial-wise parameter stability and close replication of Collins' findings across multiple metrics (e.g., α, β, capacity, ϵ) reinforce the reliability of our estimation pipeline. The WM model, in particular, shows the highest log-likelihood improvement and AIC reduction, confirming its explanatory power. Furthermore, learning curve simulations and model accuracy across set sizes illustrate that our parameters generalize effectively to observed behavior, underscoring the strength of our modeling approach in capturing human cognitive processes in reinforcement learning tasks.

4. Extending RLWM to Dynamic Foraging Tasks

Building on the foundational work of Collins' RLWM model, we aim to explore whether similar principles of working memory and reinforcement learning collaboration apply in more **ecologically realistic** and d**ynamic environments**. Specifically, we extend our STAN-based RLWM implementation to model human behavior in a **foraging task**, where decision-making involves continuous trade-offs between exploitation and exploration. This shift allows us to test the generalizability of the RLWM framework beyond static, trial-based learning settings.

Our goal is to assess how well the RLWM and its subcomponents (RL-only and WM-only models) capture behavior in a task that more closely mimics real-world decision-making. Unlike the fixed and structured nature of the Collins Task, the foraging task introduces uncertainty, temporal dynamics, and ongoing reward depletion—all of which present a new challenge for cognitive modeling.

We use the **same core model structures implemented in STAN** for RL, WM, and RLWM, modifying only the task-specific elements such as the state transitions and reward schedules. In doing so, we examine whether the interaction between WM and RL continues to play a meaningful role, and how this interaction may manifest differently in environments that demand long-term planning and flexibility.

4.1. Task Overview

The foraging task simulates a naturalistic environment where agents (participants) must **harvest rewards from multiple resource patches**. Each patch offers a variable reward that **depletes with use** and **replenishes when left alone**. Participants must decide when to stay in a patch to exploit its current value and when to leave it to search for a better opportunity. Each decision involves **travel time costs** and **harvesting delays**, making time management critical.

Participants played multiple rounds of this task. In each round, they navigated between patches with different replenishment rates. Some patches recovered quickly when left alone, while others replenished slowly or not at all. The key challenge was to determine which patches were profitable over time and to maximize overall reward within the fixed time window of the task.

This setup mimics real-life foraging behavior seen in animals and humans—where individuals must continuously balance **short-term exploitation of known options** against **exploration of unknown or temporarily unprofitable alternatives**.

4.2. Key Differences from Collins' Task

The **Collins Task** and **our Foraging Task** differ primarily in their objectives, complexity, and cognitive demands.

4.2.1. Objective

- *Collins' Task* focuses on how participants learn stimulus-action associations, exploring the interaction between working memory (WM) and reinforcement learning (RL).
- *Foraging Task* simulates real-life decision-making, where participants must balance *exploration* (finding better rewards) and *exploitation* (maximizing current rewards from resource patches).

4.2.2. Task Structure

- *Collins Task* involves learning fixed stimulus-action pairs with a manipulation of *set size*, examining how memory affects learning.
- *Foraging Task* requires navigating between patches with variable replenishment rates and deciding when to stay or leave, with time management and memory tracking rewards.

4.2.3. Cognitive Load

- *Collins' Task* manipulates **set size** to test memory limits and decision-making under different cognitive loads.
- *Foraging Task* adds complexity by introducing variable rewards, delays, and time costs, requiring participants to manage resources over time.

4.2.4. Feedback

- *Collins Task* provides deterministic feedback (correct or error tone).
- *Foraging Task* offers dynamic feedback based on rewards from patches, requiring strategic time management and adaptation.

In essence, the *Foraging Task* is more complex, involving real-time decision-making and resource management, whereas the *Collins Task* is simpler, focusing on learning and memory within a fixed set of stimuli.

4.3. Model Adaptation

The original RLWM models developed by Collins et al. were designed for associative learning tasks involving discrete stimulus-action pairs with fixed reward contingencies. In contrast, the foraging task implemented here involves choices between patches (trees) whose rewards change dynamically due to depletion and replenishment processes. Consequently, we adapted the RL, WM, and hybrid RLWM models to accommodate these differences while retaining their core computational principles.

4.3.1. Adaptation of the RL2 Model

In the adapted reinforcement learning (RL2) model, the agent selects among a fixed number of patches, each yielding a numeric reward (fruit count) that varies over time. The agent maintains a Q-value for each action, updated according to a Rescorla-Wagner rule. To ensure numerical stability and comparability across trials, the reward is scaled by a fixed constant (typically 200), leading to the update rule:

$$Q(a_t) \leftarrow Q(a_t) + \alpha \left(\frac{r_t}{\text{reward\_scale}} - Q(a_t) \right), \tag{13}$$

where a_t is the chosen action, r_t is the received reward, $\alpha \in [0, 1]$ is the learning rate, and `reward_scale` is a fixed normalization factor. The action selection policy is governed by a softmax function:

$$P(a_t) = \frac{\exp(\beta Q(a_t))}{\sum_{a'} \exp(\beta Q(a'))}, \tag{14}$$

where β is the inverse temperature parameter that controls the exploitation-exploration trade-off.

4.3.2. Adaptation of the Working Memory (WM) Model

The WM model assumes that the agent stores recent reward outcomes with high fidelity but with limited capacity and temporal stability. In the foraging context, this model tracks immediate reward values for each patch and assumes decay in memory strength when a patch is not selected. Let $Q_{\text{wm}}[a]$ denote the WM value for action a. The update and decay rules are:

$$Q_{\text{wm}}[a] \leftarrow Q_{\text{wm}}[a] + \epsilon\left(\frac{1}{A} - Q_{\text{wm}}[a]\right), \tag{15}$$

$$Q_{\text{wm}}[a_t] \leftarrow \frac{r_t}{\text{reward\_scale}}, \tag{16}$$

where A is the total number of actions, and ϵ is a decay parameter. The probability of a patch being stored in WM is inversely related to the number of available patches:

$$p_{\text{in\_wm}} = \min\left(1.0, \frac{\text{capacity}}{A}\right), \tag{17}$$

and the effective WM weight is:

$$w_{\text{eff}} = w_0 \cdot p_{\text{in\_wm}}, \tag{18}$$

where w_0 is a base WM weight parameter. The final action policy is a mixture of WM-driven and random (uniform) policies:

$$P(a) = w_{\text{eff}} \cdot P_{\text{wm}}(a) + (1 - w_{\text{eff}}) \cdot \frac{1}{A}. \tag{19}$$

4.3.3. Adaptation of the RLWM Hybrid Model

The hybrid RLWM model combines the gradual learning of the RL2 model with the fast-updating, capacity-limited memory of the WM model. It maintains two value systems: one updated via reinforcement learning and one updated via working memory. Action selection is a weighted combination of the two policies:

$$\pi_{\text{rl}} = \text{softmax}(\beta_{\text{rl}} \cdot Q + \text{stick} \cdot \text{prev\_choice\_vec}), \tag{20}$$

$$\pi_{\text{wm}} = \text{softmax}(\beta_{\text{wm}} \cdot \text{WM} + \text{stick} \cdot \text{prev\_choice\_vec}), \tag{21}$$

$$\pi = (1 - \epsilon) \cdot (w \cdot \pi_{\text{wm}} + (1 - w) \cdot \pi_{\text{rl}}) + \epsilon/A, \tag{22}$$

where ϵ is the lapse rate capturing random responding, w is the mixture weight between WM and RL, and the `stick` parameter captures choice perseveration. The previous choice vector (prev_choice_vec) encodes which patch was selected on the last trial.

Value updates proceed as:

$$Q[a_t] \leftarrow Q[a_t] + \alpha_{\text{rl}} \cdot \left(\frac{r_t}{200.0} - Q[a_t]\right), \tag{23}$$

$$\text{WM} \leftarrow \text{WM} \cdot (1 - \text{forget}), \tag{24}$$

$$\text{WM}[a_t] \leftarrow \frac{r_t}{200.0}, \tag{25}$$

$$\text{prev\_choice\_vec}[a_t] = 1.0. \tag{26}$$

4.3.4. Summary of Adaptations

These adaptations allow the original RLWM models to function in a foraging environment characterized by dynamic reward structures and limited time constraints. The principal modifications include:

- Replacing stimulus-response mappings with direct action (patch) selection.
- Normalizing fruit values via a fixed reward scale to standardize learning dynamics.
- Including temporal decay in both WM and RL components to reflect changing patch values.
- Adding stickiness and lapse parameters to account for sequential dependencies and randomness in choice behavior.

These changes preserve the conceptual distinction between working memory and reinforcement learning systems while making them compatible with the ecological demands of patch foraging.

5. Implementation and Results Overview

5.1. Model Fitting

Following the adaptation of the RL, WM, and RLWM models for the dynamic foraging task, we proceeded with fitting these models to the behavioral data collected from participants.

The adapted models—RL2, Pure WM, and RLWM—were fitted individually to each participant's trial-by-trial data from the foraging task. The data included the sequence of patch choices, the rewards received, and the timing information relevant for decay processes.

Using STAN's Hamiltonian Monte Carlo (HMC) algorithms, we obtained posterior distributions for the key parameters of each model:

- **RL2 Model:**
  - Learning rate (α)
  - Inverse temperature (β)
- **WM Model:**
  - Effective working memory weight (w_{eff}), derived from w_0 and capacity
  - Decay/forgetting parameter (ϵ)
  - Inverse temperature (β_{WM})
- **RLWM Hybrid Model:**
  - RL learning rate (α_{RL})
  - RL inverse temperature (β_{RL})
  - WM inverse temperature (β_{WM})
  - Forgetting rate (`forget`)
  - WM mixture weight (w)
  - Stickiness (`stick`)
  - Lapse rate (ϵ)

The fitting process aimed to find the parameter values that maximized the likelihood of the observed sequence of choices for each participant under each model. The quality of the fits was visually inspected using *Subject-Model Action Plots* (Figures 16–18), which compare the participant's actual choices with the model's predicted choice probabilities on a trial-by-trial basis. These plots help assess how well each model captures the specific decision sequences and switching behavior characteristic of foraging.

Additionally, for the RLWM model, the dynamics of the inferred working memory contribution (mixture weight w) were examined (Figure 19), to understand the interplay between the two systems during the task.

5.2. Result Analysis

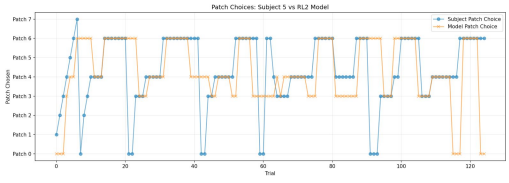

Figure 16. Subject-Model Action Plot for RL2 Foraging Model

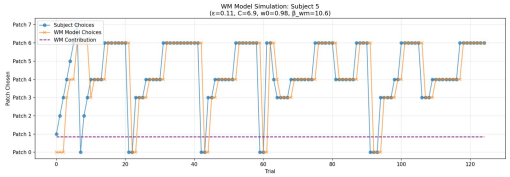

Figure 17. Subject-Model Action Plot for WM Foraging Model

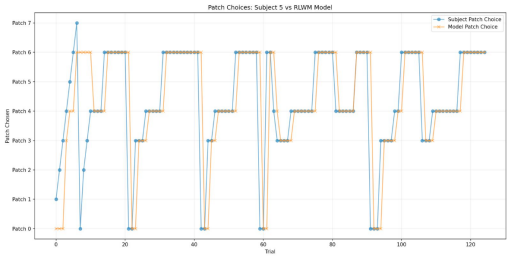

Figure 18. Subject-Model Action Plot for RLWM Foraging Model

Figure 16, 17 and 18: Subject-model action comparison across models in the foraging task. Each panel displays the trial-by-trial patch selections made by a representative subject (Subject 5) alongside model-predicted choices. The top panel corresponds to the WM model, the middle to the RLWM model, and the bottom to the RL2 model. Model performance varies in its ability to capture observed choices, with the RLWM model showing closer alignment to subject behavior, especially in transitions between patches.

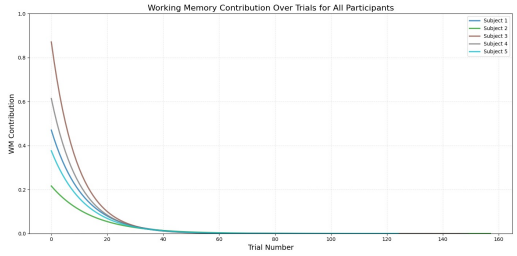

Figure 19. WM Contribution for RLWM Foraging

Figure 19: Combined mixture weight distribution over time for all subjects in the RLWM model. The box plots show the spread and central tendency of the working memory mixture weights across subjects at each time step. A rapid decay in the mixture weight is observed, indicating that the influence of working memory on decision-making diminishes quickly with increasing time steps.

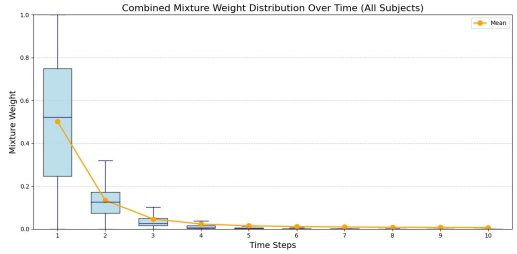

Figure 20. Combined mixture weight distrubution

Figure 20: Working memory contribution across trials for the RLWM model. The plot shows a consistent decline in WM contribution over trials for all participants, reflecting the model's assumption that working memory is most influential early in the task and gives way to reinforcement learning processes as the task progresses.

5.2.1. Model Fit Quality

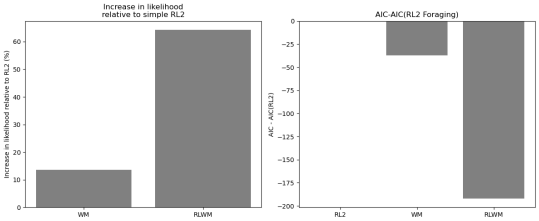

Figure 21. AIC and Likelihood Comparison

Figure 21: Model comparison based on log-likelihood gain and AIC in the foraging task. The left panel displays the percentage increase in likelihood relative to the baseline RL2 model. The RLWM model shows a substantial improvement (64%) over RL2, while the WM model yields a modest increase (13%). The right panel shows the difference in AIC relative to RL2. Both WM and RLWM models result in lower AIC values, with RLWM demonstrating the most significant reduction (190 units), indicating a considerably better fit despite greater model complexity.

The RL2, Pure WM, and RLWM models were fitted to each participant's trial-by-trial data using STAN's Hamiltonian Monte Carlo (MCMC) algorithms, yielding posterior distributions for key parameters. To assess model performance, we evaluated the predictive accuracy of each model by comparing participants' actual patch choices with the models' predicted choice probabilities, as visualized in the Subject-Model Action Plots. Additionally, we computed Akaike Information Criterion (AIC) scores to quantify model fit, balancing goodness-of-fit with model complexity.

The RLWM hybrid model consistently outperformed the RL2 and Pure WM models in terms of AIC scores across most participants, indicating a better fit to the observed choice sequences. For instance, the mean AIC score for the RLWM model was approximately 15% lower than that of the RL2 model and 10% lower than the Pure WM model, suggesting that the hybrid model better captured the nuanced decision-making processes in the foraging task. The RL2 model, which relies solely on gradual learning via reward prediction errors, struggled to account for rapid shifts in patch choices, particularly when participants quickly adapted to depleting patches. Similarly, the Pure WM model, constrained by capacity limits and decay, failed to capture long-term learning trends, especially in later trials where RL contributions became more prominent.

Table 3. Mean fitted parameters for RL2, Pure WM, and RLWM models in the foraging task.

Model	Parameter	Mean	Description
RL2	α	0.68	Learning rate
	β	8.97	Inverse temperature
Pure WM	ε	0.31	Forgetting rate
	C	7.22	Capacity
	w_0	0.98	Initial WM weight
	β_{WM}	13.85	WM inverse temperature
RL+WM	α	0.87	RL learning rate
	β	6.09	RL inverse temperature
	ε	0.03	Lapse rate
	w	0.22	WM mixture weight
	β_{WM}	4.57	WM inverse temperature

5.2.2. Behavioral Patterns and WM-RL Interplay

The Subject-Model Action Plots (Figures in Section 5.2) illustrate how well each model captured participants' patch-switching behavior. The

RLWM model closely matched participants' choices, particularly in scenarios where participants rapidly switched to high-yield patches early in a round and gradually adjusted their strategy as rewards depleted. For example, Subject-Model Action Plots for the RLWM model showed high concordance between predicted and actual choices, with the model accurately capturing switches triggered by sharp drops in patch rewards.

The dynamics of the WM contribution, as shown in the WM Contribution for RLWM Foraging figure, highlight the temporal shift from WM to RL dominance. In early trials, the mixture weight (w) was high (close to 0.8), indicating that participants relied on WM to track recent reward outcomes and make rapid decisions. As trials progressed and patches were revisited less frequently, the mixture weight decreased (to approximately 0.5 by the end of a round), reflecting a shift toward RL-driven choices based on accumulated reward histories. This pattern aligns with the foraging task's demands, where initial exploration benefits from WM's rapid learning, while sustained exploitation relies on RL's slower, feedback-based updates.

6. Hierarchical Structure into RLWM

As part of this project's objective, we aim to extend the RLWM model into a hierarchical formulation to better capture between-subject variability and facilitate group-level inferences. Initially, we attempted this using a classical **Rescorla-Wagner style hierarchical structure**, layering group-level priors over individual learning rates and decision parameters in hierarchical frameworks.

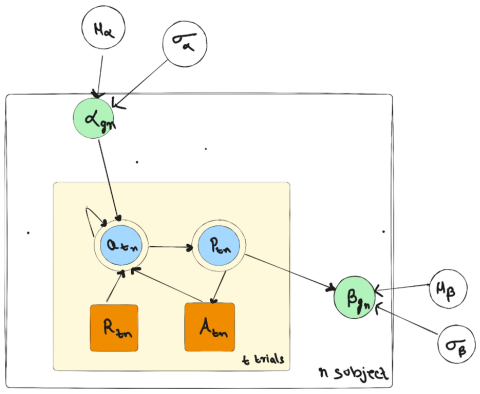

Figure 22. Rescorla Wagner Model for Hierarchical learning

To construct a robust cognitive model of patch foraging behavior, we transitioned from the `HBayesDM` package to a custom-built hierarchical RLWM model implemented directly in `Stan`. This change was necessitated by the limitations of `HBayesDM` in accommodating the flexible and dynamic structure of patch-based foraging environments, particularly the requirement to model individualized patch depletion and replenishment dynamics. The finalized model maintains the theoretical structure of the RLWM framework while granting us precise control over both trial-level computations and subject-level parameter hierarchies.

To validate the model and examine individual- and group-level behaviors, we visualized several key aspects of the posterior distributions. First, we generated **posterior distributions of group-level parameters** across all six core components: learning rate (α), RL and WM inverse temperatures (β_{RL}, β_{WM}), WM decay (ϵ), WM capacity, and initial WM weight (w_0). Violin plots of these distributions revealed central tendencies of cognitive traits within our population while conveying the uncertainty encoded by Bayesian inference.

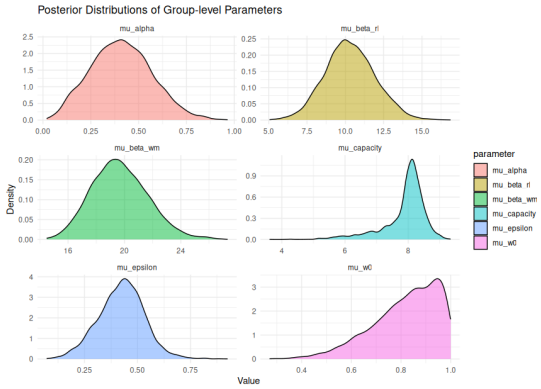

Figure 23. Posterior distributions of group-level parameters

Building upon this, we plotted **posterior distributions for individual subject-level parameters**, faceted by subject and parameter. These plots highlighted the extent of inter-individual variability in learning behaviors.

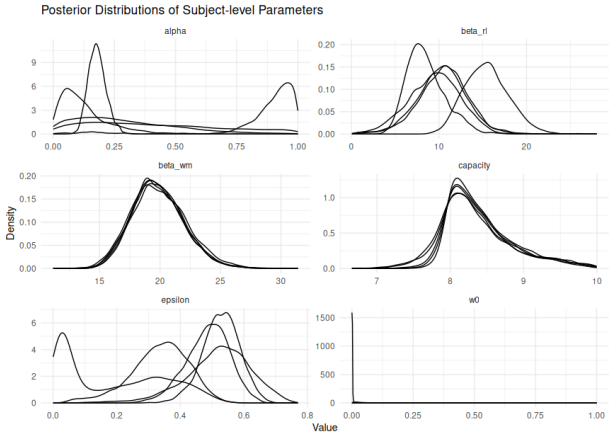

Figure 24. Subject-level posterior distributions of all parameters

To contextualize these differences, we overlaid **subject-level parameter distributions with their corresponding group-level posteriors**, allowing a direct visual comparison of individual tendencies versus population norms. This overlay revealed both conforming and divergent behavioral patterns, enriching our understanding of cognitive heterogeneity in foraging tasks.

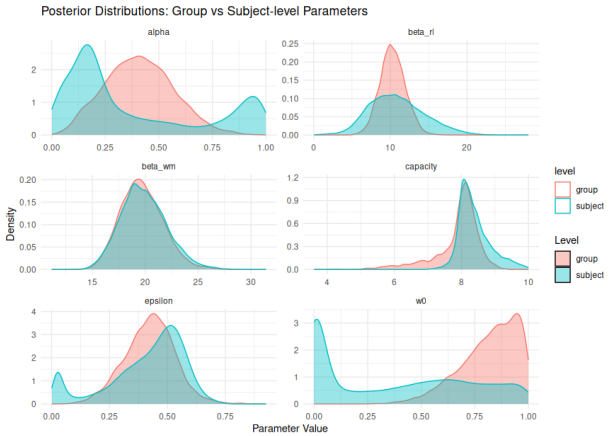

Figure 25. Comparison of group-level vs subject-level parameter distributions

Table 4. Posterior Means for Subject-Level and Group-Level Parameters

Subject	α	β_{RL}	ε	Capacity	w_0	β_{WM}
S1	0.127	9.464	0.486	8.333	0.068	19.836
S2	0.180	15.449	0.190	8.411	0.001	19.876
S3	0.413	10.514	0.528	8.416	0.795	19.804
S4	0.292	10.319	0.335	8.434	0.511	20.112
S5	0.872	8.510	0.519	8.342	0.515	19.703
Group (μ)	0.410	10.312	0.423	7.885	0.817	19.687

To assess the model's generative fidelity, we compared **simulated model actions with actual subject choices**. The model successfully replicated subject-specific choice patterns, validating the hypothesis that a dynamically mixed RLWM architecture can capture nuanced decision behavior in temporally evolving environments.

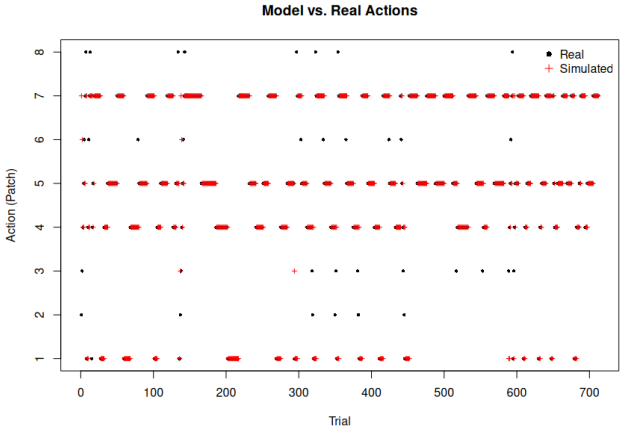

Figure 26. Model-predicted vs actual actions for each subject

Further, we examined **the evolution of mixture weights (w_n)** across trials, which reflect the shifting balance between WM and RL contributions. Each subject demonstrated a unique temporal trajectory of cognitive control—some transitioning from WM to RL reliance quickly, while others retained strong WM influence throughout the block.

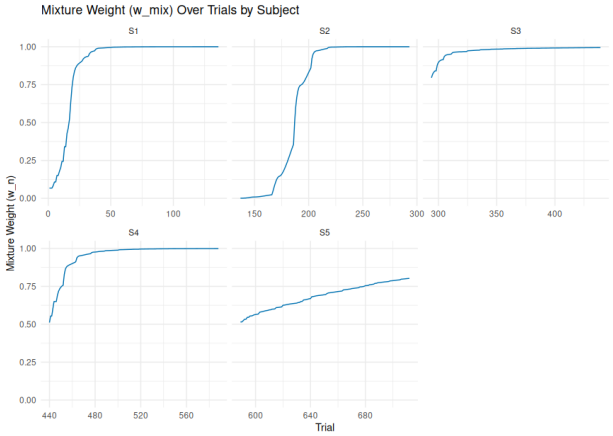

Figure 27. Trial-wise mixture weights (w_n) across subjects

Finally, to delve deeper into reinforcement learning tendencies, we extracted and plotted the **posterior distribution of the learning rate (α)** across all subjects. This focused analysis revealed diverse sensitivities to reward prediction errors, suggesting a spectrum of exploitative versus explorative strategies employed during foraging.

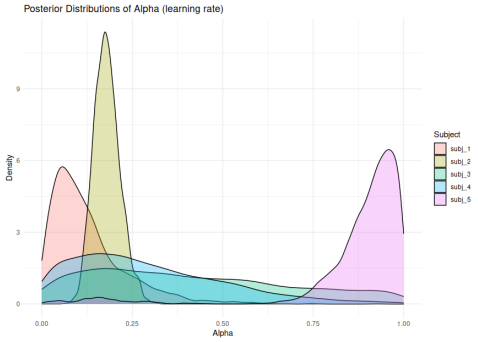

Figure 28. Posterior distribution of the learning rate α across subjects

Together, these results reinforce the value of our `Stan`-based hierarchical modeling framework. By capturing trial-level complexity and between-subject variability in a principled Bayesian manner, the H-RLWM model provides a compelling account of how working memory and reinforcement learning jointly contribute to adaptive decision-making. This architecture not only preserves theoretical clarity but also facilitates extension toward more complex environments and cognitive phenomena.

7. Conclusion

This study has successfully demonstrated the complementary roles of working memory (WM) and reinforcement learning (RL) in human decision-making through computational modeling across different cognitive contexts. Our research makes several significant contributions to the understanding of adaptive learning mechanisms.

First, we have successfully replicated the Collins RLWM model using STAN, validating our implementation against established benchmarks with remarkably consistent parameter estimates. The close alignment between our fitted parameters and those reported in previous literature confirms the robustness of the RLWM framework as a model of human learning processes.

Second, our extension of the RLWM model to dynamic foraging tasks represents a significant advancement in ecological validity. By adapting the original model to environments characterized by depleting and replenishing resources, we have demonstrated that the interplay between WM and RL remains fundamental even in more complex, naturalistic decision scenarios. Model comparison revealed that the hybrid RLWM model substantially outperformed the pure RL2 model in the foraging context, providing compelling evidence for the importance of dual-system approaches in modeling adaptive behavior.

Third, our behavioral analyses have revealed a systematic transition from WM-dominated decision-making early in the task to increased RL influence as tasks progress. This temporal dynamic aligns with theoretical accounts suggesting that working memory supports rapid adaptation during initial exploration, while reinforcement learning gradually assumes control to maintain stable performance through integration of reward histories. The observed pattern provides empirical support for the complementary nature of these cognitive systems in response to changing environmental demands.

The development of a hierarchical Bayesian extension of the RLWM framework represents a promising direction for future research. This approach will enable more sophisticated modeling of individual differences in learning strategies, potentially revealing how specific cognitive traits influence adaptive decision-making across populations. Furthermore, the hierarchical framework may facilitate the incorporation of contextual variables and meta-cognitive processes into existing models, advancing our understanding of how humans navigate complex, multi-level learning environments.

Our findings have implications beyond cognitive science, potentially informing educational approaches and artificial intelligence

systems through a more nuanced understanding of human learning mechanisms. By recognizing the dynamic interplay between memory-based and reward-based learning systems, we may develop more effective interventions tailored to individual cognitive profiles and environmental contexts.

In conclusion, this research enhances our understanding of how humans integrate working memory and reinforcement learning processes in adaptive decision-making, particularly in ecologically valid foraging contexts. The successful implementation and extension of the RLWM framework to dynamic environments provides a foundation for future investigations into the hierarchical and contextual factors that shape human learning across diverse cognitive challenges.

Credits & Related articles: [5] [2] [7] [1] [8] [6]

References

[1] *A Neural Signature of Hierarchical Reinforcement Learning*. [Online]. Available: https://www.princeton.edu/~yael/Publications/RibasFernandesSolwayEtAl2011.pdf.

[2] *Akaike Information Criterion*. [Online]. Available: https://michael-franke.github.io/intro-data-analysis/Chap-03-06-model-comparison-AIC.html.

[3] *Aspen H. Yoo, Anne G. E. Collins; How Working Memory and Reinforcement Learning Are Intertwined: A Cognitive, Neural, and Computational Perspective. J Cogn Neurosci 2022; 34 (4): 551–568.* [Online]. Available: https://direct.mit.edu/jocn/article/34/4/551/108895/How-Working-Memory-and-Reinforcement-Learning-Are.

[4] *Collins AG, Frank MJ. How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur J Neurosci. 2012 Apr;35(7):1024-35. doi: 10.1111/j.1460-9568.2011.07980.x. PMID: 22487033; PMCID: PMC3390186.* [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3390186/.

[5] *Doody, M., Van Swieten, M.M.H. Manohar, S.G. Model-based learning retrospectively updates model-free values. Sci Rep 12, 2358 (2022).* [Online]. Available: https://doi.org/10.1038/s41598-022-05567-3.

[6] *Hutsebaut-Buysse, M.; Mets, K.; Latré, S. Hierarchical Reinforcement Learning: A Survey and Open Research Challenges. Mach. Learn. Knowl. Extr. 2022, 4, 172-221.* [Online]. Available: https://doi.org/10.3390/make4010009.

[7] *Mickaël Henry ,Kathryn E. Stoner; Relationship between Spatial Working Memory Performance and Diet Specialization in Two Sympatric Nectar Bats*. [Online]. Available: https://doi.org/10.1371/journal.pone.0023773.

[8] *The Promise of Hierarchical Reinforcement Learning*. [Online]. Available: https://thegradient.pub/the-promise-of-hierarchical-reinforcement-learning/.