HUMAN STRATEGY ADAPTATION IN REINFORCEMENT LEARNING RESEMBLES POLICY GRADIENT ASCENT

Hua-Dong Xiong[*]
School of Psychology
Georgia Institute of Technology
hdx@gatech.edu

Li Ji-An[*]
Neurosciences Graduate Program
University of California San Diego
jil095@ucsd.edu

Robert C. Wilson[†]
School of Psychology
Georgia Institute of Technology
rwilson337@gatech.edu

Marcelo G. Mattar[†]
Department of Psychology
New York University
marcelo.mattar@nyu.edu

ABSTRACT

A hallmark of intelligence is the ability to adapt behavior to changing environments, which requires adapting one's own learning strategies. This phenomenon is known as learning to learn in cognitive science and meta-learning in artificial intelligence. While this phenomenon is well-established in humans and animals, no quantitative framework exists for characterizing the trajectories through which biological agents adapt their learning strategies. Previous computational studies that either assume fixed strategies or use task-optimized neural networks do not explain how humans refine strategies through experience. Here we show that humans adjust their reinforcement learning strategies resembling principles of gradient-based online optimization. We introduce DynamicRL, a framework using neural networks to track how participants' learning parameters (e.g., learning rates and decision temperatures) evolve throughout experiments. Across four diverse bandit tasks, DynamicRL consistently outperforms traditional reinforcement learning models with fixed parameters, demonstrating that humans continuously adapt their strategies over time. These dynamically-estimated parameters reveal trajectories that systematically increase expected rewards, with updates significantly aligned with policy gradient ascent directions. Furthermore, this learning process operates across multiple timescales, with strategy parameters updating more slowly than behavioral choices, and update effectiveness correlates with local gradient strength in the reward landscape. Our work offers a generalizable approach for characterizing meta-learning trajectories, bridging theories of biological and artificial intelligence by providing a quantitative method for studying how adaptive behavior is optimized through experience.

Keywords Computational Modeling · Reinforcement Learning · Decision Making · Meta-Learning

1 Introduction

Intelligence is flexible and dynamic, shaped by evolutionary pressures to adapt to an ever-changing environment. Adaptation must shape not just one's behavior, but also the process of learning itself. Anderson (2013) articulated a general principle of rationality: "The cognitive system operates at all times to optimize the adaptation of the behavior of the organism." In line with this principle, Harlow's pioneering work (Harlow, 1949) and numerous subsequent studies have demonstrated that humans and animals improve performance by dynamically adjusting their learning strategies — i.e., they *learn to learn* (Behrens et al., 2007; Browning et al., 2015; Nassar et al., 2010, 2012; Piray and Daw, 2020; Soltani et al., 2006; Hattori et al., 2023). For example, in rapidly changing environments, individuals place more weight

[*]Co-first authors
[†]Co-senior authors.

on recent experiences by adopting a higher learning rate, whereas in stable contexts, they use a lower learning rate to guide decisions (Piray and Daw, 2021). This strategic flexibility extends from simple bandit tasks to complex real-world situations, yet it is not captured by traditional modeling frameworks that assume fixed parameters throughout learning. This discrepancy highlights a fundamental limit in our understanding of the biological principles of learning.

Recent advances in artificial intelligence (AI) offer a compelling clue to these principles. In AI, *learning to learn* has featured prominently in recent studies under the label "meta-learning", with algorithms spanning a range of tasks settings and performance objectives (Andrychowicz et al., 2016; Finn et al., 2017; Li et al., 2018; Ortega et al., 2019; Wang et al., 2017). Notably, AI models of meta-learning have offered valuable insights into human and animal adaptation, with task-optimized AI strategies accurately predicting biological learning in extensively-trained individuals (Wang et al., 2018). In large language models (LLMs), a similar process labeled "in-context learning" (ICL) enables models to rapidly adapt to new examples online without without model retraining (Lampinen et al., 2024), akin to human-like rapid generalization. This flexible, context-sensitive learning behavior is shown to resemble a form of online, gradient-based optimization in response to feedback (Ahn et al., 2023; Cheng et al., 2024; von Oswald et al., 2023a,b). Moreover, there is evidence that LLMs implicitly improve strategies (Coda-Forno et al., 2023) in ICL. These results suggest a striking possibility: that the refinement of learning strategies in humans can be similarly interpreted as online, gradient-based optimization.

In this study, we explore whether human strategy adaptation in value-based decision-making aligns with this `gradient-based meta-learning` principle. We investigate this within the reinforcement learning (RL) framework (Rescorla and Wagner, 1972; Sutton and Barto, 1998), where strategies for learning from reward feedback are characterized by a small set of parameters. To capture how these strategies evolve over time, we developed the `DynamicRL` framework, which employs neural networks to estimate RL parameters on a trial-by-trial or block-by-block basis (see also Xiong et al., 2025a).

Across four diverse value-based decision-making tasks, our findings provide evidence for the `gradient-based meta-learning` principle. We first show that the `DynamicRL` framework consistently outperforms traditional RL models with fixed parameters, indicating that participants dynamically adjusted their learning strategies during the task. These dynamically estimated RL parameter trajectories exhibit a systematic progression toward higher expected rewards, and, together with improved task performance, suggest that this adaptation reflects a functional refinement of learning strategies in real time. Critically, we find that the direction of parameter updates aligns with the policy gradient—the gradient of expected reward with respect to the RL parameters. Our results support a two-level learning process in which individuals not only learn action values but also gradually refine their learning strategies by tuning RL parameters in a gradient-guided manner. Finally, we show that this meta-learning process operates across multiple timescales, with internal strategy parameters adapting more slowly than behavioral choices, further supporting the biological plausibility of our account.

Taken together, these findings motivate a broader hypothesis: that human strategy adaptation in decision-making may reflect a form of gradient-based meta-learning. This view offers a unified computational account of how learning strategies themselves evolve in response to environmental feedback, bridging biological and artificial systems via shared adaptive principles.

2 Results

2.1 Gradient-based meta-learning

In psychology and neuroscience, a common paradigm for studying reinforcement learning behavior is via multi-armed bandit tasks, where participants choose between several options with unknown expected rewards. The expected rewards of different options may drift over time, reflecting environmental volatility. Additionally, the reward delivered in each sample is often stochastic, introducing noise into the reward feedback. Below, we elucidate the two levels in our gradient-based meta-learning hypothesis.

First-level: Learning action values

The first level of our model addresses how agents learn the value of their actions directly from experience. To solve a bandit task, an agent must learn the expected reward, or value, of each option from noisy feedback (Fig. 1a). This is formalized using a standard Rescorla-Wagner learning rule, where the value of a chosen action i at time t, $Q_{i,t}$, is updated based on the reward prediction error:

$$Q_{i,t} = Q_{i,t-1} + \alpha \left(r_{t-1} - Q_{i,t-1} \right) \ , \tag{1}$$

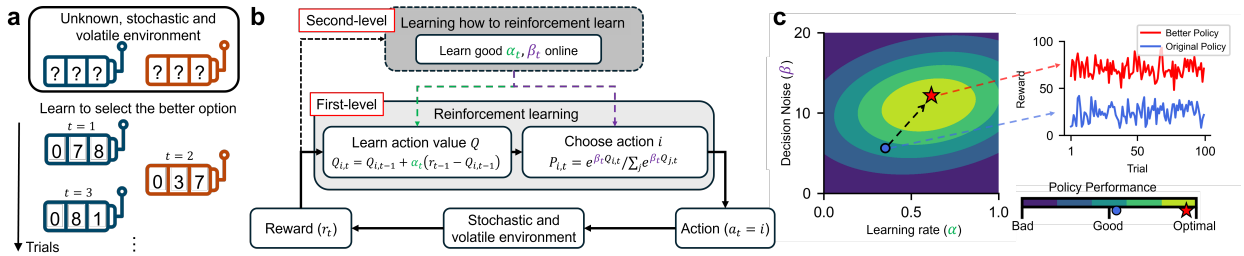

Figure 1: **Illustration of learning to reinforcement learn in a two-armed bandit task. a**, Example trial of a two-armed bandit task: participants interact with two slot machines whose reward probabilities are initially unknown. Through repeated choices, they learn to estimate the value of each machine. **b**, Diagram of our hypothetic meta-RL process: At the first level, participants take actions, interact with the environment, and receive rewards according to a reinforcement learning (RL) model (solid-line components). At the second level, they learn to optimize their RL parameters (α_t, β_t) to improve long-term task performance. **c**, A conceptual "reward landscape" illustrates how different RL parameter settings yield different expected returns. The agent begins with suboptimal parameters (blue dot), evaluates reward signals, and updates parameters in a direction that improves performance. Over time, it converges toward better settings (red star). Top right inset: hypothetical reward comparison between the original and optimized policy, demonstrating the advantage of better RL parameter selection.

where r_{t-1} is the reward received on the previous trial. Here, the learning rate α governs how much weight is given to the most recent outcome.

To translate these learned values into choices, a softmax function determines the probability of selecting each action, balancing exploitation of the highest-value option with exploration of alternatives:

$$P_{i,t} = \frac{e^{\beta Q_{i,t}}}{\sum_j e^{\beta Q_{j,t}}}, \tag{2}$$

The inverse temperature β controls the degree of decision stochasticity. Together, α and β define the agent's learning strategy. In the following sections, we focus on this two-parameter model (α and β) for its simplicity and normality.

Second-level: Learning to learn by adapting RL parameters

We posit that humans engage in a second-level learning process, dynamically adapting their learning strategies by tuning the parameters α and β to improve performance based on task feedback. The choice of these RL parameters is critical, as their suitability depends on environmental statistics (e.g., volatility; Piray and Daw, 2021). We conceptualize this as a "reward landscape" (Fig. 1c), in which each combination of α and β corresponds to a specific expected reward.

Most cognitive modeling studies adopt fixed parameters for each subject, with an implicit assumption that these parameters reflect the agents' stable traits. However, it remains unclear how rational agents (and human participants) select appropriate parameter values without prior knowledge of the task's environmental statistics.

Since the optimal parameters cannot be known in advance, we hypothesize that humans adapt their learning strategies through a second-level online learning process, i.e., by adjusting α and β as a form of learning-to-learn. Reward feedback serves as a signal that enables the agent to sense the local slope of the reward landscape and adjust their parameters in a direction that increases expected reward. Our central hypothesis is that this continuous refinement of human learning strategies resembles an online, gradient-based meta-learning process.

2.2 Humans adjust RL parameters throughout the experiment

Estimating individual dynamic RL parameters via the DynamicRL framework

To test whether humans dynamically adjust their RL strategies, we analyzed behavior across four diverse multi-armed bandit tasks. These datasets were chosen to cover a range of environmental conditions, including differences in the number of choices, reward volatility, and task length (see Sec. 4.1 for details). Specifically, we focused on two widely studied categories of bandit tasks: stationary and restless. Stationary bandit tasks assume stable reward distributions, whereas restless bandit tasks involve dynamically shifting rewards that demand continual adaptation. These paradigms

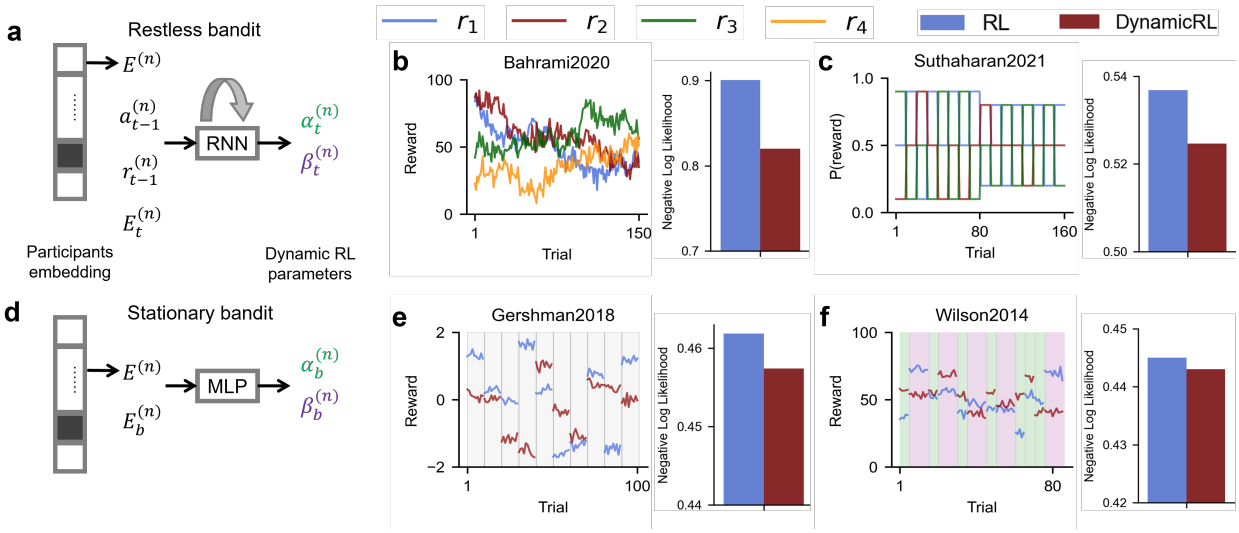

Figure 2: **Estimating dynamic RL parameters in four different bandit tasks. a, d**, Architecture of the `DynamicRL` model. **a**, A gated recurrent unit (GRU) is used to estimate RL parameters at the trial level. **d**, A two-layer perceptron is used to estimate RL parameters at the block level. **b, c, e, f**, Bandit tasks. **Left panels:** Reward structures for the four datasets. `Bahrami2020`: a 4-armed restless bandit task where participants choose among actions A_1–A_4, each with rewards ranging from 0 to 100 that evolve over time via an Ornstein–Uhlenbeck process centered at 50. `Suthaharan2021`: a 3-armed probabilistic reversal learning task where participants choose among actions A_1–A_3, each associated with binary rewards and time-varying reward probabilities. `Gershman2018`: a 2-armed bandit task where the expected rewards for the two options are sampled from a hierarchical Gaussian distribution. `Wilson2014`: a 2-armed bandit task where the difference in expected rewards between arms is sampled from a predefined set. The first 4 choices in each block are forced, with block horizons of either 1 (green shading) or 6 (purple shading). For `Gershman2018` and `Wilson2014`, the reward standard deviation is divided by 4 in this figure for visualization purposes. **Right panels:** Performance of the `DynamicRL` model on unseen test data (lower values indicate better fitting performance) exceeds that of classic RL models with fixed parameters.

provide insight into biological value-based learning under stable versus volatile conditions. For the restless tasks, we use `Bahrami2020` (Bahrami and Navajas, 2020) and `Suthaharan2021` (Suthaharan et al., 2021). `Bahrami2020` involves a 4-armed bandit task in which each option's mean reward drifts over time (Fig. 2b). `Suthaharan2021` features a 3-armed probabilistic reversal learning task, where choice probabilities change abruptly and unpredictably. For the stationary tasks, we use `Gershman2018` (Gershman, 2018) and `Wilson2014` (Wilson et al., 2014). `Gershman2018` includes a 2-armed bandit task with Gaussian-distributed rewards whose mean is resampled each block. `Wilson2014` presents a 2-armed bandit task with varying block lengths and differing numbers of forced-choice trials to manipulate uncertainty.

A key challenge in testing our hypothesis is that estimating learning parameters on a trial-by-trial basis is difficult with traditional methods because each trial provides only one data point per participant. To overcome this, we developed the `DynamicRL` framework (see Sec. 4.2; see also Xiong et al., 2025a), which uses a neural network to robustly estimate the parameters of a learning algorithm dynamically over the course of an experiment. By training a single model across all participants, DynamicRL leverages pooled data to generate stable and reliable parameter estimates (see also Xiong et al., 2025b). Note that, in principle, one could estimate parameters using a sliding window without neural networks; however, such approaches typically perform poorly and prone to overfit due to the limited data available in each window (see details in Sec. A.2).

We tailored the architecture of DynamicRL to the temporal structure of the tasks. In the restless bandit tasks (`Bahrami2020`, `Suthaharan2021`), the expected value of each option changes over time within each block. We therefore define the dynamic RL strategy using trial-level parameters α_t and β_t. To estimate these parameters, we use a gated recurrent unit (GRU) that takes as input the previous action $a_{t-1}^{(n)}$, previous reward $r_{t-1}^{(n)}$, participant information $E^{(n)}$, and trial level task information $E_t^{(n)}$, estimating RL parameters for each participant trial-by-trial (Fig. 2a).

In the stationary bandit tasks (`Gershman2018`, `Wilson2014`), the expected value of each option remains constant within a block but varies across blocks (Fig. 2e,f). Accordingly, we define the RL strategy at the block level, with parameters $\alpha_b^{(n)}$ and $\beta_b^{(n)}$ estimated using only block-level task information $E_t^{(n)}$ via a two-layer perceptron (Fig. 2d). This formulation allows us to examine RL strategy adaptation on a slower, block-level timescale. Consequently, trial-by-trial inputs such as actions and rewards are not used in the estimation process (see Sec. 4.2.1 for rationale).

DynamicRL better explains human behavior and reveals RL strategy adaptation

We first examined whether RL models with dynamic parameters could better explain human behavior than a traditional RL model with fixed parameters. A formal model comparison on held-out test data revealed that `DynamicRL`, which achieved lower test negative log-likelihood, consistently outperformed the static RL model across all four datasets (Fig. 2b,c,e,f). We also conducted a model recovery analysis using synthetic data generated from RL agents with predefined parameter trajectories—constant, increasing, or decreasing—for both α and β (see Sec. A.1 for full results). `DynamicRL` accurately recovered the qualitative structure and linear trends of the ground-truth parameters in most cases, demonstrating robustness across a range of dynamic regimes. The superior predictive performance of `DynamicRL` provides evidence that incorporating temporal dynamics in RL parameters yields a more accurate account of human behavior. These results support the view that humans do not rely on a fixed strategy, but instead continuously adapt their learning parameters throughout the experiment.

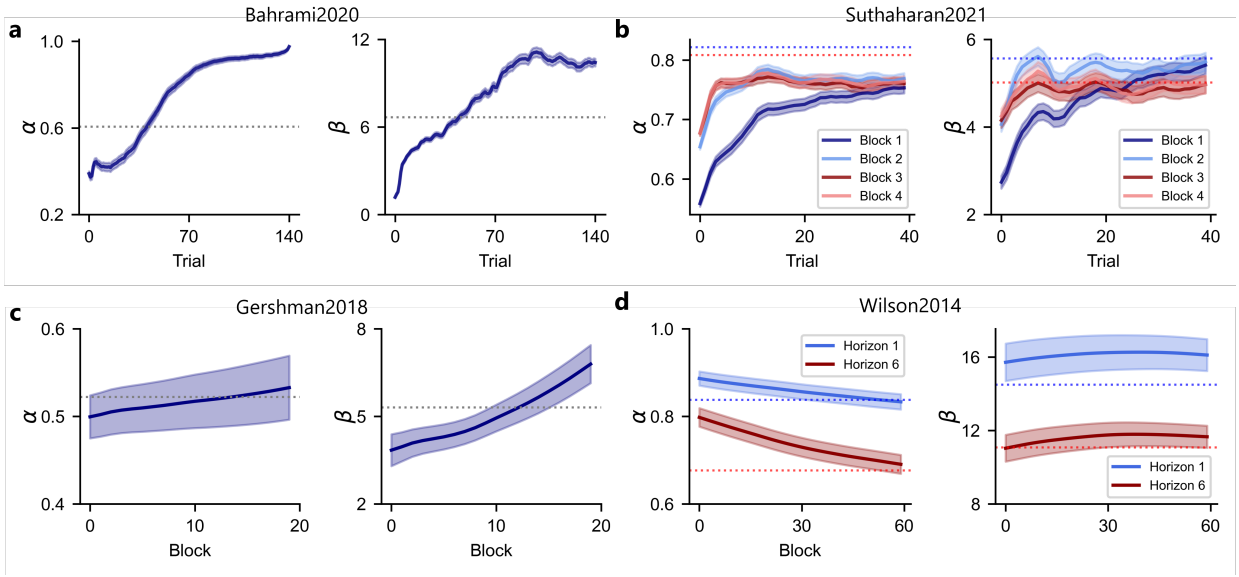

Figure 3: **DynamicRL reveals human RL strategy adaptation. a–d**, Solid lines show RL parameters (α and β) estimated from `DynamicRL`, averaged across participants over the course of the experiment. Shaded areas indicate 95% confidence intervals. Dotted lines represent RL parameters estimated from classical models with fixed parameters.

We next examined the trajectories of these RL parameter changes over time. The `DynamicRL` framework revealed clear, systematic patterns of adaptation in both α and β (Fig. 3). For example, in the restless bandit tasks, both parameters tended to increase across trials in `Bahrami2020` (Fig. 3a) and across sessions in `Suthaharan2021` (Fig. 3b). Notably, in blocks 2–4 of `Suthaharan2021`, RL parameters plateau more quickly, suggesting faster adaptation as participants become familiar with the task structure. In the stationary bandit tasks, `Gershman2018` and `Wilson2014`, we also observe systematic changes in α and β across blocks (Fig. 3c,d), further supporting the notion that participants adjust their RL strategies over time. Together, these findings confirm that humans continuously adjust their reinforcement learning strategies at both trial and block levels.

2.3 Humans improve their strategies by adjusting RL parameters online

Having established that learning strategies are dynamic, how should we interpret the observed changes in RL parameters? Our *gradient-based meta-learning* hypothesis predicts that participants should perform better in the later stages of an experiment, as changes in their RL parameters functionally lead to improved performance. We thus conducted a

model-agnostic analysis, comparing within-subject performance in the first versus second half of each experiment, while controlling for task conditions using a mixed effects model (see Sec. 4.3 for details). We found significant improvements in second-half performance relative to the first half in `Bahrami2020`, $t(950) = 18.14, p < .001$; `Suthaharan2021`, $z = 5.37, p < .001$; `Gershman2018`, $t(43) = 2.34, p = .024$; and `Wilson2014`, $z = 4.736, p < .001$. These results confirm that participants' decision-making improves systematically over time.

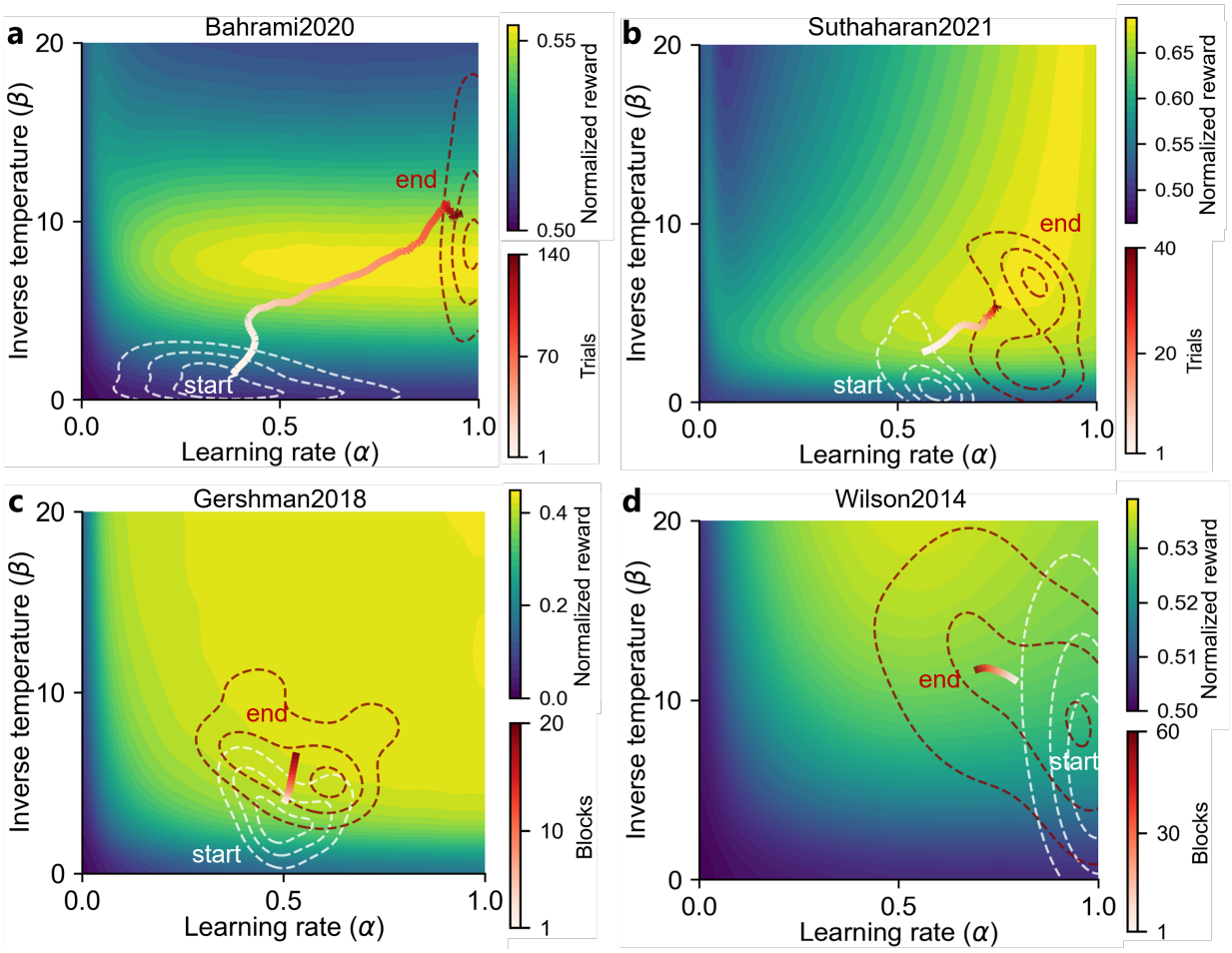

Figure 4: **Trajectories of RL parameters over the course of the experiment.** Each panel shows the 2D RL parameter space defined by learning rate (α) and inverse temperature (β). The colored background and solid contour lines depict the reward landscape computed from RL simulations, with color intensity and contour density indicating the expected reward associated with each parameter combination. Dashed contours represent kernel density estimates of participants' RL parameters at the start (white) and end (dark red) of the experiment. The white-to-red solid lines traces the average temporal evolution of participants' estimated RL parameters.

To understand how this behavioral improvement was achieved, we visualized the evolution of the estimated RL parameters (α, β) on a "reward landscape" that shows the expected reward for any given strategy (Fig. 4). This landscape was generated by simulating an RL agent with different combinations of of learning rate (α) and inverse temperature (β) to map the effectiveness of every possible strategy (see Sec. 4.4 for details). When we overlaid participants' parameter trajectories onto this landscape, we observed a clear and consistent pattern across all four datasets: participants' averaged RL strategies evolved from lower-reward regions to higher-reward regions over the course of the experiment (white-to-red lines in Fig. 4). This was further confirmed by kernel density plots showing that the distribution of RL parameters shifted from suboptimal starting points toward more reward-maximizing configurations by the end of the experiment (dashed contours in Fig. 4). This analysis demonstrates that the observed improvement in task performance is directly linked to a systematic adaptation of RL parameters toward more optimal strategies.

2.4 Comparing RL parameter change with policy gradient ascent

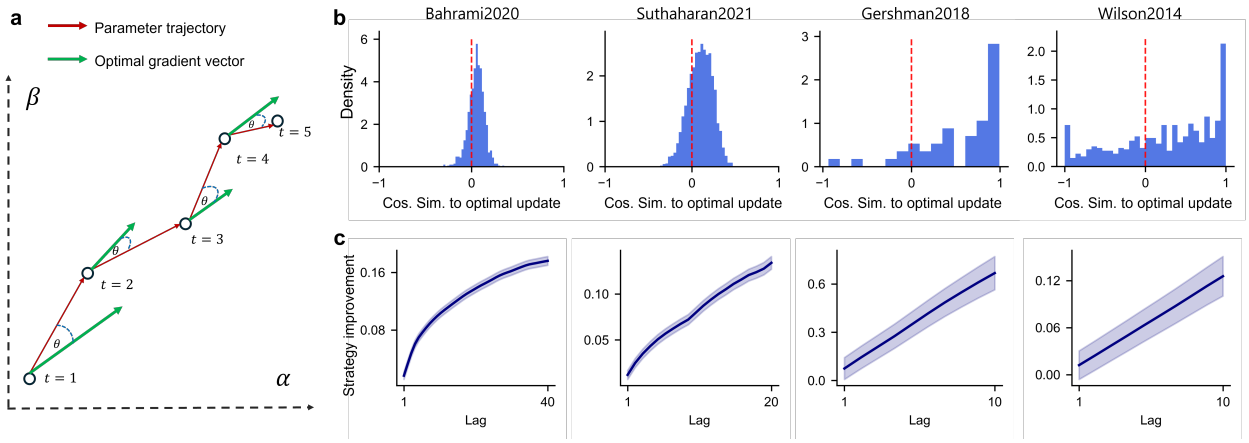

Figure 5: **Optimality of RL parameter updates.** **a**, Schematic illustrating cosine similarity between human RL parameter updates and the optimal gradient direction. Each circle represents an update step; the red arrow indicates the actual RL parameter update, and the green arrow shows the optimal update according to gradient ascent. θ indicates difference between actual RL parameter update and optimal gradient ascent direction. **b**, Distribution of cosine similarity values across participants, showing population-level alignment with the optimal gradient. **c**, Effect size of alignment increases with larger window sizes, indicating consistent improvement in RL strategies.

Having shown that humans adapt their learning strategies to improve performance, we next investigated the central claim of our hypothesis: that this adaptation process is analogous to policy gradient ascent. Visual inspection of the parameter trajectories suggests they follow the reward gradient (Fig. 4), but to quantify this, we developed two complementary metrics.

First, we measured the alignment in parameter space between RL parameter updates and the optimal gradient direction using cosine similarity (Fig. 5a). To ensure a stable comparison, we calculated the optimal direction from a simulated "optimal gradient vector" rather than from noisy empirical data (see Sec. 4.5.1). Because the optimal gradient is statistical, policy gradient ascent typically requires averaging over many repeated episodes in reality. In contrast, each human participant completed only a single episode per task, producing just one action–outcome sequence. This limited data naturally results in high variance in inferred learning signals, so we do not expect perfect alignment with the optimal gradient at the individual level. However, at the population level, we hypothesize that RL parameter updates should exhibit positive alignment with the optimal gradient. As predicted, we found a statistically significant positive alignment at the population level across all four tasks, with the distribution of cosine similarity values skewed significantly above zero (Fig. 5b).

While directional alignment is informative, cosine similarity can be sensitive to differences in the numerical scale of the learning parameters (α and β). To provide a more direct and scale-independent test, we introduced a second metric, "strategy improvement," which measures the change in expected reward produced by a parameter update in reward space. This metric quantifies whether the current strategy yields a higher expected reward than the strategy from the previous trial or block (see Sec. 4.5.2 for details). This analysis revealed that strategy improvement was significantly positive at every step across all four datasets: `Bahrami2020` ($t(143{,}784) = 14.15, p < .001$), `Suthaharan2021` aggregated across all blocks ($t(157{,}871) = 4.48, p < .001$), `Gershman2018` ($t(835) = 10.72, p < .001$), and `Wilson2014` for both horizon conditions (Horizon 1: $t(14{,}615) = 32.01, p < .001$; Horizon 6: $t(14{,}615) = 16.63, p < .001$). p-values were obtained via permutation tests in which the temporal order of RL parameters was shuffled 10,000 times per participant to construct a null distribution. These results demonstrate that participants' parameter updates systematically guided them toward more rewarding strategies.

To ensure these single-step improvements reflect a consistent trend rather than random fluctuations, we assessed the strength of *strategy improvement* over longer timescales using a sliding-window analysis. This approach aggregates updates to smooth out single-step noise. We found that the effect size of the improvement became progressively stronger as the window size increased across all four tasks (Fig. 5c). This indicates that while individual updates may be noisy, they reflect a robust and consistent trajectory of improvement over time. Taken together, the positive alignment in parameter space and the consistent *strategy improvement* in reward space provide converging evidence that human

strategy adaptation is functionally aligned with the gradient of expected reward, as predicted by our gradient-based meta-learning hypothesis.

2.5 Gradient structure of the reward landscape correlates with human meta-RL performance

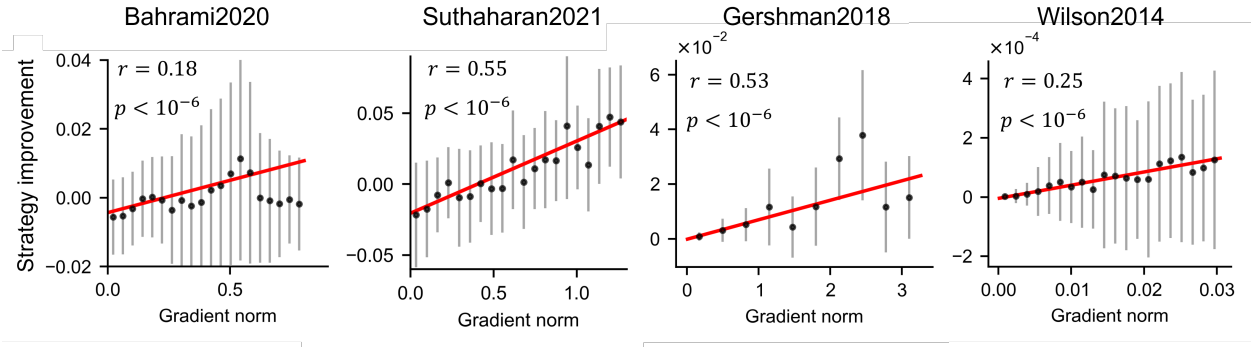

Figure 6: **Correlation between gradient norm and meta-RL strategy improvement.** For each dataset, we computed the empirical correlation between the magnitude of the local reward gradient and the optimality of meta-RL updates. The scatter plot shows binned data with mean (± 1 SD) of strategy improvement scores (black dots with error bars) across gradient norm. The red line indicates the linear fit to the original (unbinned) data.

A key prediction of gradient-based non-convex optimization is that the magnitude of improvement should scale with the magnitude of the local gradient; updates should be more effective where the reward landscape is steeper. To test this, we examined the relationship between the norm of the local reward gradient and the degree of strategy improvement for each step in a participant's trajectory (see Section 4.6 for details). Across all four datasets, we found a significant positive correlation between the gradient norm and the magnitude of strategy improvement (Fig. 6). That is, participants made more effective updates to their strategy when the local reward landscape provided a stronger directional signal. These correlations were highly robust, far exceeding null distributions generated by permutation tests (Fig. A.4). This "dose-dependent" relationship between gradient strength and update effectiveness provides strong evidence that human meta-learning is not just directionally aligned with the reward gradient, but is also sensitive to its local structure, further supporting a policy gradient-like framework.

2.6 Temporal structure of human meta-RL

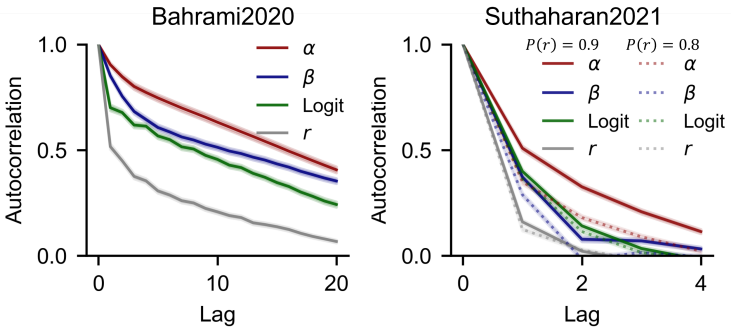

Figure 7: **Timescale of human meta-RL processes.** Autocorrelation analysis of learning rate α, inverse temperature β, logits and reward r reveals distinct temporal dynamics. For `Suthaharan2021`, solid and dotted lines denote two experimental conditions based on the maximum reward probability ($P(r)$) of the options.

A hallmark of biological learning is the coordination of learning processes that operate on distinct timescales in response to environmental dynamics (Binz et al., 2022; Hattori et al., 2023; Wang et al., 2018). Our two-level learning framework predicts that the meta-learning process of refining a strategy (how to learn) should operate more slowly than the process of learning action values (what to choose). To test this prediction, we performed a temporal autocorrelation analysis

on trial-level data from the restless bandit tasks (Bahrami2020 and Suthaharan2021), as block-level estimation of RL parameters would naturally reflect slower timescales. We predicted that the RL parameters (α and β) would be more stable over time (i.e., exhibit slower autocorrelation decay) than trial-by-trial variables such as environmental rewards (reflecting environmental structure) and action probabilities (logits, reflecting choice probabilities). Our results confirmed a clear temporal hierarchy (Fig. 7). In both datasets, the learning rate (α) was the most stable parameter, evolving more slowly than the inverse temperature (β), which itself evolved more slowly than the choice-related logits. These distinct timescales (which emerged naturally from the behavioral data without being imposed by the model) support a hierarchical learning architecture.

We next examined whether these learning timescales are themselves adaptive. An effective agent should calibrate the speed of its strategy updates to the environmental dynamics. The timescales of α and β serve as components of an agent's internal model of environmental statistics, reflecting a belief about how frequently this model should be updated. Updating too quickly can lead to chasing noise, while updating too slowly can result in missing meaningful changes in the environment. We therefore tested whether individual differences in parameter timescales correlated with overall task performance (see Sec. 4.7 for details). In Bahrami2020, we found no significant correlation between the timescale of α and total reward ($r = 0.037$, $p = 0.256$), but observed a significant positive correlation for β ($r = 0.256$, $p < .001$). This suggests that flexible adaptation of decision noise or exploration, as governed by β, is particularly important in this task. In Suthaharan2021, we found modest but significant negative correlations between total reward and the timescales of both parameters: α ($r = -0.063$, $p = 0.004$) and β ($r = -0.312$, $p < .001$). These results indicate that participants with more adaptive (i.e., faster-changing) RL parameters tend to perform better, particularly when β adjusts more rapidly. This supports the role of β in tracking and responding to environmental volatility.

Finally, we tested whether participants actively adjust their adaptation speed in response to environmental statistics. In the Suthaharan2021 task, we compared parameter timescales between conditions with different maximum reward probabilities. We found that participants' strategy updates were significantly faster in the more informative, high-reward condition ($P_{\max}(r) = 0.9$) compared to the lower-reward condition. This was true for the timescales of both the learning rate α ($t(2023) = 27.965$, $p < .001$, Cohen's $d = 0.617$) and the inverse temperature β ($t(2023) = 16.049$, $p < .001$, Cohen's $d = 0.384$). This finding further supports that humans flexibly modulate the timescale of meta-learning itself, adapting not only their strategy but also the speed of their strategic adjustments in response to the environment's structure.

3 Discussion

This work addresses a fundamental gap in our understanding of cognitive flexibility: how do humans dynamically adapt their own learning strategies? We proposed and tested the *gradient-based meta-learning* hypothesis, positing that humans tune their learning strategies online in a process analogous to gradient ascent. Across four diverse decision-making tasks, we found three key lines of evidence supporting this claim. First, using our novel DynamicRL framework, we showed that models with dynamic learning parameters (α, β) better explain human behavior than traditional static models. Second, we demonstrated that this adaptation is functional, as the parameter trajectories systematically evolved toward more optimal, high-reward regions of the strategy space. Finally, and most critically, we found that the direction of these strategy updates was significantly aligned with the gradient of expected reward, providing strong evidence for a policy gradient-like learning mechanism.

These discoveries were enabled by DynamicRL, a novel framework that overcomes the static assumptions of conventional cognitive models. DynamicRL is designed to robustly track latent cognitive variables as they evolve over time, even when behavioral data are sparse and noisy. By doing so, it captures residual variance attributable to dynamic learning strategies, yielding improved behavioral fits compared to fixed-parameter RL models. DynamicRL offers a middle ground between traditional interpretable models and "black-box" RNNs; it retains the normative and well-established structure of an RL model while using a neural network to flexibly capture behavioral variability, thereby avoiding reliance on opaque latent representations of neural networks. This methodological advance supports more nuanced interpretations of behavior by directly modeling the temporal dynamics of strategy adaptation and is generalizable to other cognitive models with differentiable likelihoods.

The architecture of DynamicRL offers key advantages over existing methods for modeling dynamic processes. Unlike naive sliding-window estimators, it avoids overfitting by jointly estimating parameter dynamics across participants and conditions using shared representations. In contrast to task-specific hierarchical models that rely on strong priors over hierarchical generative processes (Mathys et al., 2011, 2014), and to amortized simulation-based inference approaches (Schumacher et al., 2023; Ger et al., 2024) that learn from synthetic data generated by predefined generative models, DynamicRL trains directly on human data without imposing such strong constraints on the generative process.

Our findings speak to a broader question: what computational principles unify biological and artificial adaptation? The gradient-aligned parameter updates we observed parallel recent findings in large language models, which exhibit rapid "in-context learning" during inference. Recent work shows that hierarchical meta-learning structures emerge within large language models to refine strategies over time (Coda-Forno et al., 2023). The convergence of these findings suggests that gradient-based meta-learning may reflect a general principle of adaptive intelligence: both biological and artificial systems may converge on similar computational mechanisms, improving behavior based on feedback gradients. This framework also resonates with longstanding multi-timescale learning theories in cognitive neuroscience (e.g., the complementary learning systems McClelland et al., 1995). Our results provide quantitative evidence for such dual-timescale learning, showing a fast process for learning action values supported by a slower meta-learning process that refines the learning strategy itself.

While our findings robustly support the gradient-based meta-learning hypothesis, several caveats remain. First, for the sake of interpretability and visualization, we used a simple two-parameter RL model. Although this choice reduces estimation variance, it also limits model expressivity; future work could apply `DynamicRL` to track more strategies. Second, the observed alignment with gradient ascent is statistically robust, but our claim is analogical rather than a direct assertion that the brain explicitly computes gradients. A comprehensive comparison with alternative processes—such as random search or other heuristic rules—is needed to fully clarify the computational basis of human meta-learning.

Finally, the `DynamicRL` framework enables powerful new applications in neuroscience and psychiatry. In cognitive neuroscience, these trial-level RL parameter estimates can serve as regressors in neural data analyses (e.g., functional magnetic resonance imaging) to help identify the specific brain regions involved in meta-strategy encoding. In computational psychiatry, the framework offers a quantitative tool to study how abnormal meta-learning dynamics—such as excessive or insufficient strategy plasticity—may underlie conditions like anxiety, depression, or paranoia, providing a new way to characterize individual differences.

In sum, this work demonstrates that humans adapt their learning strategies dynamically, in a manner aligned with gradient ascent. By introducing the `DynamicRL` framework and validating the gradient-based meta-learning hypothesis across multiple tasks, we provide a novel and generalizable account of how humans learn to learn. This framework advances our understanding of adaptive intelligence and lays the computational groundwork for linking cognitive models with artificial systems.

4 Methods

4.1 Datasets

We analyze four human behavioral datasets collected from different bandit tasks with varying reward structures, including differences in environmental volatility, stochasticity, and horizon length. These tasks fall into two categories: restless bandit and stationary bandit. To facilitate model fitting and reward landscape simulations, rewards were normalized using $r \leftarrow \frac{r}{C}$, with task-specific constants C chosen to ensure comparable reward ranges where applicable."

`Bahrami2020` (Four-armed restless bandit task): In this task (Bahrami and Navajas, 2020), participants choose among four options whose mean rewards evolve stochastically over time according to an Ornstein–Uhlenbeck process centered at $\mu = 50$, with a decay rate of $\gamma = 0.9836$ and process noise $\sigma_{\mathrm{OU}} = 2.8$. At each time step, rewards are sampled as $r \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma = 4$. Each participant is required to complete $B = 1$ block of $T = 150$ trials. Due to timeouts or delayed responses, some participants completed fewer trials. 14 participants are removed from our analysis due to low performance (averaged raw reward lower than 43 points, which approximately corresponds to 1.5 standard deviations below the mean) and too many missing trials (fewer than 80 finished trials), resulting in $T_{\mathrm{total}} = 139{,}816$ trials from $N = 965$ participants. Rewards were normalized using $r \leftarrow \frac{r}{100}$, yielding a baseline mean of 0.5. In our analysis, we focused on the first 140 trials, as these were completed by the majority of participants, ensuring consistent data for comparison.

`Suthaharan2021` (Three-armed probabilistic reversal learning task): This task (Suthaharan et al., 2021) consists of $B = 4$ blocks of $T = 40$ trials each. Participants choose among three options with probabilistic rewards. In the first two blocks, reward probabilities are fixed at [0.9, 0.5, 0.1]; in the final two blocks, they change to [0.8, 0.5, 0.2]. After ten consecutive selections of the highest-reward option, reward contingencies reverse, enabling the study of behavioral flexibility in volatile environments. The dataset includes $N = 1{,}012$ participants and $T_{\mathrm{total}} = 173{,}120$ trials. Because the baseline average reward is already 0.5, no normalization was applied.

`Gershman2018` (Two-armed fixed-horizon bandit task): This dataset is based on condition 2 of Gershman (2018), which features a symmetric task design. Condition 1, which is asymmetric because one option maintains a constant value throughout the experiment, was excluded for modeling simplicity. In each block, the expected rewards for the two arms are drawn from $\mu_b \sim \mathcal{N}(\mu, 10^2)$, and actual rewards are sampled as $r \sim \mathcal{N}(\mu_b, 10)$. Participants complete

$B = 20$ blocks of $T = 10$ trials each. The dataset includes $N = 44$ participants and $T_{\text{total}} = 8{,}800$ trials. Rewards were normalized via $r \leftarrow \frac{r}{10}$, setting the baseline mean to 0.

`Wilson2014` (Two-armed horizon task): This task investigates exploration under uncertainty through a forced-choice phase followed by a free-choice phase (Wilson et al., 2014). In each block, participants choose between two arms: one arm's mean reward is drawn from {40, 60}, and the other arm's mean differs by an amount sampled from {4, 8, 12, 20, 30}. Rewards are sampled from a Gaussian distribution $r \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma = 8$. During the forced-choice phase (first 4 trials), actions are pre-assigned to create either symmetric (2 samples per arm) or asymmetric (3 vs. 1) information conditions. The subsequent free-choice phase varies in length, with horizons of either 1 or 6 trials. We aggregate data from $N = 643$ participants across six experiments (Wilson et al., 2014; Feng et al., 2021; Somerville et al., 2017; Waltz et al., 2020; Sadeghiyeh et al., 2020; Zajkowski et al., 2017). Of these, 34 participants were excluded if they met at least three of these criteria: low performance (average score below 52 raw points), highly asymmetric free-choice behavior (choosing the same option in more than 85% or exhibiting a choice autocorrelation greater than 0.85 or less than 0.15), and excessive fast responses (more than 20% of trials with reaction times under 30 ms). Due to differences in experimental design, the number of blocks (B) completed varies across participants, ranging from 300 to 1,200 free-choice trials, yielding a total of $T_{\text{total}} = 754{,}890$ trials. Rewards were normalized using $r \leftarrow \frac{r}{100}$, resulting in a baseline mean of 0.5. Although participants completed different numbers of blocks and encountered horizon conditions (1 or 6) in varying sequences, our `DynamicRL` model receives only subject identity, block number, and horizon condition as input. We restrict our analysis to blocks 0 through 60 for both horizon conditions to ensure comparability across participants. These early blocks are representative, as they were completed by the majority of participants.

4.2 DynamicRL framework

4.2.1 Modeling choices for restless vs. stationary bandit tasks

4.2.2 Model structures

In restless tasks, such as `Bahrami2020` and `Suthaharan2021`, the expected value of each option varies across trials within a block, requiring continuous adaptation. In this setting, we consider trial-level parameters α_t and β_t that define the RL strategy, allowing us to track strategy evolution over trials. We use a gated recurrent unit (GRU) to estimate these RL parameters on a trial-by-trial basis. The model takes as input the previous action $a_{t-1}^{(n)}$, previous reward $r_{t-1}^{(n)}$, a learned participant embedding $E^{(n)}$ (capturing individual differences), and trial-level contextual information $E_t^{(n)}$. For `Bahrami2020`, $E_t^{(n)}$ consists of the current trial number. For `Suthaharan2021`, $E_t^{(n)}$ includes the current trial number, block number, and reward probability condition. The GRU updates its hidden state h_t, which is passed to a linear readout $g(\cdot)$ to generate the RL parameters $\alpha_t^{(n)}$ and $\beta_t^{(n)}$ for participant n at trial t:

$$h_t = \text{GRU}(a_{t-1}^{(n)}, r_{t-1}^{(n)}, E^{(n)}, E_t^{(n)}, h_{t-1}), \quad \alpha_t^{(n)}, \beta_t^{(n)} = g(h_t) \tag{3}$$

Different from restless tasks, stationary tasks (`Gershman2018` and `Wilson2014`) assume that expected rewards are stable within each block, but vary across blocks. To examine how participants' RL parameters adapt to task statistics that vary across blocks, we consider block-level parameters α_b and β_b, providing a parsimonious description of the strategy within a block. Under this formulation, the model inputs are limited to block-level information (e.g., block number, horizon condition), excluding trial-level information (e.g., actions and rewards). Since this block-level parameter estimation does not require processing temporal sequences of actions and rewards, we use a two-layer feedforward network (multilayer perceptron; MLP) to estimate RL parameters at the block level. The model receives the participant embedding $E^{(n)}$ and block-level contextual information $E_b^{(n)}$. For `Gershman2018`, $E_b^{(n)}$ contains only the current block number. For `Wilson2014`, it includes the block number and the horizon condition. The MLP outputs the RL parameters $\alpha_b^{(n)}$ and $\beta_b^{(n)}$ for participant n at block b:

$$\alpha_b^{(n)}, \beta_b^{(n)} = \text{MLP}(E^{(n)}, E_b^{(n)}) \tag{4}$$

4.2.3 Model fitting

To train our `DynamicRL` model, we use the negative log-likelihood (NLL) loss to maximize the likelihood of observed actions under the model's predicted action probabilities. For a dataset with N participants and T trials per participant, the mean NLL loss is defined as:

$$\mathcal{L}_{\text{NLL}} = -\frac{1}{N \cdot T} \sum_{n=1}^{N} \sum_{t=1}^{T} \log P_{\text{RL model}}(a_{i,t}^n | Q_{i,t}^n, \alpha_t^n, \beta_t^n), \tag{5}$$

where $P_{\text{RL model}}(a_{i,t}^n)$ is the probability of choosing action i at trial t for participant n, as defined in Equation (2). The model parameters α_t^n and β_t^n are estimated for each trial t in restless bandit tasks, or estimated for each block b in stationary bandit tasks.

To mitigate overfitting and ensure fair comparison across optimization methods, we employed a nested cross-validation protocol (Ji-An et al., 2025). Specifically, 70% of the data was used for training, 12.5% for validation (for early stopping and hyperparameter tuning), and 17.5% for testing. The outer loop used eight folds, and the inner loop used five.

Due to differences in task design across datasets, we tailored the cross-validation strategy accordingly. For `Gershman2018`, data were partitioned by block. For `Wilson2014`, data were segmented into non-overlapping 40-trial chunks. To preserve structural consistency, padding was introduced when needed, and padded trials were excluded from loss computation. For `Bahrami2020` and `Suthaharan2021`, trials were split into non-overlapping sets using distinct masks for training, validation, and testing (i.e., an interspersed split protocol Ji-An et al., 2025; Hans et al., 2024).

To improve robustness during training, we added Gaussian noise $\mathcal{N}(0, 0.01)$ to reward values. Early stopping was applied if validation performance did not improve for 20 consecutive epochs. We used the AdamW optimizer with default settings, a batch size of 64, and a participant embedding size of 20. Other hyperparameters were selected via grid search over the following values: learning rate in {0.005, 0.01, 0.02, 0.03}, weight decay in {0.01, 0.1}, dropout in {0, 0.1, 0.2}, and hidden layer size $h \in \{50, 100\}$. To constrain RL parameter estimates, we applied a sigmoid activation to ensure $\alpha \in [0, 1]$, and a ReLU activation to enforce $\beta \geq 0$. The best model was selected based on the lowest average NLL on the test data across all folds.

4.3 Comparing early and late task performance

To assess whether participants improved their decision-making performance over time, we examined a metric termed *advantage*, defined as the difference between the reward obtained by a participant and their average reward. To ensure comparability across datasets with differing reward structures, advantage scores were normalized: for `Gershman2018`, the baseline was set to 0; for the remaining datasets, it was set to 0.5. This normalization was intended to reduce cross-subject variance arising from heterogeneous reward trajectories.

For datasets without distinct experimental conditions, we conducted paired t-tests to compare early versus late task performance. For datasets with within-subject manipulations, we employed linear mixed-effects models to account for repeated measures. Specifically, for `Suthaharan2021`, we tested whether performance in the final 20 trials exceeded that in the first 20 trials using the following model:

$$\texttt{advantage} \sim \texttt{early_late_trial} \times \texttt{reward_probability} + (1|\texttt{subject}),$$

implemented using `statsmodels.MixedLM` (Seabold and Perktold, 2010). For `Wilson2014`, we examined whether participants performed better in later blocks with:

$$\texttt{advantage} \sim \texttt{early_late_block} \times \texttt{horizon} + (1|\texttt{subject}).$$

In `Bahrami2020` and `Wilson2014`, not all participants completed enough trials or blocks. To ensure the robustness and interpretability of the model estimates, we included only task segments with sufficient data coverage across participants.

4.4 Reward landscape

To understand the task performance of different RL parameters under each task, we constructed reward landscapes by simulating an RL agent with varying combinations of RL parameters. For each task, we systematically varied the learning rate α from 0 to 1 in increments of 0.025. The inverse temperature parameter β was varied across a broader range (0 to 60) using non-linear increments in three regions: 0 to 16 in steps of 0.5, 16 to 32 in steps of 1, and 32 to 60 in steps of 3, enabling finer resolution in regions of high sensitivity to β. This process yields a total of 1,302 parameter combinations. For each combination, we simulated behavior over 200 blocks of 120 trials and recorded the resulting average reward R. We then plotted reward landscapes R as a function of (α, β).

To ensure smooth visualization, we applied cubic interpolation to produce a dense grid with 100 points for α (from 0 to 1) and 200 points for β (from 0 to 60). To reduce high-frequency noise that arises from finite simulations, a two-dimensional Gaussian filter with a standard deviation of 3 was applied to the interpolated reward landscapes.

For tasks with multiple experimental conditions, we display representative conditions. Specifically, for `Suthaharan2021`, we show the reward landscape for the first two blocks under the [0.9, 0.5, 0.1] reward probability condition; the landscape under the [0.8, 0.5, 0.2] condition exhibits similar characteristics. For `Wilson2014`, we present results for the horizon-6 condition, as the horizon-1 condition produces comparable reward landscapes.

4.5 Alignment of human RL parameter updates with the reward gradient

The alignment of human RL parameter updates with the reward gradient was calculated at the trial level for `Bahrami2020` and `Suthaharan2021` and at the block level for `Gershman2018` and `Wilson2014`.

4.5.1 Cosine similarity with the reward gradient in parameter space

We first computed the cosine similarity between participants' RL parameter updates and the corresponding gradient direction on the reward landscape $R(\alpha, \beta)$. Specifically, the gradient of the reward surface was calculated using finite differences:

$$\frac{\partial R}{\partial \alpha} \approx \frac{R(\alpha + \Delta\alpha, \beta) - R(\alpha, \beta)}{\Delta\alpha}, \quad \frac{\partial R}{\partial \beta} \approx \frac{R(\alpha, \beta + \Delta\beta) - R(\alpha, \beta)}{\Delta\beta}.$$

This yields a gradient vector field $\mathbf{G}(\alpha, \beta) = (\frac{\partial R}{\partial \alpha}, \frac{\partial R}{\partial \beta})$ that estimates the direction in the parameter space along which the expected obtained reward is maximally increased.

For each participant s, we denote their RL parameters at time t (either a trial or a block) as $\boldsymbol{\theta}_t^{(s)} = (\alpha_t^{(s)}, \beta_t^{(s)})$. The parameter update vector is:

$$\Delta\boldsymbol{\theta}_t^{(s)} = \boldsymbol{\theta}_{t+1}^{(s)} - \boldsymbol{\theta}_t^{(s)},$$

and we evaluate the reward gradient $\mathbf{g}_t^{(s)}$ at the midpoint of each segment:

$$\mathbf{c}_t^{(s)} = \frac{1}{2}(\boldsymbol{\theta}_t^{(s)} + \boldsymbol{\theta}_{t+1}^{(s)}), \quad \mathbf{g}_t^{(s)} = \mathbf{G}(\mathbf{c}_t^{(s)}).$$

We then compute their cosine similarity:

$$\rho_t^{(s)} = \frac{\Delta\boldsymbol{\theta}_t^{(s)}}{\|\Delta\boldsymbol{\theta}_t^{(s)}\|} \cdot \frac{\mathbf{g}_t^{(s)}}{\|\mathbf{g}_t^{(s)}\|}.$$

The overall alignment score for participant s is defined as the mean cosine similarity across all valid segments:

$$\rho_{\text{mean}}^{(s)} = \frac{1}{T^{(s)} - 1} \sum_{t=1}^{T^{(s)}-1} \rho_t^{(s)}.$$

This metric quantifies the degree to which human RL parameter updates follow the optimal gradient direction. A value near 1 indicates strong alignment, while values near 0 or negative suggest deviation from the reward-maximizing direction. Perfect alignment is not expected due to the stochasticity inherent in online learning, but positive alignment at the group level provides evidence for gradient-like updates.

4.5.2 Optimality of meta-learning update in reward space

To provide a measure of alignment that is invariant to the scale of parameters, we also evaluated whether participants' RL parameter updates shift parameters toward regions with higher expected reward in the parameter space.

Using the precomputed reward landscape $R(\alpha, \beta)$, we estimated the expected rewards associated with each participant's RL parameters at different time points. For participant s at time t (trial or block), the estimated parameters $(\alpha_t^{(s)}, \beta_t^{(s)})$ were mapped to their corresponding expected rewards using a nearest-neighbor lookup on a 100×200 grid:

$$\hat{r}_t^{(s)} = R[\alpha_t^{(s)}, \beta_t^{(s)}].$$

This yielded a reward trajectory $\hat{r}_t^{(s)}$ over time for each participant.

To assess whether these trajectories reflect systematic improvement, we conducted a lagged effect analysis. For each participant and a time lag Δt (either at the trial or block level), we computed Cohen's d (i.e., differences normalized by standard deviations) between expected rewards $\hat{r}_t^{(s)}$ at time t and $\hat{r}_{t+\Delta t}^{(s)}$ at $t + \Delta t$:

$$d_{\Delta t}^{(s)} = \texttt{Cohen's d}(\hat{r}_{t+\Delta t}^{(s)}, \hat{r}_t^{(s)}).$$

A positive $d_{\Delta t}$ indicates that RL parameter updates tend to increase expected reward, consistent with a reward-directed optimization process.

Together, these complementary measures—cosine similarity in the parameter space and reward improvement in the outcome space—allow us to rigorously assess whether human strategy updates are aligned with gradient directions.

4.6 Correlation between gradient norm and RL parameter strategy improvement

Here, we examined the relationship between the magnitude of the reward gradient along the estimated RL trajectories and the optimality of RL parameter updates (measured by the change in expected reward).

Specifically, for each participant s at each trial or block t, we pre-computed the reward gradient $\mathbf{g}_t^{(s)}$ (see Section 4.5.1). The optimality of the corresponding parameter update is measured by the change in expected reward:

$$\Delta\hat{r}_t^{(s)} = \hat{r}_{t+1}^{(s)} - \hat{r}_t^{(s)} = R(\alpha_{t+1}^{(s)}, \beta_{t+1}^{(s)}) - R(\alpha_t^{(s)}, \beta_t^{(s)}),$$

where $\hat{r}_t^{(s)}$ denotes the estimated expected reward at time t for participant s (see Section 4.5.2).

This procedure yielded a set of pairs $\left(\left\|\mathbf{g}_t^{(s)}\right\|, \Delta\hat{r}_t^{(s)}\right)$ across all participants and time points. We then computed the Pearson correlation coefficient between $\left\|\mathbf{g}_t^{(s)}\right\|$ and $\Delta\hat{r}_t^{(s)}$.

To evaluate the statistical significance, we conducted a permutation test with 10,000 iterations. In each iteration, we randomly shuffled the correspondence between $\left\|\mathbf{g}_t^{(s)}\right\|$ and $\Delta\hat{r}_t^{(s)}$ along the trial dimension for each participant to generate a null distribution of correlation coefficients. The empirical correlation was then compared against this null distribution to obtain a p-value.

4.7 Autocorrelation and timescale

We computed autocorrelation for each participant using the standard implementation from `statsmodels.tsa.stattools.acf`. Due to differences in trial count per block, we set the maximum number of time lags T to 20 for `Bahrami2020` and 4 for `Suthaharan2021`. We then averaged the autocorrelation across participants to produce group-level autocorrelation $\rho(t)$ at lag t.

To quantify the effective timescale of each variable, we computed the integrated autocorrelation time (Sokal, 1997):

$$\tau_{\text{int}} = 1 + 2\sum_{t=1}^{T} \rho(t).$$

Larger τ_{int} indicates greater temporal persistence and slower variation in the corresponding variable.

Acknowledgments

We thank the researchers who generously shared their datasets. This work was primarily conducted while HDX and RCW were at the University of Arizona and was supported by NIH grant R01AG061888 awarded to RCW. We also acknowledge the University of Arizona's High Performance Computing (HPC) facilities for providing essential computational resources.

Declaration of Interests

All authors declare no competing interests.

Data Availability

The horizon task, `Wilson2014` dataset (Wilson et al., 2014), is available upon request. The two-armed bandit task, `Gershman2018` dataset (Gershman, 2018), was sourced from task2 in and is available at https://github.com/sjgershm/exploration. The three-armed probabilistic reversal learning task, `Suthaharan2021` dataset

(Suthaharan et al., 2021), was sourced from and is available at https://github.com/psuthaharan/covid19paranoia. The four-arm restless bandit task, `Bahrami2020` dataset (Bahrami and Navajas, 2020), is available at https://osf.io/f3t2a/.

Author Contributions

H.D.X., L.J.A. conceived the study and drafted the initial manuscript. H.D.X. created and applied the models, conducted the data analysis. R.C.W. provided funding support. R.C.W., M.G.M supervised the project. All authors contributed to revising the manuscript and approved the final version.

References

Ahn, K., Cheng, X., Daneshmand, H., and Sra, S. (2023). Transformers learn to implement preconditioned gradient descent for in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 45614–45650.

Anderson, J. R. (2013). *The Adaptive Character of Thought*. Psychology Press, New York.

Andrychowicz, M., Denil, M., Gómez, S., Hoffman, M. W., Pfau, D., Schaul, T., Shillingford, B., and de Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Bahrami, B. and Navajas, J. (2020). 4 Arm Bandit Task Dataset. Publisher: OSF.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9):1214–1221. Publisher: Nature Publishing Group.

Binz, M., Gershman, S. J., Schulz, E., and Endres, D. (2022). Heuristics from bounded meta-learned inference. *Psychological Review*, pages No Pagination Specified–No Pagination Specified. Place: US Publisher: American Psychological Association.

Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X., and Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience*, 18(4):590–596. Publisher: Nature Publishing Group.

Cheng, X., Chen, Y., and Sra, S. (2024). Transformers Implement Functional Gradient Descent to Learn Non-Linear Functions In Context. arXiv:2312.06528 [cs].

Coda-Forno, J., Binz, M., Akata, Z., Botvinick, M., Wang, J. X., and Schulz, E. (2023). Meta-in-context learning in large language models. arXiv:2305.12907 [cs].

Feng, S. F., Wang, S., Zarnescu, S., and Wilson, R. C. (2021). The dynamics of explore–exploit decisions reveal a signal-to-noise mechanism for random exploration. *Scientific Reports*, 11(1):3077.

Finn, C., Abbeel, P., and Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1126–1135. PMLR. ISSN: 2640-3498.

Ger, Y., Nachmani, E., Wolf, L., and Shahar, N. (2024). Harnessing the flexibility of neural networks to predict dynamic theoretical parameters underlying human choice behavior. *PLOS Computational Biology*, 20(1):e1011678. Publisher: Public Library of Science.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, 173:34–42.

Hans, A., Wen, Y., Jain, N., Kirchenbauer, J., Kazemi, H., Singhania, P., Singh, S., Somepalli, G., Geiping, J., Bhatele, A., and Goldstein, T. (2024). Be like a Goldfish, Don't Memorize! Mitigating Memorization in Generative LLMs. arXiv:2406.10209 [cs].

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, 56(1):51–65.

Hattori, R., Hedrick, N. G., Jain, A., Chen, S., You, H., Hattori, M., Choi, J.-H., Lim, B. K., Yasuda, R., and Komiyama, T. (2023). Meta-reinforcement learning via orbitofrontal cortex. *Nature Neuroscience*, 26(12):2182–2191. Publisher: Nature Publishing Group.

Ji-An, L., Benna, M. K., and Mattar, M. G. (2025). Discovering cognitive strategies with tiny recurrent neural networks. *Nature*, pages 1–9. Publisher: Nature Publishing Group.

Lampinen, A. K., Chan, S. C. Y., Singh, A. K., and Shanahan, M. (2024). The broader spectrum of in-context learning. arXiv:2412.03782 [cs].

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. (2018). Learning to Generalize: Meta-Learning for Domain Generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Mathys, C., Daunizeau, J., Friston, K. J., and Stephan, K. E. (2011). A Bayesian Foundation for Individual Learning Under Uncertainty. *Frontiers in Human Neuroscience*, 5. Publisher: Frontiers.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., and Stephan, K. E. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*, 8. Publisher: Frontiers.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457. Place: US Publisher: American Psychological Association.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., and Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, 15(7):1040–1046.

Nassar, M. R., Wilson, R. C., Heasly, B., and Gold, J. I. (2010). An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. *Journal of Neuroscience*, 30(37):12366–12378.

Ortega, P. A., Wang, J. X., Rowland, M., Genewein, T., Kurth-Nelson, Z., Pascanu, R., Heess, N., Veness, J., Pritzel, A., Sprechmann, P., Jayakumar, S. M., McGrath, T., Miller, K., Azar, M., Osband, I., Rabinowitz, N., György, A., Chiappa, S., Osindero, S., Teh, Y. W., van Hasselt, H., de Freitas, N., Botvinick, M., and Legg, S. (2019). Meta-learning of Sequential Strategies. arXiv:1905.03030 [cs, stat].

Piray, P. and Daw, N. D. (2020). A simple model for learning in volatile environments. *PLOS Computational Biology*, 16(7):e1007963.

Piray, P. and Daw, N. D. (2021). A model for learning based on the joint estimation of stochasticity and volatility. *Nature Communications*, 12(1):6587.

Rescorla, R. A. and 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. Publisher: Appleton-Century-Crofts.

Sadeghiyeh, H., Wang, S., Alberhasky, M. R., Kyllo, H. M., Shenhav, A., and Wilson, R. C. (2020). Temporal discounting correlates with directed exploration but not with random exploration. *Scientific Reports*, 10(1):4020. Publisher: Nature Publishing Group.

Schumacher, L., Bürkner, P.-C., Voss, A., Köthe, U., and Radev, S. T. (2023). Neural superstatistics for Bayesian estimation of dynamic cognitive models. *Scientific Reports*, 13(1):13778. Publisher: Nature Publishing Group.

Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. pages 92–96, Austin, Texas.

Sokal, A. (1997). Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms. In DeWitt-Morette, C., Cartier, P., and Folacci, A., editors, *Functional Integration*, volume 361, pages 131–192. Springer US, Boston, MA. Series Title: NATO ASI Series.

Soltani, A., Lee, D., and Wang, X.-J. (2006). Neural mechanism for stochastic behaviour during a competitive game. *Neural Networks*, 19(8):1075–1090.

Somerville, L. H., Sasse, S. F., Garrad, M. C., Drysdale, A. T., Abi Akar, N., Insel, C., and Wilson, R. C. (2017). Charting the expansion of strategic exploratory behavior during adolescence. *Journal of Experimental Psychology: General*, 146(2):155–164.

Suthaharan, P., Reed, E. J., Leptourgos, P., Kenney, J. G., Uddenberg, S., Mathys, C. D., Litman, L., Robinson, J., Moss, A. J., Taylor, J. R., Groman, S. M., and Corlett, P. R. (2021). Paranoia and belief updating during the COVID-19 crisis. *Nature Human Behaviour*, 5(9):1190–1202. Publisher: Nature Publishing Group.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press Cambridge.

von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. (2023a). Transformers Learn In-Context by Gradient Descent. In *Proceedings of the 40th International Conference on Machine Learning*, pages 35151–35174. PMLR. ISSN: 2640-3498.

von Oswald, J., Niklasson, E., Schlegel, M., Kobayashi, S., Zucchet, N., Scherrer, N., Miller, N., Sandler, M., Arcas, B. A. y., Vladymyrov, M., Pascanu, R., and Sacramento, J. (2023b). Uncovering mesa-optimization algorithms in Transformers. arXiv:2309.05858 [cs].

Waltz, J. A., Wilson, R. C., Albrecht, M. A., Frank, M. J., and Gold, J. M. (2020). Differential Effects of Psychotic Illness on Directed and Random Exploration. *Computational Psychiatry*, 4(0).

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., and Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, 21(6):860–868.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., and Botvinick, M. (2017). Learning to reinforcement learn. arXiv:1611.05763 [cs, stat].

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

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., and 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.

Xiong, H.-D., Ji-An, L., Mattar, M., and Wilson, R. (2025a). DynamicRL: Data-Driven Estimation of Trial-by-Trial Reinforcement Learning Parameters.

Xiong, H.-D., Ji-An, L., Mattar, M. G., and Wilson, R. C. (2025b). Deep Learning Improves Parameter Estimation in Reinforcement Learning Models. Pages: 2025.03.21.644663 Section: New Results.

Zajkowski, W. K., Kossut, M., and Wilson, R. C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. *eLife*, 6:e27430. Publisher: eLife Sciences Publications, Ltd.

A Supplementary Material

A.1 Dynamic RL parameter recovery

To evaluate the validity and robustness of the `DynamicRL` framework, we conducted a model recovery analysis using synthetic behavioral data generated from RL agents with predefined dynamic RL parameter trajectories. This analysis assesses whether the model can reliably recover the temporal dynamics of RL parameters when the ground-truth values are known (Wilson and Collins, 2019).

We tested model recovery on two datasets with trial-level dynamics: `Bahrami2020` and `Suthaharan2021`. In `Bahrami2020`, we used the original 4-armed restless bandit task without controlling for the three identical reward schedules. For `Suthaharan2021`, we focused on the condition with reward probabilities $[0.9, 0.5, 0.1]$.

For each dataset, we defined a dynamic range of α and β, and generated linear RL parameter trajectories (increasing, decreasing, or constant) across trials. Initial and final parameter values were sampled from a 3-by-3 grid within empirically observed ranges: $[0.4, 0.8]$ for α in both datasets, $[4, 10]$ for β in `Bahrami2020`, and $[2, 8]$ for β in `Suthaharan2021`. Each condition was repeated ten times using different synthetic agent IDs, simulating distinct participants employing the same strategies and producing a diverse set of RL parameter trajectories. The parameter ranges were selected based on the interquartile range (second to third quartile) of estimates from the fixed-parameter models and were slightly expanded to ensure broader coverage.

We then applied the `DynamicRL` model to these synthetic datasets, using the same fitting and selection procedure described in Sec. 4.2.3, and compared the recovered trajectories to the ground-truth values. In both datasets, the model successfully captured the qualitative trend of parameter evolution. Constant parameters were recovered as stable, flat trajectories. Monotonic changes in α and β were consistently reflected in the recovered trajectories (Fig. A.1, A.2). Quantitatively, recovered trajectories were strongly correlated with the true values. Across conditions, most Pearson correlations exceeded 0.95. Recovery accuracy was slightly lower in `Suthaharan2021`, likely due to fewer trials and discontinuities introduced by block structures. Nonetheless, overall parameter recovery performance remained high, demonstrating that `DynamicRL` effectively identifies dynamic RL parameters under a broad range of conditions.

A.2 Alternative moving average RL model

A natural baseline for estimating time-varying RL parameters is the sliding window approach: partition the trial sequence into overlapping windows and fit standard RL models within each window. This method is conceptually simple and can, in principle, capture non-stationary changes in behavior over time.

To empirically evaluate this approach, we applied a sliding-window RL estimator to the `Bahrami2020` dataset, which contains over 100 trials per participant. We selected this dataset because the relatively long trial sequences permit the use of larger window sizes. Other datasets were excluded due to insufficient trial counts (`Suthaharan2021` has 40 trials) or block counts (`Gershman2018` has 20 blocks; `Wilson2014` has 60 blocks), which would restrict window size and reduce the reliability of parameter estimation—particularly given the stochastic nature of rewards and the heightened risk of overfitting when data within each window are sparse.

We systematically varied the window size, using a window stride of 5 trials (e.g., for a window size of 30, windows covered trials 0–30, 5–35, etc.). Within each window, we fit fixed RL parameters using maximum likelihood estimation with the Nelder–Mead optimization method, initialized from 16 grid-based starting points. We constructed this grid by uniformly sampling four values for each parameter: $\alpha \in [0, 1]$, $\beta \in [0, 60]$. This initialization strategy established a strong baseline for comparison.

Because a principled train-test split is not straightforward in this setting, we directly compared the training negative log-likelihood (NLL) of the sliding-window model to the test NLL of both the standard fixed-RL model and the `DynamicRL` model. We notice that the training NLL usually exaggerates the predictive performance due to overfitting.

As shown in Fig. A.3, small windows (e.g., 10 trials) yield lower training NLL but produce noisy estimates and are potentially prone to overfit. Further, these still perform worse than the test loss of `DynamicRL`. Larger windows improve stability but fail to capture fine-grained trial-level dynamics, sacrificing temporal resolution. Even at optimal window size, the sliding-window model underperforms relative to `DynamicRL`.

Therefore, the neural network architecture in `DynamicRL` is necessary to provide robust and accurate estimation by learning to infer RL parameters across the full dataset. It leverages shared strategies across participants and conditions, producing smooth and context-sensitive parameter trajectories that generalize well.

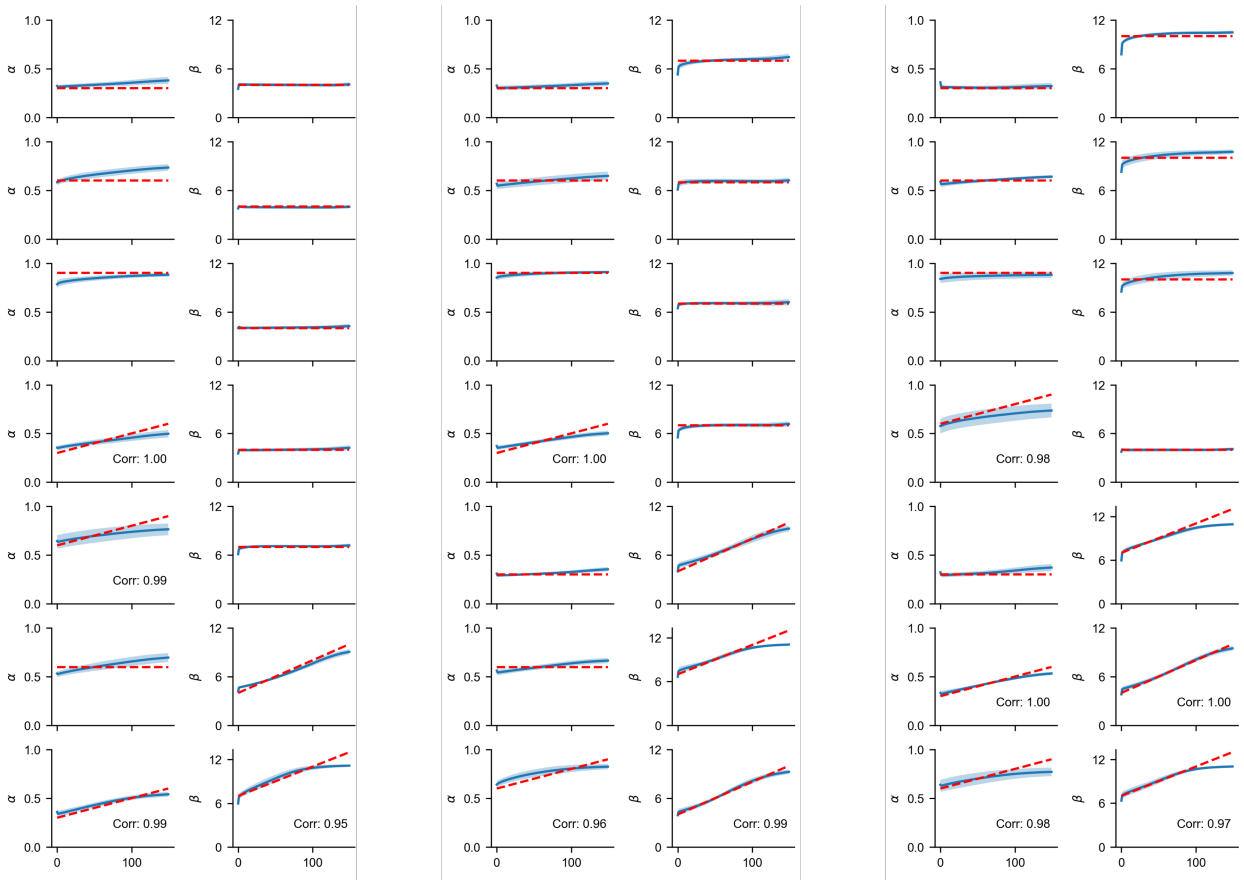

Figure A.1: **Dynamic RL parameter for** `Bahrami2020`**.** Red dashed lines indicate ground-truth RL parameter values; blue lines are estimates by `DynamicRL` from data.

A.3 Supplementary figures

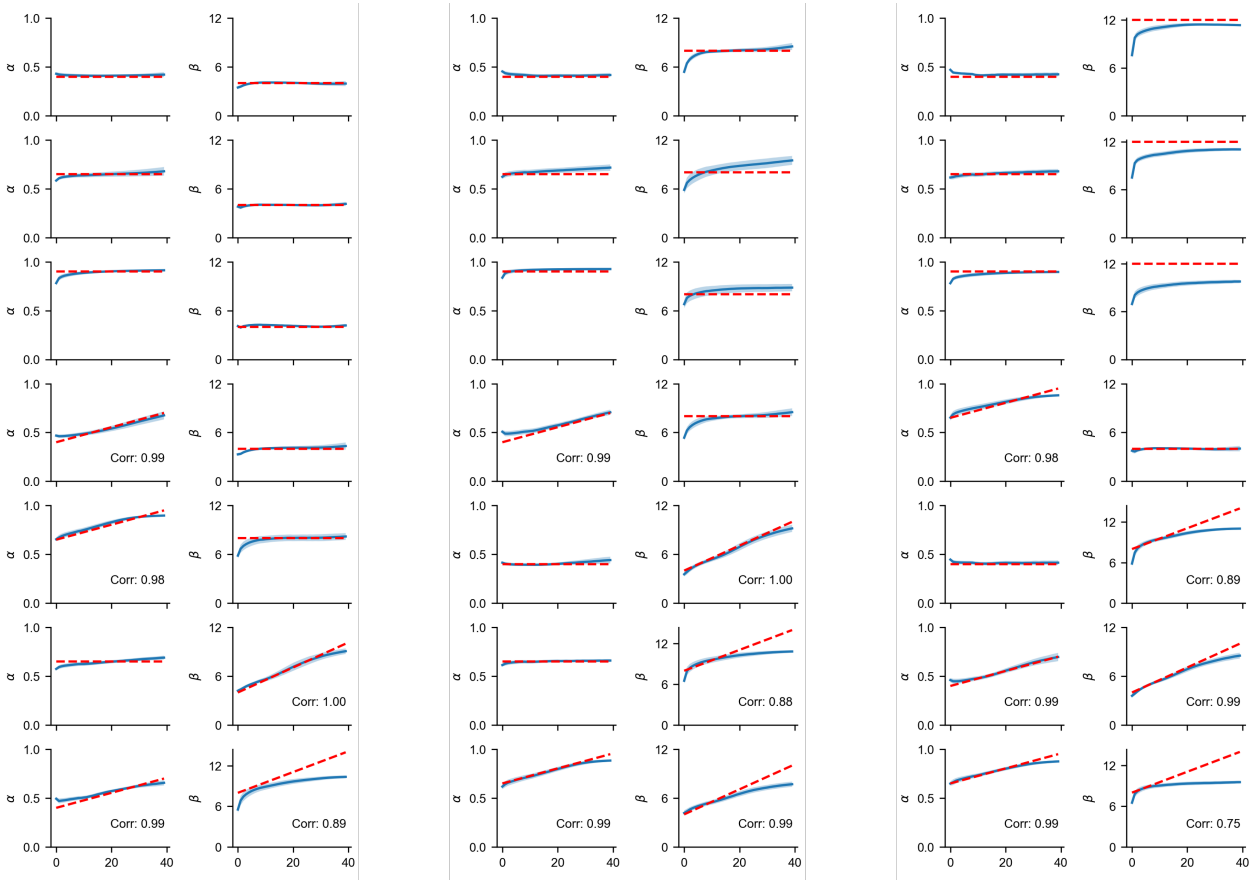

Figure A.2: **Dynamic RL parameter recovery for** `Suthaharan2021`**.** Red dashed lines indicate ground-truth RL parameter values; blue lines are estimates by `DynamicRL` from data.

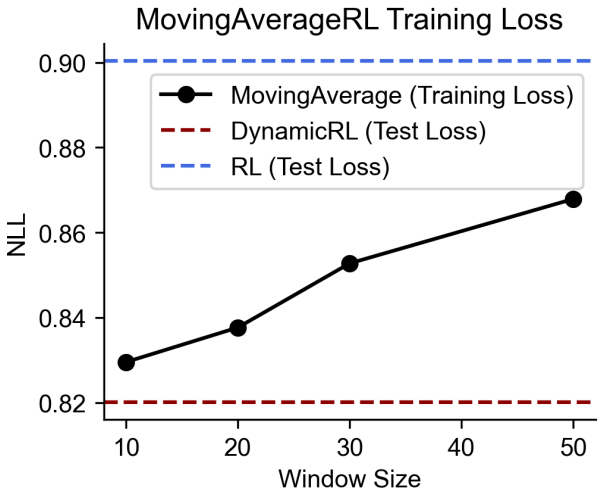

Figure A.3: **Performance of sliding-window RL estimation on** `Bahrami2020`**.** Training NLL increases with window size (i.e., fit performance decreases). Smaller windows overfit and produce unstable estimates. Even the best-performing sliding-window model underperforms relative to `DynamicRL` on held-out data.

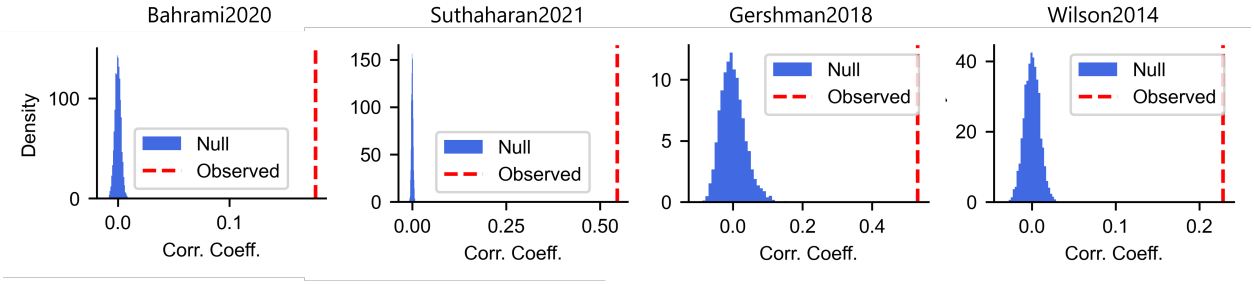

Figure A.4: **Correlation between gradient norm and meta-RL strategy improvement.** For each dataset, we computed the empirical correlation between the magnitude of the local reward gradient and the optimality of meta-RL updates. The blue histogram shows the null distribution obtained via permutation testing. The red dashed line indicates the observed correlation coefficient. In all datasets, the observed correlation significantly exceeds the null distribution.