DynamicRL: Data-Driven Estimation of Trial-by-Trial Reinforcement Learning Parameters

Hua-Dong Xiong[1]
School of Psychology
Georgia Tech
hdx@gatech.edu

Li Ji-An[1]
Neurosciences Program
UC San Diego
jil095@ucsd.edu

Marcelo G. Mattar[2]
Department of Psychology
NYU
marcelo.mattar@nyu.edu

Robert C. Wilson[2]
School of Psychology
Georgia Tech
rwilson337@gatech.edu

Abstract

In uncertain and dynamic environments, biological agents must adapt their decision-making strategies to maximize rewards. Traditional reinforcement learning (RL) models typically assume that such adaptation is governed by dynamic value updates controlled by fixed parameters or predefined schedules. However, these assumptions limit the models' ability to capture the flexible and context-sensitive nature of biological decision-making. To overcome this limitation, we introduce *DynamicRL*, a novel framework that estimates RL parameters from behavioral data on a trial-by-trial basis. We demonstrate that DynamicRL substantially improves the predictive performance of standard RL models across eight decision-making tasks, thereby reducing scientific regret.

DynamicRL captures the rich temporal variability inherent in decision-making behavior, achieving predictive performance comparable to that of recurrent neural networks trained directly on the data, while preserving the interpretability and theoretical grounding of RL models. Moreover, it enables the examination of how agents dynamically adjust RL parameters in response to environmental changes, offering insights into the cognitive mechanisms underlying such adaptations. Thus, DynamicRL serves as an efficient data-driven framework for estimating RL parameters, facilitating fine-grained behavioral analysis with potential applications in computational psychiatry and neuroscience.

Keywords: computational modeling; reinforcement learning; decision-making; multi-armed bandit task; neural network

In uncertain and dynamic environments, biological agents must continuously refine their decision-making strategies to adapt. Reinforcement learning (RL) has long been recognized as a powerful framework for explaining how agents learn from feedback by updating value estimates based on experience (Rescorla & Wagner, 1972). However, in RL models, parameters such as learning rates or choice stochasticity are typically treated as fixed over time or are allowed to vary only according to a predefined algorithm, such as Kalman filter-like learning algorithms (Nassar, Wilson, Heasly, & Gold, 2010; Speekenbrink & Konstantinidis, 2015). Such fixed or algorithmically constrained parameters limit models' ability to capture how agents adapt their decision-making strategies in response to task demands, such as volatile reward structures.

Recently, recurrent neural networks (RNNs) have emerged as a powerful paradigm for modeling decision-making behavior (Ji-An, Benna, & Mattar, 2023; Miller, Eckstein, Botvinick, & Kurth-Nelson, 2023; Xiong, Ji-An, Mattar, & Wilson, 2023). Due to the flexibility of these models, they can often achieve higher predictive performance than classical cognitive models such as RL. However, better prediction accuracy may not always lead to better understanding without the lens of normative explanations. This raises a question: can we improve the predictive power of the RL framework while preserving the interpretability of the normative RL framework?

Here, we propose a novel modeling framework, *DynamicRL*, which dynamically estimates trial-by-trial reinforcement learning (RL) parameters from behavioral data. We demonstrate that DynamicRL outperforms classical RL models with fixed parameters across eight decision-making datasets on unseen test data, while achieving predictive performance comparable to that of recurrent neural network models. These results suggest that DynamicRL offers a general framework for minimizing scientific regret (Agrawal, Peterson, & Griffiths, 2020).

To illustrate how DynamicRL captures parameter dynamics, we use the Horizon Task (Wilson, Geana, White, Ludvig, & Cohen, 2014) as a case study, revealing subjects' ongoing adaptation to task structure. We further examine factors influencing the adaptation process across tasks and find that rewards and reward prediction errors show distinct regression patterns. In summary, DynamicRL provides a data-driven approach for modeling complex temporal structures in behavioral data, enabling more nuanced characterization of individual differences, with potential applications in computational psychiatry and neuroscience.

Results

Reinforcement learning model

We consider a model-free RL model with three parameters for all decision-making tasks we studied. The action value $Q_t(a)$ for action a at time t is updated by the reward r_t:

$$Q_t(a) = Q_{t-1}(a) + \alpha(r_t - Q_{t-1}(a)), \tag{1}$$

where α denotes the learning rate that determines the degree to which beliefs are updated by new observations.

For choice selection, we use a softmax function:

$$p_t(a_t) = \frac{e^{\beta(Q_t(a_t)+\kappa I(a_t=a_{t-1}))}}{\sum_{a'} e^{\beta(Q_t(a')+\kappa I(a'=a_{t-1}))}} \tag{2}$$

where β is the inverse temperature parameter controlling the stochasticity of choices (i.e., the exploration-exploitation

[1]Co-first author
[2]Co-senior author

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

tradeoff), and κ captures choice perseveration — the tendency to repeat the previous action. The indicator function $I(a_t = a_{t-1})$ equals 1 if the current action matches the previous one, and 0 otherwise.

DynamicRL: estimating trial-by-trial RL parameters

Classical RL models assume that learning parameters remain fixed across trials. In contrast, *DynamicRL* assumes that model-free RL agents dynamically adapt their decision-making strategies by adjusting parameters such as the learning rate (α_t), choice temperature (β_t), and choice perseverance (κ_t). DynamicRL estimates these parameters on a trial-by-trial basis using gated recurrent units (GRUs). Conceptually, this approach is related to the fast weights framework (Schmidhuber, 1992) and hypernetworks (Ha, Dai, & Le, 2017), in which one model generates parameters for another.

In our implementation, the learning rate α_t and choice temperature β_t are modeled as functions of past rewards and actions (Eq. 3). In contrast, the perseveration parameter κ_t depends solely on past actions, consistent with its interpretation as capturing reward-independent choice repetition (Miller, Shenhav, & Ludvig, 2019). This design choice is critical: since κ_t is directly added to the action logits, allowing it to depend on reward history could enable it to absorb the functional roles of α_t and β_t, effectively reducing the model to an unconstrained RNN and undermining interpretability. By constraining κ_t's inputs, we preserve functional separation among the parameters and retain the interpretability grounded in normative RL theory.

$$\begin{aligned} \alpha_t, \beta_t, h_t &= \mathrm{GRU}_h(a_{t-1}, r_{t-1}, E_\mathrm{p}, E_{\mathrm{task},t}, h_{t-1}), \\ \kappa_t, g_t &= \mathrm{GRU}_g(a_{t-1}, E_\mathrm{p}, E_{\mathrm{task},t}, g_{t-1}), \end{aligned} \tag{3}$$

where E_p is a subject embedding learned from data to capture individual differences, $E_{\mathrm{task},t}$ encodes task-related information such as experimental conditions and trial numbers, h and g denote the hidden states of GRU_h and GRU_g, respectively (Fig. 1).

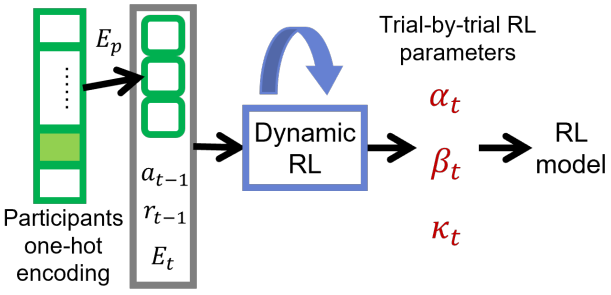

Figure 1: **DynamicRL model.** Recurrent neural networks are used to estimate RL parameters on a trial-by-trial basis.

DynamicRL outperforms classical RL models across eight decision-making tasks

We systematically examined the performance of DynamicRL compared to classical RL models with fixed parameters across a diverse set of datasets from eight decision-making experiments. All datasets involved the commonly studied multi-armed bandit task, where subjects choose between multiple options with unknown expected rewards. These expected rewards can drift over time, reflecting changes in the volatile environments. Additionally, rewards may be stochastic, introducing noise into observations. In the model-free RL framework, solving this task is framed as learning the value of different options. Thus, subjects should sample options to estimate their expected rewards from noisy observations, and make decisions to maximize outcomes.

The eight datasets include the Horizon Task in humans (Wilson et al., 2014), a two-armed restless bandit task in mice (Chen, Knep, Han, Ebitz, & Grissom, 2021), a four-armed restless bandit task in humans (Bahrami & Navajas, 2020), three-armed restless probabilistic reversal learning tasks in both humans (Suthaharan et al., 2021; Reed et al., 2020) and mice (Groman, Rich, Smith, Lee, & Taylor, 2018), two two-armed bandit tasks in humans (Gershman, 2018), a two-armed instrumental learning task in humans (Dezfouli, Griffiths, Ramos, Dayan, & Balleine, 2019), and the Iowa Gambling Task in humans (Steingroever et al., 2015).

To assess model performance, we conducted nested cross-validation (where the outer loop selects testing folds, and the inner loop further selects training and validation folds) (Fig. 2). The results (Figure 3) demonstrate that DynamicRL models consistently outperform classical RL models across all eight datasets ($t(7) = -5.863$, $p = 0.001$, Cohen's $d = -0.433$). Moreover, DynamicRL models achieved performance almost comparable to, though slightly worse than, GRU models that are directly fitted to behavioral data ($t(7) = 4.160$, $p = 0.004$, Cohen's $d = 0.087$).

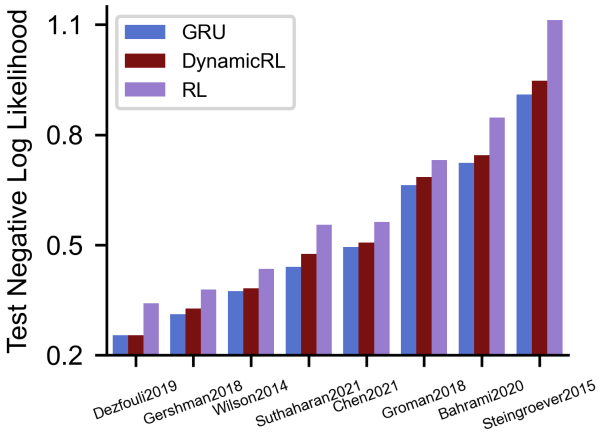

Figure 2: **Model fitting performance across eight datasets.** The DynamicRL model's performance, evaluated with nested cross-validation (lower negative log-likelihood indicates better fitting performance), surpasses that of the classical RL model with fixed parameters.

The average fitting performance across all tasks is summarized in Fig. 3. We assume that intrinsic noise in the true

data-generating process sets an upper bound on predictive performance (hypothetical green bar). The best-fitting GRU models represent the highest level of predictability attainable given the available data. Compared to classical RL models, DynamicRL models significantly reduce scientific regret (Agrawal et al., 2020) by explaining more variance that is predictable by GRU models. Importantly, DynamicRL models preserve interpretability and adhere to the normative RL framework.

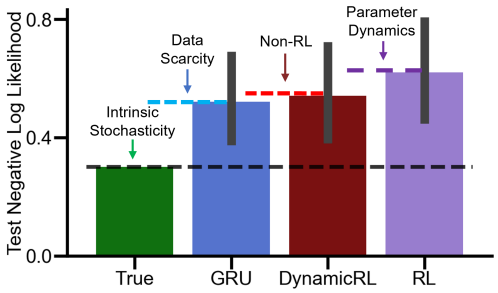

Figure 3: **Behavioral variability explained by models.** The true data-generating model (green) is *hypothetical* and unknown. Previous research commonly used RL models (purple) to explain behavioral data, while the GRU model (blue) offers superior predictive accuracy but lacks interpretability. Our DynamicRL model (red) minimizes scientific regret while retaining interpretability. Error bars indicate 95% confidence intervals (CI).

DynamicRL captures richer temporal variability

Researchers typically use reinforcement learning (RL) parameters to characterize individual differences in decision-making strategies. *DynamicRL* offers a more flexible and temporally precise framework for capturing dynamic strategy adaptation. Using this approach, we aim to quantify how much variability in strategy—reflected in fluctuations of RL parameters—stems from temporal dynamics rather than from across-subject differences alone.

To this end, we partition the total variance of RL parameters estimated by DynamicRL into subject-level and block-level components, following the law of total variance. For the subject-level component, we averaged parameters within each subject and computed the variance across subjects. For the block-level component, we computed the variance within each subject across blocks and then averaged these values across subjects.

Our analysis shows that the proportion of RL parameter variance attributable to subject- and block-level sources varies across tasks (Fig. 4). Notably, block-level variance is often substantial and can even exceed subject-level variance. This indicates that DynamicRL captures meaningful temporal fluctuations in RL parameters—an aspect often overlooked by conventional models.

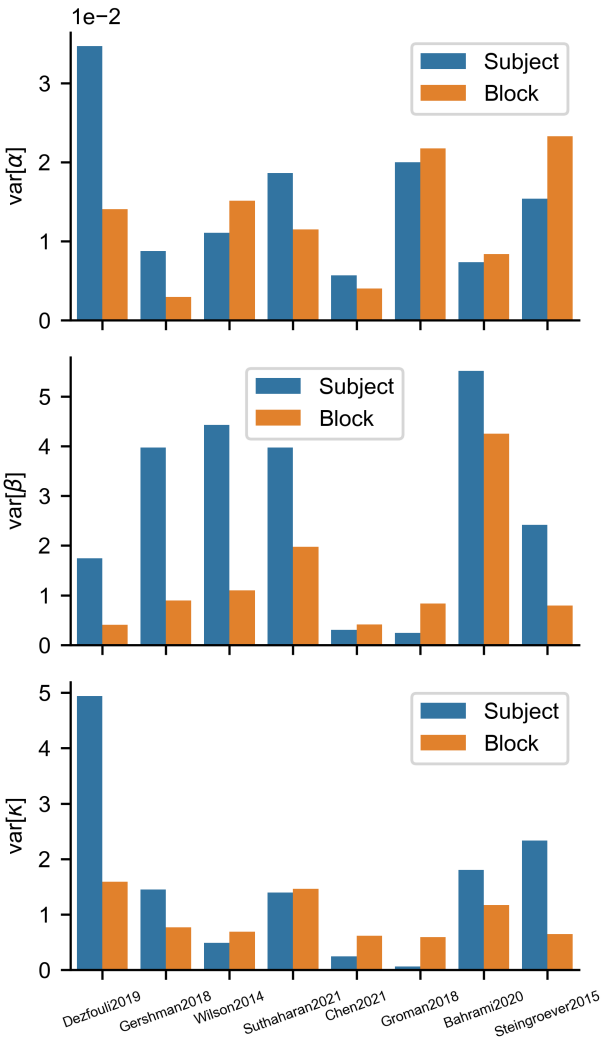

Figure 4: **Subject-level and block-level variance of RL parameters.** The proportion of variance in RL parameters attributed to subject-level and block-level components varies across tasks and RL parameters.

Case study: horizon task

After demonstrating that DynamicRL captures a substantial amount of behavioral variability, we present an example to illustrate how dynamic RL parameters evolve in a value-based decision-making task. We use the Horizon Task (Wilson et al., 2014) as a concrete case, as its well-defined structure enables clear examination of how subjects' estimated parameters change under different conditions. Each block begins with four forced-choice trials, during which subjects passively observe actions and outcomes, ensuring equal information about both options. This is followed by either 1 or 6 free-choice trials, corresponding to the horizon-1 and horizon-6 conditions (Fig. 5a). Option rewards are sampled independently from a distribution with a fixed mean within each block (Fig. 5b). For DynamicRL models, we provide the task inputs $E_{\text{task},t}$, including the current trial number, horizon condition, and trial

type (forced or free).

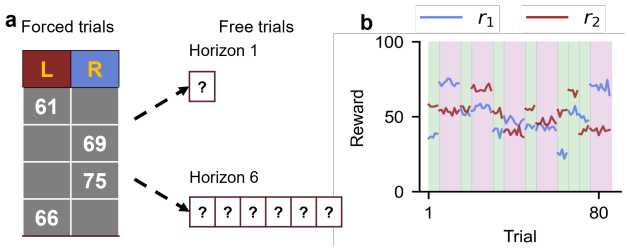

Figure 5: **Horizon task. a,** Subjects first passively observed four forced-choice trials. They then made free choices for either one or six trials, corresponding to the horizon-1 and horizon-6 conditions, respectively. **b,** Example reward structure in the horizon task, showing interleaved horizon-1 (light green) and horizon-6 (light red) blocks. Subjects estimated the value of the two options (resampled each block) based on noisy reward outcomes (r_1 and r_2).

We present the RL parameters estimated by DynamicRL for the Horizon Task, grouped by condition and averaged across trials (Fig. 6). During the four forced-choice trials, the learning rate α in the horizon-1 condition is higher than in the horizon-6 condition (Fig. 6a), suggesting that subjects update reward values more aggressively when faced with fewer opportunities to act. In both conditions, the learning rate α decreases over time, consistent with predictions from the Kalman filter (Piray & Daw, 2021), which posits that learning rates decline as uncertainty about expected rewards diminishes.

We also examined how parameters evolve during the free-choice trials in the horizon-6 condition (Fig. 6b,c,d). The learning rate α continues to decline, indicating a decreasing influence of new information. Meanwhile, both β and κ increase, reflecting a shift toward exploitation and perseveration as subjects become more confident in their reward estimates and the number of remaining trials decreases.

The dynamic strategy adaptation revealed by DynamicRL qualitatively aligns with rational models of behavioral adaptation (see Discussion for details, summarized in Table 1), suggesting that subjects adjust their strategies in a task-dependent and adaptive manner. These patterns are typically overlooked by RL models with fixed or predefined parameter dynamics, whereas DynamicRL flexibly captures such adaptations by modeling trial-by-trial changes in RL parameters.

What drives this adaptation process?

The results above suggest that subjects adapt their decision-making strategies dynamically—and possibly rationally—in response to task demands. This raises a critical question: what temporal strategies drive such adaptation? Specifically, do subjects proactively adopt distinct RL parameter profiles across trials, independent of task feedback, or do they adjust these parameters in a feedback-driven manner based on environmental cues?

For example, in environments that are volatile but exhibit low stochasticity, a large reward prediction error (RPE) may

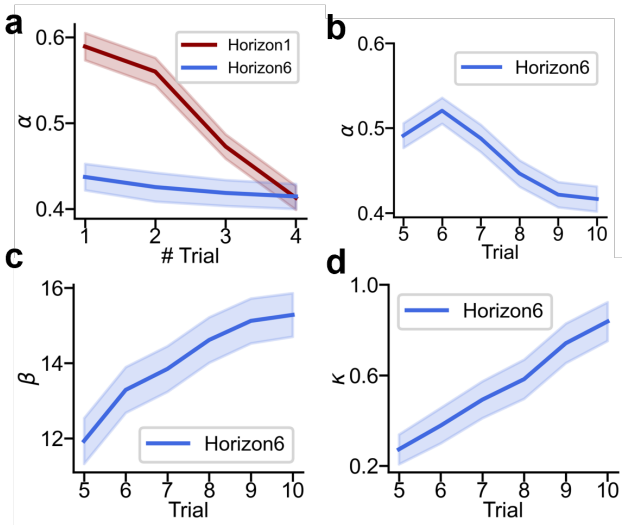

Figure 6: **RL parameters in the horizon task change over trials. a,** Average learning rate (α) across subjects during forced trials. **b,c,d** Average values of α (b), β (c), and κ (c) across subjects during free trials in horizon 6 condition. Shaded areas indicate 95% CI.

signal a shift in environmental contingencies, prompting a rapid increase in the learning rate (α). Alternatively, subjects may rely on simpler heuristics—such as consistently favoring recently high-rewarded options—without explicitly encoding environmental statistics such as volatility or stochasticity. In this case, recent reward, rather than RPE, would emerge as the main predictor of changes in RL parameters.

To investigate the drivers of dynamic parameter adaptation, we conducted a regression analysis using Lasso regularization to predict changes in RL parameters. Predictors included task-structure (e.g., trial number), stimulus (e.g., rewards), internal variables (e.g., RPEs), and previous-trial parameters. By examining how these variables contribute to temporal variability in RL parameters across tasks, we sought to identify the sources and mechanisms driving changes in RL strategies, thereby gaining insight into the cognitive processes underlying task adaptation. We found that dynamic RL parameters α, β, and κ were all well explained by the selected predictors (Fig. 7; averaged over tasks), suggesting that these variables account for the primary sources of variability in the dynamically estimated RL parameters.

To examine the relationship between predictors and changes in RL parameters, we first formulated task-specific hypotheses based on rational analysis (see Table 1). For tasks with high volatility and low stochasticity (Chen2021, Suthaharan2021, Groman2018, Bahrami2020), we predicted that RPEs would increase α and decrease κ. In contrast, for tasks with no volatility (Dezfouli2019, Gershman2018, Wilson2014, Steingroever2015), we expected rewards to increase κ. Trial number was not expected to influence RL parameters in restless bandit tasks but was predicted to decrease α and increase β and κ in non-restless tasks.

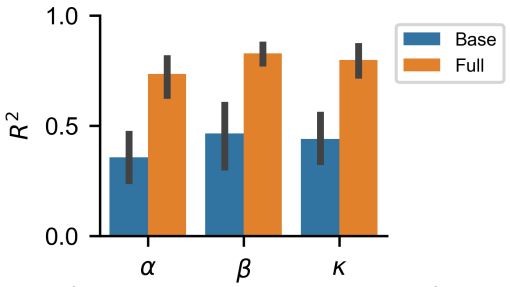

Figure 7: **R^2 of the regression analysis.** R^2 values from the regression analysis for dynamic RL parameters α, β, and κ, averaged across tasks. The base model includes only RL parameters from previous trials, while the full model incorporates all predictors. Error bars indicate 95% CI.

We then analyzed the regression coefficients across all eight tasks (Fig. 8). The results revealed notable variability in strategy adaptation across tasks, with several predicted patterns supported by the data.

As a case study, we examine the Horizon Task, where we observe a general trend of decreasing α and increasing β and κ over time (Fig. 6). A naive explanation might attribute this to the decreasing number of remaining trials. However, regression analysis (Wilson2014, third column in Fig. 8) shows that trial number does not significantly predict α or β, suggesting subjects are not relying on a preemptive countdown strategy. Instead, α is mainly driven by RPEs, while β appears influenced by unobserved latent factors. By contrast, κ increases with trial number, possibly reflecting growing choice perseverance as fewer trials remain.

In the Chen2021 and Groman2018 tasks—restless bandit paradigms with mice—β and κ were primarily predicted by reward, suggesting a reliance on external stimuli rather than internal decision variables. Similarly, in the structurally identical Suthaharan2021 (human) and Groman2018 (mouse) tasks, reward was the dominant predictor for both α and κ.

For α, this reflects rational adaptation to environments with low measurement noise but high volatility, where rapid changes make value learning less effective. In such cases, a win-stay/lose-shift heuristic suffices: it bypasses internal value representation by directly mapping recent outcomes to choices, making reward the primary driver.

Interestingly, the main predictor of β differs by species: humans rely more on RPEs, while mice rely on raw reward. This suggests humans engage internal representations of uncertainty more extensively, whereas mice respond more directly to observed outcomes.

Together, these findings clarify the cognitive processes underlying dynamic RL adaptation, reveal species-specific strategies, and provide a functional lens for comparing RL tasks. They underscore the diversity of adaptive behavior and point to new directions for understanding how strategy selection varies across contexts and species.

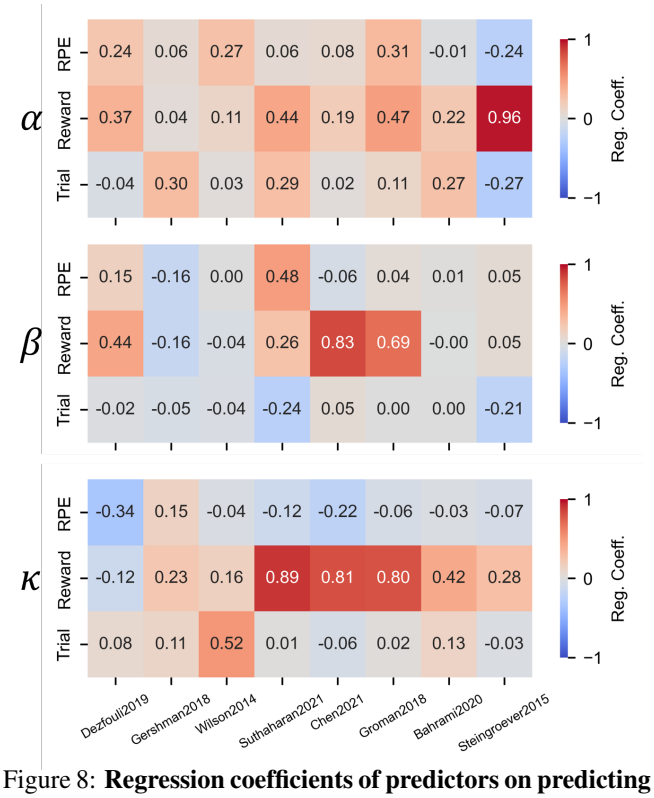

Figure 8: **Regression coefficients of predictors on predicting varying RL parameters.**

Discussion

In this work, we introduce the *DynamicRL* framework for estimating trial-by-trial RL parameters from behavioral data in decision-making tasks. Across eight datasets, DynamicRL outperforms classical RL models with fixed parameters and achieves predictive performance comparable to state-of-the-art recurrent neural networks trained directly to predict behavior, while uncovering rich temporal dynamics in strategy use.

We further show that block-level variability in parameter estimates—reflecting temporal variation across trials—is substantial compared to subject-level variance. This indicates that DynamicRL captures a critical source of behavioral variability often overlooked by traditional models. Using the Horizon Task as a case study, we demonstrate that dynamic RL parameters adapt to task structure in ways consistent with rational behavior, suggesting that subjects adjust their strategies in response to environmental demands.

To investigate the drivers of these adaptations, we conduct regression analyses identifying factors that influence dynamic changes in RL parameters. This analysis offers a useful tool for generating hypotheses about cognitive processes at fine-grained temporal resolution. For instance, in the Horizon Task, reward prediction errors primarily drive updates to the learning rate α, while trial number influences adjustments in choice perseverance κ, offering insights for future experimental design.

By equipping RL models with dynamically varying parameters, DynamicRL improves model performance and captures

strategic adaptation in changing environments. While some existing models in cognitive science incorporate time-varying parameters, they face key limitations. Hierarchical Gaussian filters (C. Mathys, Daunizeau, Friston, & Stephan, 2011; C. D. Mathys et al., 2014) impose strong structural assumptions (e.g., predefined probabilistic graphs), which can lead to inaccurate predictions if violated. Amortized simulation-based inference (Schumacher, Bürkner, Voss, Köthe, & Radev, 2023; Ger, Nachmani, Wolf, & Shahar, 2024) relies on pre-training neural networks on synthetic behavioral data, which may poorly generalize if the synthetic distribution diverges from real behavior. Hidden Markov models (Ashwood et al., 2022; Le et al., 2023) assume discrete latent states, lacking the flexibility to model continuous parameter changes.

In contrast, DynamicRL estimates continuously varying parameters directly from empirical data using flexible neural networks with minimal assumptions about task structure. This avoids the limitations of strong priors, distributional mismatches, and discretization artifacts.

Even under model misspecification, DynamicRL remains a valuable descriptive tool for capturing residual behavioral variance. For example, attentional lapses may manifest as transient reductions in α, allowing researchers to characterize variability across conditions or populations. Such trial-level characterizations of individual differences in decision-making offer promising avenues for research in computational psychiatry and neuroscience.

Our results show that subjects dynamically adjust α, β, and κ in response to task feedback. How should we interpret these adjustments? We hypothesize that these RL parameters may change in a rational manner, shaped by environmental structure and its dynamics. In the Bayesian reinforcement learning framework ((Dayan & Daw, 2008)), an ideal agent maintains posterior distributions over both environmental statistics (e.g., volatility and stochasticity) and RL parameters ($\alpha_t, \beta_t, \kappa_t$). By updating these posterior distributions of parameters dynamically in response to environmental cues, subjects may approximate rational strategies that adapt to uncertainty.

We provide the hypotheses of rational online adaptation strategies for dynamic and uncertain environments summarized in Table 1.

Concretely, the learning rate α may track environmental change-point probabilities using sliding-window variance estimators (Nassar et al., 2010), and adapt to both volatility and stochasticity (C. Mathys et al., 2011; Piray & Daw, 2021). In highly volatile environments—characterized by frequent shifts in expected values—rapid belief updating is necessary, requiring higher α. In contrast, in low-volatility but high-stochasticity settings, a lower α mitigates the influence of noisy reward fluctuations, reducing overreaction to random outcomes.

Adaptation of the inverse temperature β enables further refinement of strategies. Lower values of β promote exploration over exploitation, which is beneficial in volatile environments by preventing premature commitment to suboptimal options. Conversely, in stable settings, a higher β facilitates exploitation, accelerating convergence toward optimal choices. Similarly, adjustment of the perseverance parameter κ may also serve an adaptive function. While perseveration may appear heuristic, it can be resource-rational under computational constraints (Lieder & Griffiths, 2020), as it reduces policy complexity and cognitive burden (Gershman, 2020). Higher κ values may reduce unnecessary switching and promote consistent policy application in stable environments. However, excessive perseverance can hinder flexible adaptation under volatile conditions.

This framework aligns naturally with the control-as-inference perspective (Todorov, 2008; Levine, 2018), which recasts decision-making as probabilistic inference. From this view, participants adaptively update their strategies—in this case, RL parameters—to increase the posterior probability of actions that maximize cumulative reward.

Table 1: Hypothetical rational adaptation of RL parameters to environmental characteristics.

	Stage		**Volatility**		**stochasticity**		**Horizon**	
	Early	Late	Low	High	Low	High	Short	Long
α	$\uparrow$	$\downarrow$	$\downarrow$	$\uparrow$	$\uparrow$	$\downarrow$	$\uparrow$	$\downarrow$
β	$\downarrow$	$\uparrow$	$\uparrow$	$\downarrow$	$-$	$-$	$\uparrow$	$\downarrow$
κ	$\downarrow$	$\uparrow$	$\uparrow$	$\downarrow$	$-$	$-$	$\uparrow$	$\downarrow$

Previous research on RL has identified neural substrates underlying reward and value representation. For example, the dopaminergic midbrain encodes precision-weighted prediction errors (Schultz, Dayan, & Montague, 1997), which may influence updates to α. Similarly, prefrontal-amygdala circuits track uncertainty estimates (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006), potentially modulating β. Although *DynamicRL* is a behavioral model and does not explicitly capture neural mechanisms, the dynamic parameters it estimates offer a powerful lens for analyzing neural dynamics. Linking these parameters to neural data may clarify how adaptive RL processes are implemented in the brain and shed light on interactions between regions involved in learning and decision-making.

In conclusion, DynamicRL offers a flexible, data-driven framework for estimating RL parameters at the trial-by-trial level, capturing substantial behavioral variability that is often missed by models with fixed or predefined parameter dynamics. While grounded in the normative RL framework, our approach achieves predictive performance comparable to recurrent neural networks trained directly on behavioral data. The discovered parameter trajectories reveal adaptation to task structure and offer a means of connecting behavior to theories of rational adaptation, such as Bayesian reinforcement learning and control-as-inference. As a general modeling tool, DynamicRL has broad potential for generating hypotheses in computational psychiatry and neuroscience.

Acknowledgments

We are grateful to all the researchers who generously shared their datasets, enabling the analysis presented in this study. This work was supported by start-up funding from the Georgia Institute of Technology awarded to RCW. We also acknowledge the use of the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, which provided essential computational resources for this research.

References

Agrawal, M., Peterson, J. C., & Griffiths, T. L. (2020, April). Scaling up psychology via Scientific Regret Minimization. *Proceedings of the National Academy of Sciences*, *117*(16), 8825–8835. Retrieved 2023-09-04, from https://www.pnas.org/doi/full/10.1073/pnas.1915841117 (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1915841117

Ashwood, Z. C., Roy, N. A., Stone, I. R., Urai, A. E., Churchland, A. K., Pouget, A., & Pillow, J. W. (2022, February). Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience*, *25*(2), 201–212. Retrieved 2023-05-11, from https://www.nature.com/articles/s41593-021-01007-z (Number: 2 Publisher: Nature Publishing Group) doi: 10.1038/s41593-021-01007-z

Bahrami, B., & Navajas, J. (2020, August). *4 Arm Bandit Task Dataset*. Retrieved 2024-10-09, from https://osf.io/f3t2a/ (Publisher: OSF) doi: 10.17605/OSF.IO/F3T2A

Chen, C. S., Knep, E., Han, A., Ebitz, R. B., & Grissom, N. M. (2021, November). Sex differences in learning from exploration. *eLife*, *10*, e69748. Retrieved 2024-09-06, from https://doi.org/10.7554/eLife.69748 (Publisher: eLife Sciences Publications, Ltd) doi: 10.7554/eLife.69748

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006, June). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. Retrieved 2025-01-27, from https://www.nature.com/articles/nature04766 (Publisher: Nature Publishing Group) doi: 10.1038/nature04766

Dayan, P., & Daw, N. D. (2008, December). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453. Retrieved 2025-01-26, from https://doi.org/10.3758/CABN.8.4.429 doi: 10.3758/CABN.8.4.429

Dezfouli, A., Griffiths, K., Ramos, F., Dayan, P., & Balleine, B. W. (2019, June). Models that learn how humans learn: The case of decision-making and its disorders. *PLOS Computational Biology*, *15*(6), e1006903. Retrieved 2022-11-07, from https://dx.plos.org/10.1371/journal.pcbi.1006903 doi: 10.1371/journal.pcbi.1006903

Ger, Y., Nachmani, E., Wolf, L., & Shahar, N. (2024, January). Harnessing the flexibility of neural networks to predict dynamic theoretical parameters underlying human choice behavior. *PLOS Computational Biology*, *20*(1), e1011678. Retrieved 2024-06-06, from https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011678 (Publisher: Public Library of Science) doi: 10.1371/journal.pcbi.1011678

Gershman, S. J. (2018, April). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42. Retrieved 2024-06-02, from https://www.sciencedirect.com/science/article/pii/S0010027717303359 doi: 10.1016/j.cognition.2017.12.014

Gershman, S. J. (2020, November). Origin of perseveration in the trade-off between reward and complexity. *Cognition*, *204*, 104394. Retrieved 2025-01-26, from https://www.sciencedirect.com/science/article/pii/S0010027720302134 doi: 10.1016/j.cognition.2020.104394

Groman, S. M., Rich, K. M., Smith, N. J., Lee, D., & Taylor, J. R. (2018, March). Chronic Exposure to Methamphetamine Disrupts Reinforcement-Based Decision Making in Rats. *Neuropsychopharmacology*, *43*(4), 770–780. Retrieved 2024-09-13, from https://www.nature.com/articles/npp2017159 (Publisher: Nature Publishing Group) doi: 10.1038/npp.2017.159

Ha, D., Dai, A. M., & Le, Q. V. (2017, February). HyperNetworks.. Retrieved 2024-11-12, from https://openreview.net/forum?id=rkpACe1lx

Ji-An, L., Benna, M. K., & Mattar, M. G. (2023, May). *Automatic Discovery of Cognitive Strategies with Tiny Recurrent Neural Networks*. bioRxiv. Retrieved 2023-05-05, from https://www.biorxiv.org/content/10.1101/2023.04.12.536629v2 (Pages: 2023.04.12.536629 Section: New Results) doi: 10.1101/2023.04.12.536629

Le, N. M., Yildirim, M., Wang, Y., Sugihara, H., Jazayeri, M., & Sur, M. (2023, September). Mixtures of strategies underlie rodent behavior during reversal learning. *PLOS Computational Biology*, *19*(9), e1011430. Retrieved 2023-09-16, from https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011430 (Publisher: Public Library of Science) doi: 10.1371/journal.pcbi.1011430

Levine, S. (2018, May). *Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review*. arXiv. Retrieved 2023-02-18, from http://arxiv.org/abs/1805.00909 (arXiv:1805.00909 [cs, stat]) doi: 10.48550/arXiv.1805.00909

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1. Retrieved 2022-11-07, from https://www.cambridge.org/core/product/identifier/S0140525X1900061X/type/journal_article doi: 10.1017/S0140525X1900061X

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan,

K. E. (2011, May). A Bayesian Foundation for Individual Learning Under Uncertainty. *Frontiers in Human Neuroscience*, *5*. Retrieved 2024-10-23, from https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2011.00039/full (Publisher: Frontiers) doi: 10.3389/fnhum.2011.00039

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K. E. (2014, November). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, *8*. Retrieved 2024-10-23, from https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2014.00825/full (Publisher: Frontiers) doi: 10.3389/fnhum.2014.00825

Miller, K. J., Eckstein, M., Botvinick, M. M., & Kurth-Nelson, Z. (2023, June). *Cognitive Model Discovery via Disentangled RNNs*. bioRxiv. Retrieved 2023-07-21, from https://www.biorxiv.org/content/10.1101/2023.06.23.546250v1 (Pages: 2023.06.23.546250 Section: New Results) doi: 10.1101/2023.06.23.546250

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019, March). Habits without values. *Psychological Review*, *126*(2), 292–311. Retrieved 2025-02-02, from https://doi.apa.org/doi/10.1037/rev0000120 doi: 10.1037/rev0000120

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010, September). An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. *Journal of Neuroscience*, *30*(37), 12366–12378. Retrieved 2022-11-07, from https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0822-10.2010 doi: 10.1523/JNEUROSCI.0822-10.2010

Piray, P., & Daw, N. D. (2021, December). A model for learning based on the joint estimation of stochasticity and volatility. *Nature Communications*, *12*(1), 6587. Retrieved 2022-11-07, from https://www.nature.com/articles/s41467-021-26731-9 doi: 10.1038/s41467-021-26731-9

Reed, E. J., Uddenberg, S., Suthaharan, P., Mathys, C. D., Taylor, J. R., Groman, S. M., & Corlett, P. R. (2020, May). Paranoia as a deficit in non-social belief updating. *eLife*, *9*, e56345. Retrieved 2024-09-03, from https://doi.org/10.7554/eLife.56345 (Publisher: eLife Sciences Publications, Ltd) doi: 10.7554/eLife.56345

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical conditioning, Current research and theory*, *2*, 64–69. Retrieved 2024-10-09, from https://cir.nii.ac.jp/crid/1572543025504096640 (Publisher: Appleton-Century-Crofts)

Schmidhuber, J. (1992, January). Learning to Control Fast-Weight Memories: An Alternative to Dynamic Recurrent Networks. *Neural Computation*, *4*(1), 131–139. Retrieved 2025-01-10, from https://doi.org/10.1162/neco.1992.4.1.131 doi: 10.1162/neco.1992.4.1.131

Schultz, W., Dayan, P., & Montague, P. R. (1997, March). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. Retrieved 2025-01-26, from https://www.science.org/doi/full/10.1126/science.275.5306.1593 (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.275.5306.1593

Schumacher, L., Bürkner, P.-C., Voss, A., Köthe, U., & Radev, S. T. (2023, August). Neural superstatistics for Bayesian estimation of dynamic cognitive models. *Scientific Reports*, *13*(1), 13778. Retrieved 2024-11-04, from https://www.nature.com/articles/s41598-023-40278-3 (Publisher: Nature Publishing Group) doi: 10.1038/s41598-023-40278-3

Speekenbrink, M., & Konstantinidis, E. (2015, April). Uncertainty and Exploration in a Restless Bandit Problem. *Topics in Cognitive Science*, *7*(2), 351–367. Retrieved 2022-11-07, from https://onlinelibrary.wiley.com/doi/10.1111/tops.12145 doi: 10.1111/tops.12145

Steingroever, H., Fridberg, D. J., Horstmann, A., Kjome, K. L., Kumari, V., Lane, S. D., ... Wagenmakers, E.-J. (2015, June). Data from 617 Healthy Participants Performing the Iowa Gambling Task: A "Many Labs" Collaboration. *Journal of Open Psychology Data*, *3*(1). Retrieved 2024-10-09, from https://openpsychologydata.metajnl.com/articles/10.5334/jopd.ak doi: 10.5334/jopd.ak

Suthaharan, P., Reed, E. J., Leptourgos, P., Kenney, J. G., Uddenberg, S., Mathys, C. D., ... Corlett, P. R. (2021, September). Paranoia and belief updating during the COVID-19 crisis. *Nature Human Behaviour*, *5*(9), 1190–1202. Retrieved 2024-06-22, from https://www.nature.com/articles/s41562-021-01176-8 (Publisher: Nature Publishing Group) doi: 10.1038/s41562-021-01176-8

Todorov, E. (2008, December). General duality between optimal control and estimation. In *2008 47th IEEE Conference on Decision and Control* (pp. 4286–4292). Retrieved 2025-04-30, from https://ieeexplore.ieee.org/abstract/document/4739438 (ISSN: 0191-2216) doi: 10.1109/CDC.2008.4739438

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074–2081. Retrieved 2022-11-07, from http://doi.apa.org/getdoi.cfm?doi=10.1037/a0038199 doi: 10.1037/a0038199

Xiong, H.-D., Ji-An, L., Mattar, M. G., & Wilson, R. C. (2023, October). Distilling human decision-making dynamics: a comparative analysis of low-dimensional architectures.. Retrieved 2025-01-20, from https://openreview.net/forum?id=xW5JQo6TXO