JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Exploration Is a State, Not a Setting: A Markov-Switching Reinforcement-Learning Model of Strategy Transitions in Sequential Choice

Fathimah Al-Ma'shumah [1,2,3]*, Feneta Fidi Kirani [4], and Nita Ratnawaty [5]

[1] Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia; e-mail : fathimah.al-mashumah@binus.ac.id
[2] Center of Excellence Predictive Risk Simulation and Modelling (PRISM), Institut Teknologi Bandung, Bandung 40132, Indonesia
[3] Research Consortium Financial Modeling, Optimization and Simulation (RC-FinanMOS), Institut Teknologi Bandung 40132, Bandung, Indonesia
[4] Psychology Department, Faculty of Humanities, Bina Nusantara University, Jakarta 11480, Indonesia; e-mail : feneta.kirani@binus.ac.id
[5] Uttama Analytics Consulting, Bandung 40559, Indonesia; e-mail : nratnawaty@uttama-analytics.com
* Corresponding Author : Fathimah Al-Ma'shumah

Abstract: Computational accounts of human exploration usually assume a stationary policy, in which a single set of parameters for value sensitivity and uncertainty seeking generates every choice in a task, so that all within-task variation is treated as decision noise. Neuroscience instead treats exploration and exploitation as dissociable modes between which the brain switches, which a stationary model cannot represent. We introduce the Markov-Switching Reinforcement-Learning (MS-RL) model, in which a first-order hidden Markov chain governs transitions among a small number of latent decision regimes, each with its own softmax policy over a shared value-learning process. The three regimes are exploitation, directed exploration, and random exploration. The model contains the standard stationary account (one regime) and a temporally unstructured mixture (memoryless transitions) as nested special cases, making the stationarity assumption testable. We estimate the model using Expectation-Maximization with Viterbi decoding and select the number of regimes using the Bayesian information criterion. A parameter- and state-recovery study confirmed that the generating parameters and latent regime paths are recoverable (parameter correlations 0.88–0.94; state accuracy 86 percent; Cohen's kappa ≈ 0.79). Applied to an openly available two-armed bandit dataset (46 adults, 13,800 choices), a three-regime MS-RL model was preferred over the nested baselines, two- and four-regime variants, and a single-regime model with smoothly time-varying value sensitivity. An ablation analysis attributes the largest gains to the latent regimes and temporal switching. The decoded regime path revealed a systematic within-game shift from directed exploration toward exploitation. The estimated transition matrix provides a per-participant measure of strategy change that has no counterpart in stationary models. We conclude that exploration is better described as a dynamic state than as a fixed trait, and that modeling it as a switching process is both more accurate and useful.

Keywords: Computational behavioral modeling; Directed and random exploration; Exploration–exploitation; Hidden Markov model; Latent states; Markov switching; Reinforcement learning; Strategy dynamics.

Received: June, 25th 2026
Revised: July, 29th 2026
Accepted: July, 31st 2026
Published: August, 21st 2026

1. Introduction

Systems that adapt to people over time must track a moving target. A tutoring system, a recommender, or a clinical decision aid observes a sequence of choices and infers what a person is doing, yet the strategy generating those choices may change: the same person may spend one stretch of a session probing unfamiliar options and the next settling on a known favorite. Reading that shift correctly has practical consequences. An adaptive learning system

that introduces new material while a learner is consolidating may interrupt rather than help, while a behavioral assessment that treats every non-greedy choice as an error cannot distinguish between a person who has stopped seeking information and one who has become inattentive. Knowing when a strategy changes, and in which direction, is therefore a measurement problem with immediate applications. It calls for models that locate strategy changes in time rather than summarizing them as an average over a task.

When a decision maker faces repeated choices under uncertainty, they confront the explore–exploit dilemma: select the option that currently seems best, or sample an uncertain alternative that might turn out to be better [1], [2]. Human exploration is itself algorithmically structured, combining value-guided and uncertainty-guided components rather than reducing to simple randomness [3]. Exploration is also not merely a reaction to uncertainty; it reflects dynamic changes in cognitive control, where rising uncertainty or conflict promotes information-seeking and behavioral flexibility [4]. Formal accounts of this trade-off must therefore specify not only how a decision maker weighs value against uncertainty, but also whether that weighting can change as the task unfolds.

The dominant computational treatment of this dilemma assumes that each individual resolves it with a fixed policy: a single inverse-temperature parameter for value sensitivity and, in directed-exploration models, a single uncertainty bonus, both held constant across an entire task. This is the stationary-policy assumption. Under this assumption, every choice is drawn from the same decision rule, and all within-task variation in exploratory behavior is absorbed into independent decision noise. The assumption is convenient, but it is increasingly difficult to defend. Neuroscience describes exploration and exploitation as dissociable modes that the brain enters and leaves rather than as points on a single continuum, while recent latent-state models show that behavior once dismissed as random lapsing can reflect structured switching between discrete strategies. A model that fixes one policy for the whole task cannot represent this structure: it cannot distinguish whether a non-greedy choice reflects information-seeking or noise, whether a strategy persists across a sequence of trials, or how the balance of strategies changes as a task unfolds.

This is the gap we address. We treat the parameters that govern exploration not as fixed traits but as latent states that switch over time, making the switching dynamics themselves an object of inference. We introduce the Markov-Switching Reinforcement-Learning (MS-RL) model, in which a first-order hidden Markov chain governs transitions among a small number of latent regimes. Within each regime, choices follow a regime-specific softmax policy that combines a learned value term with a count-based information term, while value learning is shared across regimes so that the regimes represent distinct modes of action selection rather than distinct learners. With one regime, the model reduces to stationary softmax reinforcement learning, while a memoryless transition matrix reduces it to a temporally unstructured mixture. Thus, the two dominant alternatives are nested special cases, making the stationarity assumption a testable hypothesis rather than a default. We estimate the model using Expectation-Maximization with Viterbi decoding, select the number of regimes using the Bayesian information criterion, and validate identifiability through parameter- and state-recovery studies before fitting the human data. The main contributions of this work are summarized as follows:

- A Markov-Switching Reinforcement-Learning framework that allows the action-selection policy to switch among latent regimes while value learning remains shared, providing an interpretable representation of exploitation, directed exploration, and random exploration.
- A unified latent-state formulation of exploration dynamics in which stationary reinforcement learning and temporally unstructured mixtures are represented as nested special cases, allowing the stationarity assumption to be evaluated against temporal switching.
- A regime-transition representation of strategy dynamics that provides participant-specific information about strategy persistence and switching, which is not directly available from conventional stationary reinforcement-learning accounts.
- A systematic validation and empirical evaluation framework combining parameter and state recovery, model comparison, and component-level ablation to assess the identifiability, fit, and behavioral relevance of the proposed model.

Taken together, these contributions provide a framework for modeling exploration as a dynamic latent state rather than as a fixed policy parameter. The empirical analysis further

examines whether this switching structure captures within-game transitions between exploration and exploitation in human sequential choice. The remainder of the paper is organized as follows. Section 2 reviews related work and positions MS-RL against existing latent-state and decision-making models. Section 3 specifies the model and its estimation, including Algorithm 1. Section 4 reports the experimental evaluation, including the dataset and baselines, validation on simulated data, recovered regimes and their within-game dynamics, model comparison, and ablation analysis. Section 5 discusses the contributions, practical implications, and limitations, and Section 6 concludes.

2. Related Work

Three strands of prior work are particularly relevant to the present model: computational models of exploration, neural and psychological accounts of discrete decision modes, and latent-state models of behavior. We review each and then position MS-RL within these modeling directions.

2.1 Models of Exploration in Sequential Choice

Computational models of the explore–exploit trade-off typically fit a single softmax policy with a fixed inverse temperature and, in directed-exploration variants, a fixed uncertainty bonus [2], [3], [5]. These models have established that human exploration combines directed and random components and that directed exploration is driven by uncertainty [2], [6], [7], with related evidence from restless-bandit and model-based accounts of choice [8], [9]. Their shared limitation is stationarity: one policy is assumed to generate every choice, so changes in strategy within a task are represented as noise rather than as changes in state. The present model relaxes this assumption while retaining the stationary policy as its one-regime special case.

The stationarity assumption has also been questioned within this family of bandit tasks. Ger et al. [10] trained a recurrent network on synthetic data to infer trial-by-trial reinforcement-learning parameters and found, on the closely related dataset of Gershman [7], that allowing these parameters to vary over time predicted held-out choices better than fitting them as constants. We take this result as motivation rather than as a direct competitor: it provides independent evidence that a fixed policy may be an inadequate idealization for this paradigm. However, such an approach does not provide an explicit discrete strategy structure. Its recovered trajectories are continuous and network-derived, so they can show that parameters change without specifying how many strategies a person uses, how long each persists, or how often one gives way to another. The present model instead imposes a small number of regimes connected by a Markov chain and estimates their switching dynamics directly. Nor is instability confined to within-task timescales. In a twelve-week longitudinal study, the weight participants placed on relative uncertainty in a two-armed bandit varied systematically across sessions with practice and affective state [11]. Taken together with the within-task evidence, these findings suggest that variation in exploration parameters can reflect meaningful changes in the policy being measured rather than measurement noise alone.

2.2 Neural and Psychological Evidence for Discrete Modes

A parallel literature treats exploration and exploitation as distinct neural and cognitive modes rather than as points on a continuum. Dissociable activity in frontopolar and striatal systems [1], causal separability of directed and random exploration [12], and the adaptive-gain account of locus-coeruleus function [13]–[15] all describe exploration as a state that a decision system enters and leaves. Psychological accounts of cognitive control as a dynamic, demand-sensitive process [4], [16] and of novelty- and uncertainty-driven information seeking [17], [18] support a similar view at the behavioral level.

This dissociation has also been demonstrated within the task family considered here. Using fMRI on a two-armed bandit that independently manipulates relative and total uncertainty, Tomov et al. [19] found relative uncertainty represented in the right rostrolateral prefrontal cortex and associated with directed exploration, while total uncertainty was represented in the dorsolateral prefrontal cortex and associated with random exploration. Each signal predicted cross-trial variability in its corresponding strategy rather than the other. Directed and random exploration are therefore not merely labels for a statistical partition of choices; in this paradigm, they have separable neural correlates. This evidence motivates a

latent-state formulation but does not itself provide a computational model that recovers such states from behavioral choices.

2.3 Latent-State Models of Behavior

Hidden Markov and related state-switching models recover discrete behavioral states from sequential data and show that apparent lapses can reflect structured switching [5], [20], [21]. Closest to the present work, recurrent-network and sequential-exploration models capture exploration as a latent process that can be recovered from choices [5], [6]. However, these approaches do not allow reinforcement-learning policy parameters themselves to switch with a latent regime while keeping value learning shared across regimes.

This distinction is important for interpreting the resulting states. A whole-rule switching model may capture changes in behavior but does not necessarily distinguish changes in learning from changes in action selection. MS-RL instead places a hidden Markov chain over regime-specific softmax policies that share a common value-learning process. The resulting regimes therefore represent changes in action-selection strategy while preserving a common learning process, providing a direct link between latent states and interpretable exploration parameters.

2.4 Research Gap and Position of the Present Study

Prior work has advanced on two fronts that the present study brings together. Computational models of exploration have established that human choice combines directed and random components and that directed exploration is uncertainty-driven [2], [3], yet they generally assume a stationary policy, so within-task changes in strategy are represented as noise rather than as changes in state. Latent-state models, in turn, have shown that discrete behavioral states can be recovered from choices and that apparent lapsing can reflect structured switching [5], [6], [10], [20], but they do not explicitly separate a switching action-selection policy from a shared value-learning process.

Two gaps therefore remain. First, the stationarity assumption in explore–exploit models is typically imposed rather than directly tested, leaving limited scope for determining whether exploration is better characterized as a fixed setting or as a state that changes within a task. Second, existing latent-state approaches do not directly connect the recovered states to the interpretable policy parameters used in exploration models, such as value sensitivity and uncertainty seeking. The present study addresses these gaps by combining a hidden Markov state structure with regime-specific policies and shared value learning.

MS-RL is therefore positioned at the intersection of these modeling directions. It nests the stationary policy as a testable special case, assigns interpretable policy parameters to each latent regime, and estimates a transition matrix that characterizes strategy persistence and switching. Table 1 summarizes this positioning against representative approaches to sequential choice.

Table 1. Positioning of MS-RL against representative modeling directions for sequential choice

Modeling direction	Representative studies	Discrete strategy switching	Interpretable regime policy parameters	Stationarity tested (nested), not assumed	Per-participant transition matrix
Stationary softmax RL	[2], [3]	No	Yes	No	No
Memoryless policy mixture	[2], [3]	Partial (no temporal order)	Yes	No	No
Time-varying-temperature RL	[22]	No (gradual drift)	Yes	No	No
HMM latent-state choice models	[20]	Yes	No (whole rule switches)	No	Yes
Recurrent / sequential-exploration networks	[5], [6]	Yes (implicit)	No	No	No
Present study (MS-RL)	This work	Yes	Yes	Yes	Yes

As shown in Table 1, approaches that retain interpretable policy parameters generally do not explicitly model discrete temporal switching, whereas approaches that capture latent temporal structure do not directly connect the states to regime-specific exploration policies. MS-

RL combines these properties while also treating stationarity as a testable modeling assumption rather than a fixed premise. This combination enables the estimated transition matrix and decoded regime sequence to be interpreted in terms of strategy persistence and change rather than as an unlabeled sequence of latent states.

3. The Markov-Switching Reinforcement-Learning Model

3.1. Learning Component

On each trial t, the agent chooses action a_t and observes reward r_t. Action values follow a delta-rule update [23] shared across regimes, given by Eq. (1):

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

where the learning rate $\alpha \in (0,1)$, and unchosen actions retain their previous values. The reward r_t is the observed payoff, and $Q_t(a)$ is the running value estimate for action a at trial t. An information signal, defined in Eq. (2), provides a count-based proxy for uncertainty:

$$I_t(a) = \frac{1}{\sqrt{n_t(a)+1}} \tag{2}$$

where $n_t(a)$ is the number of prior selections of action a. The signal $I_t(a)$ is highest for rarely sampled actions and decreases as evidence accumulates. It is computed deterministically from the choice history rather than estimated as a separate parameter, and serves as a proxy for the uncertainty that can drive directed exploration [1], [2]. This count-based form provides a simple, parameter-free representation of relative uncertainty in the bandit task.

3.2. Latent Regime Dynamics

Let $S_t \in \{1, \ldots, K\}$ denote the latent regime on trial t, evolving as a first-order Markov chain with initial distribution π and $K \times K$ transition matrix P, whose entries are defined in Eq. (3):

$$P_{jk} = \Pr(S_{t+1} = k \mid S_t = j) \tag{3}$$

The off-diagonal entries of P quantify the propensity to switch strategies, whereas the diagonal entries quantify regime persistence. These switching dynamics are estimated rather than fixed and constitute the model's central object of inference.

3.3. Emission (Choice) Model

Conditional on regime k, choice probabilities follow a regime-specific softmax policy that combines value and information, as given in Eq. (4):

$$\Pr(a_t = a \mid S_t = k) \propto \exp[\,\beta_k\, Q_t(a) + \gamma_k\, I_t(a)\,] \tag{4}$$

where β_k is the regime-k inverse temperature (value sensitivity, indexing exploitation) and γ_k is the regime-k information weight. A positive γ_k reflects uncertainty-seeking behavior associated with directed exploration, whereas a γ_k near zero reflects value-dominated or near-uniform choice. Regimes are not labeled a priori; their interpretation is assigned post hoc from contrasts in the fitted parameters, as summarized in Table 2.

Table 2. Regime identification by parameter contrast. Labels are assigned post hoc from the fitted policy parameters rather than imposed a priori.

Regime	β (value weight)	γ (information weight)	Interpretation
Exploit	High	Low	Greedy value maximization
Directed-explore	Moderate	High (positive)	Uncertainty-driven sampling
Random-explore	Low	Low	Near-uniform choice

As shown in Table 2, the three regimes are distinguished by their relative value sensitivity and information weighting rather than by predefined state labels. This allows the behavioral interpretation of each regime to emerge from the fitted policy parameters.

3.4. Likelihood and Nested Baselines

For a participant with choice sequence $a_{1:T}$, the observed-data likelihood marginalizes over the latent regime path, as given in Eq. (5):

$$L(\theta) = \prod_{i=1}^{N} \sum_{S_{1:T}} \Pr\left(a_{1:T}^{(i)}, S_{1:T} \mid \theta\right) \tag{5}$$

and is computed using the forward recursion. Conditional on the regime path, emissions are independent across trials because the value trajectory Q_t is deterministic given the past choices and rewards. The forward–backward recursion can therefore be applied directly, with the value and information trajectories computed in a single forward sweep so that each emission term uses the corresponding Q_t and I_t. With K=1, the model reduces to stationary softmax-RL, whereas constraining every row of P to a common distribution yields a memoryless mixture. The contrast between the full MS-RL model and the memoryless mixture therefore isolates the explanatory contribution of temporal switching.

As illustrated in Figure 1, the shared learning process updates the action values from the observed reward, while the latent regime determines which regime-specific softmax policy is active. The selected policy combines the current value and information signals to generate the observed choice, after which the reward updates the shared value estimates. Across trials, the first-order Markov chain governs transitions between regimes. Thus, the model separates changes in action-selection strategy from the underlying value-learning process.

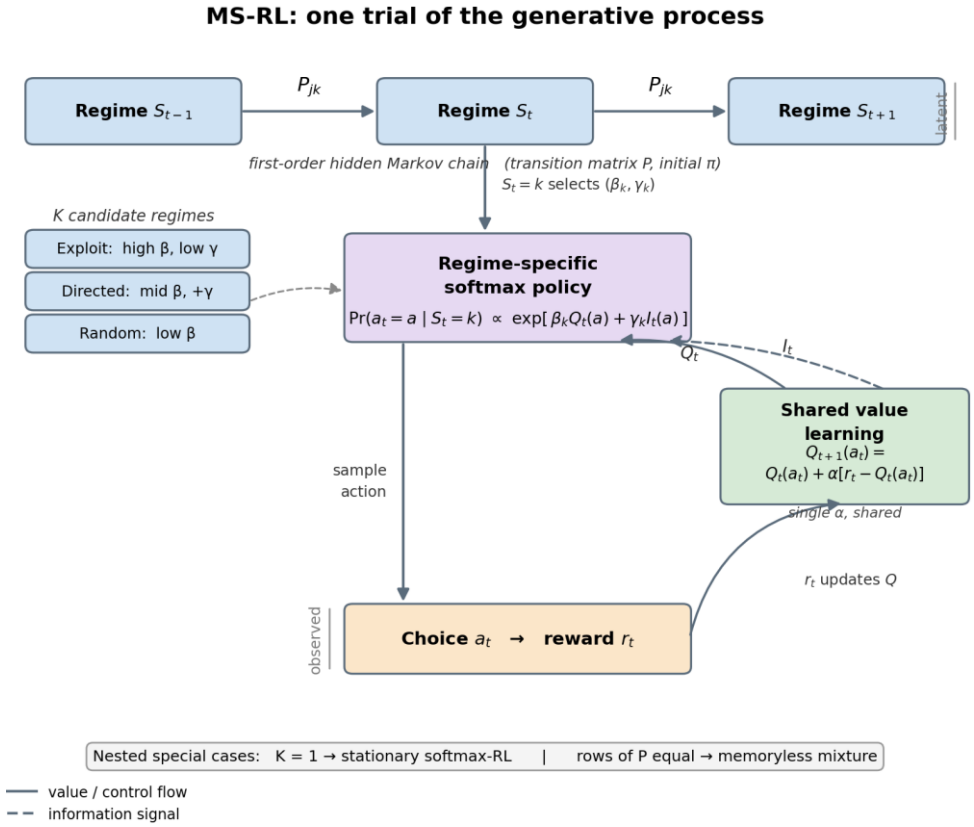

Figure 1. Generative architecture of the proposed MS-RL model.

3.5. Estimation

We estimate the parameters using Expectation-Maximization, fitting the model separately to each participant so that each participant has their own learning rate, regime-specific

policy parameters, transition matrix, and decoded regime path; no parameters are pooled across participants. The E-step uses a log-space forward–backward recursion [24], after which the transition matrix and initial distribution have closed-form updates. The policy parameters $(\alpha, \{\beta_k\}, \{\gamma_k\})$ have no closed-form solution because they enter the softmax nonlinearly and are recursive through the value update. They are therefore obtained by numerically maximizing the responsibility-weighted expected emission log-likelihood under bound constraints [25].

Two choices are specific to this model. The learning rate α is shared across regimes, which allows the regimes to represent modes of action selection rather than separate learners. In addition, regimes are ordered by descending β to resolve label switching; because the recovered β values are well separated (Section 4.4), this ordering is unambiguous. We use multiple random restarts because the likelihood is multimodal, decode the most likely regime path for each participant using the Viterbi algorithm, obtain standard errors from the observed information with a subject-level bootstrap as a robustness check, and select K using the Bayesian information criterion, with K=1 nested as a special case.

The complete estimation procedure is summarized in Algorithm 1. For each random restart, the model parameters are initialized within the specified bounds, followed by alternating E- and M-steps until convergence. The E-step computes the value and information trajectories and the posterior responsibilities using the log-space forward–backward recursion. The M-step updates the initial distribution and transition matrix in closed form and numerically optimizes the policy parameters using L-BFGS-B. The fit with the highest log-likelihood is retained across restarts, after which the most likely regime path is decoded using the Viterbi algorithm. Finally, the number of regimes is selected using BIC, with K=1 corresponding to stationary softmax-RL.

Algorithm 1. Expectation-Maximization estimation of the MS-RL model

INPUT: choice/reward sequences $\{a_{1:T}, r_{1:T}\}$ per participant; number of regimes K
OUTPUT: learning rate α, policies $\{\beta_k, \gamma_k\}$, initial distribution π, transition matrix P, and regime paths

for each random restart do
  initialize $\alpha, \{\beta_k\}, \{\gamma_k\}, \pi, P$ (descending β to fix label order)
  repeat (EM until Δ log-likelihood < ε)
    E-step: forward sweep to compute value Q_t and information I_t trajectories
    compute emission log-probabilities $\log p(a_t \mid S_t = k)$ for all t, k
    run log-space forward–backward ⇒ responsibilities $\gamma_t(k), \xi_t(j,k)$
    M-step: $\pi \leftarrow \gamma_1$; $P_{jk} \leftarrow$ normalized $\Sigma_t\, \xi_t(j,k)$ (closed form)
    $(\alpha, \{\beta_k\}, \{\gamma_k\}) \leftarrow \arg\max$ responsibility-weighted emission log-likelihood (L-BFGS-B)
  until converged
  keep fit if log-likelihood exceeds best-so-far
end for
decode most-likely regime path per participant with the Viterbi algorithm
select K by BIC (K=1 nested = stationary softmax-RL)

3.6. Reproducibility and Implementation Configuration

To make the procedure reproducible, we report the full estimation configuration in Table 3, which is applied consistently across the validation study (Section 4.3) and the analysis of human choices (Sections 4.4 to 4.6). The policy parameters are bounded during optimization (value sensitivity within a nonnegative range and the information weight within a symmetric interval around zero), and each participant is fitted from twenty random restarts to guard against local optima arising from the multimodal likelihood; the restart with the highest log-likelihood is retained. EM iterates until the change in log-likelihood falls below 1e−6 or a maximum of two hundred iterations is reached. Random seeds are fixed for every simulation and bootstrap resample, so the recovery study, model fits, and reported intervals can be reproduced exactly. The full code is available on request, and the analyzed data are public (see the Data Availability statement), allowing the entire pipeline to be rerun end to end.

Table 3. Estimation and implementation settings for the MS-RL model. These settings apply to both the validation study (Section 4.3) and the analysis of human choices (Sections 4.4 to 4.6).

Setting	Configuration
Estimation method	Expectation-Maximization (EM)
E-step	Log-space forward–backward recursion
M-step (transitions, initial)	Closed-form responsibility-weighted updates
M-step (policy $\alpha, \beta_k, \gamma_k$)	L-BFGS-B with bound constraints and numerical gradient
Parameter bounds	$\alpha \in (0,1)$; $\beta_k \in [0,15]$; $\gamma_k \in [-3,3]$
Initialization	Each restart draws $\alpha \sim \mathrm{U}(0.3,0.8), \beta_k \sim U(0.1,5.0)$ sorted in descending order to fix labels, $\gamma_k \sim \mathrm{U}(-0.5,0.5)$; P initialized near-diagonal, π uniform
Random restarts	20 per participant; restart with the highest log-likelihood retained
Convergence criterion	Δ log-likelihood $< 1e-6$ between EM iterations
Maximum EM iterations	200
Label-switching control	Regimes ordered by descending β (value sensitivity)
Latent-path decoding	Viterbi (most-likely regime sequence per participant)
Model selection	Bayesian information criterion (BIC); $K \in \{1, 2, 3, 4\}$
Uncertainty quantification	Observed-information Hessian; subject-level bootstrap (1,000 resamples)
Random seed	Fixed (seed = 42) for all simulations and bootstrap resampling
Software	Python (NumPy, SciPy optimize); code available on request

3.7. Evaluation Metrics

We evaluate the model primarily by how well it explains observed human choices and whether the structure it recovers can be trusted [26], [27]. Conventional reinforcement-learning measures such as cumulative reward, regret, and task accuracy are not directly applicable because rewards are determined by participants' actual choices. We therefore distinguish validation, which assesses recovery from simulated data with known ground truth, from comparative performance, which assesses model fit on human choices. Comparative performance is assessed primarily using the Bayesian information criterion (BIC):

$$\mathrm{BIC} = -2\log\hat{L} + k\log n \tag{6}$$

where L is the maximized likelihood, k is the number of free parameters, and n is the number of observed choices. BIC balances model fit and complexity and is suitable for comparing nested and non-nested cognitive models [27]. Because stationary softmax-RL is the nested K=1 case of MS-RL, BIC provides a direct test of whether the switching structure is supported by the data. BIC is used instead of AIC because its stronger complexity penalty is more conservative when models differ in the number of latent regimes. The original analysis of these data selected the same model under both criteria [28]. We also report McFadden pseudo-R^2 against a random-choice baseline and a likelihood-ratio test comparing MS-RL with its stationary special case.

Validation is assessed using simulated data with known ground truth [26]. Section 4.3 reports correlations between generated and recovered parameters and the accuracy of Viterbi-decoded regime labels. Because accuracy can be affected by unequal regime frequencies, Cohen's kappa is also reported:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{7}$$

where p_o is observed agreement and p_e is agreement expected by chance. A value of zero indicates chance agreement, whereas one indicates perfect agreement.

BIC and likelihood-ratio tests rely on regularity assumptions that are only approximate for hidden-Markov and mixture models. They are therefore interpreted as comparative evidence rather than exact model-evidence or reference-test quantities and are considered together with the recovery analyses in Section 4.3.

4. Experimental Evaluation

4.1 Dataset and Experimental Setup

We evaluate the framework on the openly available dataset of Gershman [28], in which 46 adults recruited online completed a two-armed bandit task. Each participant played 30 games of 10 trials, yielding 1,380 games and 13,800 choices in total. At the start of each game, the mean payoff of each arm was drawn from a Gaussian distribution with mean zero and variance 100, and each arm was independently labeled as safe or risky. A risky arm returned payoffs drawn around its mean with variance 16, whereas a safe arm returned its mean deterministically. The labels were resampled at the beginning of each game, so participants encountered all four safe–risky combinations across the session. Each game therefore represents a self-contained explore–exploit problem that begins with limited information and ends with one option typically becoming better characterized.

The design makes this dataset particularly well suited to the present research question. It was designed to manipulate two forms of uncertainty orthogonally: the difference in uncertainty between the arms, which drives uncertainty-directed choice, and the total uncertainty across arms, which influences choice stochasticity [28]. These quantities provide experimental leverage for distinguishing directed from random exploration rather than relying on model fit alone. The ten-trial game structure also provides a within-game time course with a known starting point, allowing regime occupancy to be examined as a function of trial position rather than only across the session. This temporal structure is particularly relevant to a switching framework because it allows changes in strategy to be localized within individual games. In addition, the dataset is publicly available and has been analyzed using established choice models, allowing the present comparisons to be conducted on the same choices and reproduced by other researchers. In the present analysis, choices are treated as actions, observed payoffs as rewards, and the count-based bonus as the information signal. Rewards are z-scored across the dataset before model fitting.

4.2. Baseline Models

MS-RL is compared with three baselines representing distinct accounts of within-task variation. Each baseline is either nested within the proposed framework or directly comparable using the same observed choices. The stationary softmax policy is the de facto standard in computational accounts of human exploration and represents the model that would be applied to these data by default [2], [3]. It corresponds to the $K=1$ case of MS-RL, making the comparison nested and allowing stationarity to be tested rather than assumed. The memoryless policy mixture permits multiple policies but removes their temporal order, thereby testing whether the data require temporal switching or merely multiple policies. The time-varying-temperature model represents the main continuous alternative to discrete switching, in which value sensitivity changes gradually as the task unfolds [22]. Together, these baselines distinguish stationary behavior, multiple but temporally unstructured policies, and continuous parameter drift from discrete temporal switching.

Recurrent and sequential-exploration networks have also been applied to comparable choice data and can achieve high predictive accuracy [5], [6]. They are not included as direct baselines because they do not provide the quantities central to the present research question, such as regime-specific value sensitivity, information weighting, and an interpretable transition matrix. A comparison with these models would therefore primarily address predictive accuracy rather than whether strategy switching can be identified and interpreted.

The same dataset has also been analyzed using probit choice models in which the regressors implement established accounts of uncertainty-guided choice, including estimated value difference, relative uncertainty between arms, and value scaled by total uncertainty, corresponding to upper-confidence-bound and Thompson-sampling strategies and a hybrid of the two [28]. These models provide complementary evidence that both forms of uncertainty influence choice, but they are not included as direct baselines because they model choice probabilities using uncertainty quantities from an assumed Bayesian learner and do not represent latent temporal states. Consequently, they provide neither trial-level regime assignments nor a transition matrix and cannot be nested against the switching framework in the same way as the baselines above.

The comparison set has also been extended in later work. Binz and Schulz [29] benchmarked meta-learned and resource-rational exploration algorithms against Boltzmann exploration, upper-confidence-bound algorithms, and Thompson sampling on this bandit family, while Ger et al. [10] compared a recurrent network with a stationary hybrid exploration model on the same data. These studies are relevant to the broader modeling landscape, but they address a different axis of comparison: whether a particular fixed or learned rule describes choice, rather than whether a single rule remains valid throughout a task. The baselines used here are therefore selected to test the central assumption of the present study—stationary versus temporally structured strategy selection—rather than to provide an exhaustive survey of predictive models.

4.3. Validation on Simulated Data

A latent-variable model can fit well while reporting quantities that are not identifiable. Before analyzing human choices, we therefore assessed whether the framework could recover the quantities it was designed to measure [26]. Synthetic participants were generated from a three-regime MS-RL process with known parameters, and the model was then refitted without access to the ground truth using the settings in Table 3.

Recovery was high. The generating learning rate, regime-specific policy parameters, and transition matrix were recovered with strong correspondence (parameter correlations: α_r=0.94, β_r=0.91, and γ_r=0.88). The latent regime path was recovered with 86% accuracy against a chance level of 33% (Figure 2), corresponding to substantial agreement beyond chance (Cohen's κ≈0.79). Residual errors were concentrated in the directed-exploration regime, which was the most difficult to identify because its information weight was only weakly constrained by the data. Exploitation and random exploration, which were more clearly separated by value sensitivity, were recovered more accurately. The recovered transition matrices also reproduced the high-persistence diagonal structure of the generating process, indicating that the temporal structure was identifiable rather than imposed by the model.

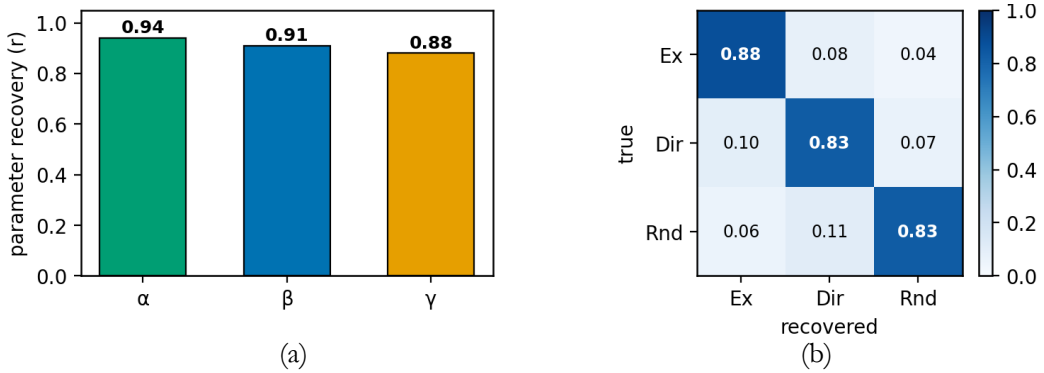

Figure 2. Parameter and state recovery from simulated data. (**a**) Parameter correlations; (**b**) regime-decoding confusion matrix.

4.4. Recovered Regimes

A three-regime MS-RL model was preferred over the two- and four-regime variants according to the Bayesian information criterion (Section 4.6). The recovered regimes were consistent with the interpretive scheme in Table 2. The exploitation regime showed high value sensitivity (β=4.98, 95% CI [4.18, 5.78]) and an information weight opposing uncertainty. The directed-exploration regime showed intermediate value sensitivity (β=1.71, 95% CI [1.42, 2.00]) and a positive information weight, whereas the random-exploration regime showed near-zero value sensitivity (β=0.18, 95% CI [0.06, 0.29]) and near-uniform choice (Figure 3). The shared learning rate was α=0.65 (95% CI [0.60, 0.70]). The three value-sensitivity intervals do not overlap, indicating clear separation among the regimes on the parameter that most strongly defines their behavioral interpretation.

The information-weight estimates reached the estimation bound of γ=±3 for all three regimes. Consequently, γ is interpreted primarily by its sign and behavioral role rather than its precise magnitude, and its standard error is not considered interpretable. This boundary behavior is expected in a standard two-armed bandit, where information value is not

manipulated strongly enough to constrain its magnitude. The low-sensitivity regime is therefore labeled random exploration; however, its near-uniform choices cannot distinguish value-insensitive sampling from attentional lapses.

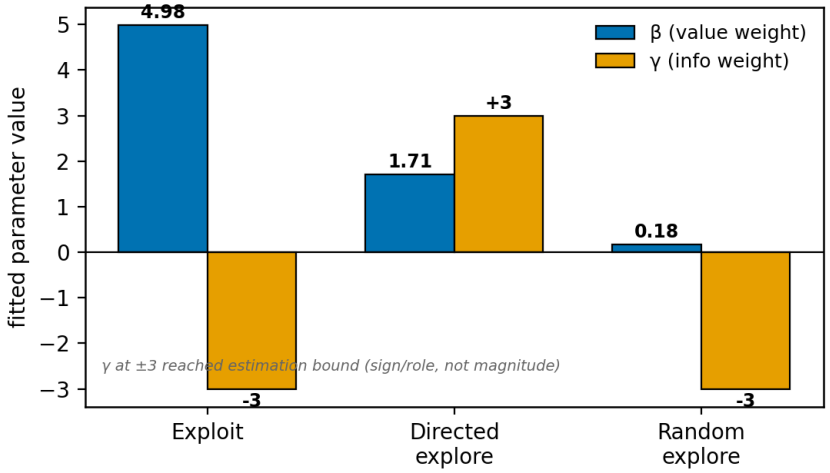

Figure 3. Recovered regime-specific policy parameters from real data.

Because the model is fitted separately to each participant, each participant has an individual transition matrix. We summarize these matrices by averaging them across participants. The resulting transition matrix showed strong diagonal persistence, with mean self-transition probabilities of 0.96 for exploitation, 0.69 for directed exploration, and 0.86 for random exploration (Figure 4). The regimes were therefore persistent, with decision makers remaining within a strategy for extended runs rather than re-randomizing on every trial. Exploitation was both the most persistent regime and the most common destination, consistent with a process that settles into value maximization once a good option has been identified.

This pattern is consistent with other latent-state accounts of choice, in which mice and humans alternate among discrete strategies that persist over multiple trials rather than switching randomly [20]. The high persistence of exploitation (0.96) is also consistent with sustained, goal-directed engagement described as proactive control in dual-mechanisms accounts of cognitive control [30], whereas the lower persistence of directed exploration (0.69) is consistent with its role as a transient information-gathering state that becomes less necessary as uncertainty is reduced.

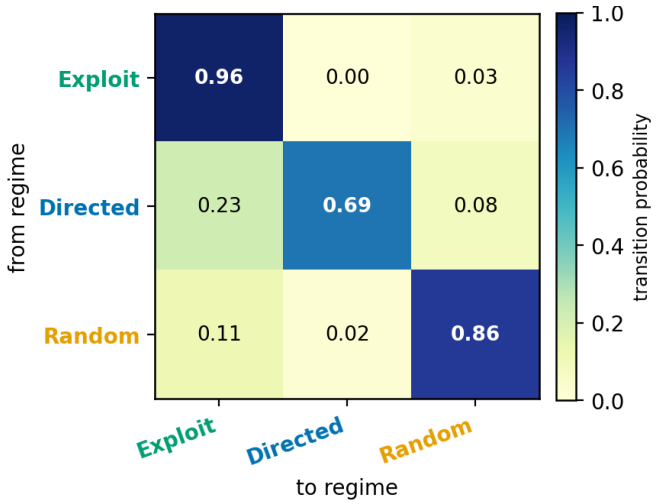

Figure 4. Participant-averaged regime transition matrix showing regime persistence and switching.

4.5. Within-Game Strategy Dynamics

Decoding the most likely regime for every trial revealed a systematic within-game progression (Figure 5). The first trial of each game was occupied almost entirely by the directed-

exploration regime, consistent with the rational opening move when both options are unknown. Across successive trials, occupancy of the exploitation regime rose monotonically from roughly zero on trial 1 to 82% by trial 10, while both exploratory regimes declined. Averaged across all trials, participants spent 64% of the time exploiting, 28% in directed exploration, and 8% in random exploration.

MS-RL therefore localizes the explore-to-exploit transition to a specific decoded quantity: the changing occupancy of latent regimes over time, rather than a change absorbed into a static temperature. A stationary model must represent every trial with the same policy and cannot express this temporal progression. As shown in Figure 5, the trajectory tracks the resolution of uncertainty. At the opening of a game, both options are unknown, and information is most valuable, so the policy that samples for information dominates. As outcomes accumulate, the uncertainty that initially justified exploration is reduced, and control shifts toward the policy that exploits what has been learned. On this interpretation, the within-game shift is not a decay of engagement but a reallocation of decision strategy that follows the information available at each point in the game.

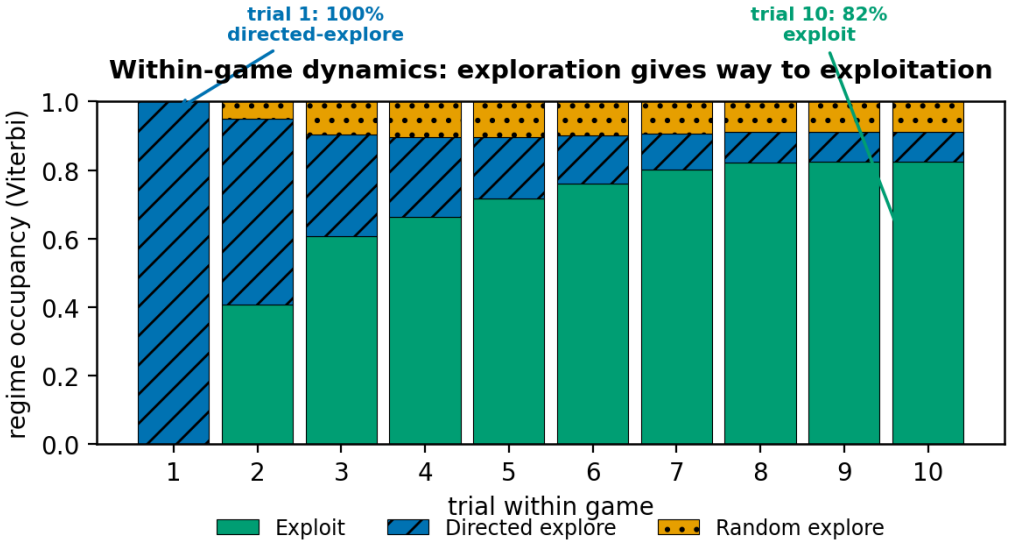

Figure 5. Within-game regime occupancy by trial position (Viterbi-decoded, real data).

4.6. Model Comparison and Ablation Analysis

The critical test is whether temporal switching earns its additional complexity. Across the full dataset, the three-regime MS-RL model achieved the lowest Bayesian information criterion among the candidate models: BIC = 12,295, compared with 12,766 for two-regime MS-RL, 12,347 for four-regime MS-RL, 13,095 for the memoryless mixture, and 13,259 for the stationary single-policy model (Figure 6). A single-regime model with smoothly time-varying value sensitivity was also fitted to distinguish discrete switching from gradual drift. This model performed no better than the stationary model (BIC = 13,270), with a negligible estimated decay.

The comparison with the memoryless mixture is particularly informative because both models contain three policies, while only the full MS-RL model includes temporal dependence between regimes. The improvement of approximately 800 BIC points therefore indicates that the advantage of MS-RL is not attributable simply to having multiple policies, but to modeling how those policies switch over time. Similarly, the preference for K=3 over K=4 indicates that the additional regime does not provide sufficient improvement to justify its complexity. The comparison against the stationary model further supports the need for a non-stationary policy: McFadden pseudo-R^2 increased from 0.31 for the stationary policy to 0.37 for three-regime MS-RL, while a likelihood-ratio test against the nested stationary special case was decisive (χ^2 =1078, 12 degrees of freedom), corresponding to an improvement of approximately 0.04 in per-choice log-likelihood.

These comparisons also provide an ablation of the major components of the framework. Removing the latent regimes by setting $K=1$ produces the largest degradation (964 BIC points), indicating that a single policy does not adequately represent the observed choices. Removing temporal dependence while retaining the three policies produces the memoryless mixture and costs approximately 800 BIC points, isolating the contribution of temporal switching. Replacing discrete switching with continuous drift costs 975 BIC points and yields a negligible decay, indicating that the observed strategy changes are not adequately captured by a smooth change in value sensitivity. Finally, reducing the model to two regimes costs 471 BIC points, whereas adding a fourth regime costs 52 points. Together, these ablations indicate that both the latent regime structure and its temporal dynamics contribute substantially to model fit, while three regimes provide the best balance between flexibility and complexity. As shown in Figure 6, the three-regime MS-RL model is preferred over the stationary, memoryless-mixture, time-varying, and alternative-regime models, while the ablation comparisons clarify which components account for this advantage.

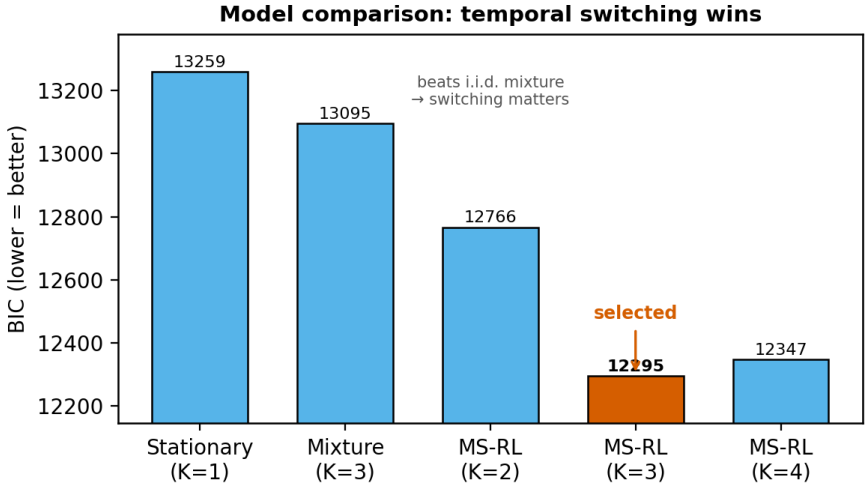

Figure 6. BIC comparison of the candidate models (lower is better).

5. Discussion

Across the validation study on simulated data and the application to open human choices, MS-RL supports a simple thesis: exploration is better described as a state that a decision maker passes through than as a setting held fixed throughout a task. This view is consistent with a broader psychological perspective in which behavior emerges from temporally extended states rather than invariant decision parameters. Cognitive control, for example, is increasingly understood as a dynamic process that varies with task demands and internal states rather than as a fixed capacity applied uniformly [30]. In dual-mechanisms accounts, behavior is supported by different control modes depending on goals and demands, implying that processing need not be governed by a single stable mode over time. The present results extend this perspective to exploration.

5.1. Contributions and Relation to Prior Models

The model addresses the exploration-versus-noise ambiguity that has limited the interpretation of bandit choices [2] by assigning non-greedy choices to identifiable regimes (Figure 3) and providing a transition matrix (Figure 4), a quantity absent from stationary accounts. A stationary softmax model [2], [3] represents the entire task with a single inverse temperature and therefore cannot determine whether non-greedy choices reflect information seeking or random responding, nor whether exploration is concentrated early and exploitation later. A memoryless mixture can represent several policies but, without temporal structure, cannot describe their persistence or systematic changes across trials. Similarly, a single-regime model with smoothly time-varying value sensitivity can represent gradual parameter drift, but the model comparison indicates that it does not capture the discrete structure supported by the data.

MS-RL combines these capabilities by estimating which regime generates each choice, how long regimes persist, and how their occupancy changes within a game (Figures 4 and 5). Recurrent-network and sequential-exploration models [5], [6] recover latent exploration processes, but they do not separate a switching action-selection policy from a shared learning process. This distinction allows the transition matrix in MS-RL to be interpreted as a measure of strategy change rather than as a consequence of changes in learning.

The recovered regimes are also consistent with independently motivated behavioral and neural constructs. Exploitation and exploration are neurally dissociable [1], directed and random exploration are causally separable [12], and explore–exploit transitions have been linked to hypothesized neuromodulatory switching processes [13]. The estimated transition matrix is therefore consistent with such a switching account, but it should not be interpreted as a direct estimate of any neuromodulatory variable. The adaptive-gain account describes a continuous gain signal, whereas the present model represents behavioral strategy using discrete latent states. The mapping is consequently a motivated and falsifiable interpretation rather than a mechanistic identification.

The within-game transition from directed exploration toward exploitation further illustrates the explanatory value of this representation. Rather than interpreting the change as a gradual cooling of a single decision temperature, the model represents it as a reallocation among strategies as information accumulates. This pattern is consistent with accounts in which exploration serves to reduce uncertainty and is reduced as uncertainty is resolved [17]. The relatively lower persistence of directed exploration is also consistent with its interpretation as a transient information-gathering strategy rather than a stable preference for exploration.

5.2. Practical Implications

The framework has several implications for the analysis and application of sequential choice. Because stationary softmax-RL is the K=1 special case of MS-RL, researchers can test for strategy switching rather than assuming stationarity. When switching is supported, the decoded regime path provides participant-level measures such as time to first exploitation, switching frequency, and opening-trial exploration. This localization can be particularly useful in clinical and developmental research, where differences in exploration may arise from reduced directed exploration, increased random exploration, or altered switching dynamics rather than from a single temperature change. Populations with documented exploratory differences, such as attention-deficit/hyperactivity disorder, depression, or healthy aging, could therefore be examined in terms of which regimes and transitions differ.

The transition matrix may also provide a behavioral measure of cognitive flexibility by quantifying individual differences in regime persistence and switching propensity, connecting the framework to research on executive function [31]. The same information could support adaptive systems. A system that can distinguish information-seeking from preference exploitation could time interventions accordingly—for example, delaying new options while a user is consolidating a preference and presenting alternatives when the user is already exploring. Similarly, an instructional system could distinguish systematic hypothesis testing from random guessing, which may require different interventions. The weak identification of the uncertainty bonus also has a practical implication for experimental design. Paradigms that manipulate information value parametrically, such as the Horizon Task, may identify the directed-exploration regime more precisely than a standard bandit.

5.3. Limitations

Several limitations constrain the interpretation of the findings. First, the regimes are inferred rather than directly observed. Their mechanistic interpretation therefore requires converging neural or physiological evidence, and the psychological labels assigned to the regimes should be treated as hypotheses rather than direct observations. In particular, the random-exploration regime may reflect value-insensitive sampling or attentional disengagement, which can produce similar behavioral patterns [32]. Pupillometry provides a potentially discriminating test: because pupil size in this paradigm tracks total rather than relative uncertainty, and total uncertainty drives random rather than directed exploration [33], pupil dilation should be more closely associated with the decoded random-exploration regime than with

directed exploration if the proposed interpretation is correct. Such evidence would provide a stronger test of the psychological meaning of the latent states.

Second, the model assumes that value learning is shared across regimes. If learning also changes across states, the current specification could attribute some learning variation to action-selection dynamics. The information signal is also only an approximation to the uncertainty manipulated by the task. A count-based bonus decreases with sampling, whereas uncertainty also depends on whether an arm is safe or risky: a risky arm retains residual uncertainty even after repeated choices, whereas a safe arm becomes known after a single observation. This mismatch may contribute to the information weights reaching their estimation bounds, leaving the magnitude of γ weakly identified. Its consistent sign and the early dominance of directed exploration nevertheless support its interpretation as an information-seeking component. Replacing the count-based signal with posterior variance from a Bayesian learner, together with paradigms that manipulate information value parametrically such as the Horizon Task [2], would provide a stronger test.

Third, the observed within-game shift toward exploitation has an alternative explanation based on the horizon effect. Because each game contains only ten trials, participants may anticipate the end of the game and exploit for that reason rather than solely because uncertainty has been reduced. Designs that vary the number of remaining choices [2] would help distinguish these explanations.

Finally, the framework assumes discrete strategy changes. This assumption makes the transition matrix interpretable, but a genuinely gradual process would be represented as a sequence of transitions between fixed regimes. The ablation against the drifting-temperature model (Section 4.6) shows that discrete switching fits these data better than smooth drift alone, but it does not establish that continuous variation is absent. A hybrid model in which regime-specific policies can also drift would provide a more direct test. The framework should likewise be evaluated in richer settings, including n-armed or restless bandits, where the best option can change, and Markov decision processes, where exploration operates across states rather than only among arms.

5.4. Scientific Insights and Trade-offs

Three broader insights emerge from the analysis. First, the switching representation changes how the explore–exploit transition can be interpreted. The decoded regime path shows that the transition is not simply a gradual cooling of a single temperature but a reallocation among discrete strategies: directed exploration dominates the opening stage and gives way to exploitation as information accumulates, while random exploration remains a minority state. A stationary temperature averages over this process and therefore cannot represent the strategy occupying any particular trial.

Second, the combination of recovery and model comparison provides evidence that the switching structure is not merely an artifact of model flexibility. Parameter and state recovery establish that the latent quantities can be recovered from known generating processes, while the preference for $K=3$ over $K=4$ and the improvement over the memoryless mixture provide complementary evidence from human choices. Together, these results support interpreting the transition matrix as a measure of behavioral strategy change rather than as an unlabeled state sequence produced by overfitting.

Third, the findings extend a broader pattern in which discrete, persistent strategy states have been recovered across species and tasks. Mice and humans in perceptual decisions alternate among a small number of states that persist for tens of trials [20], while exploration in restless bandits has also been modeled as a latent process recoverable from behavior [6]. MS-RL extends this perspective to value-based exploration by allowing the action-selection policy to switch among interpretable regimes while keeping value learning shared. This separation addresses an important question: whether observed switching can be explained by changes in how agents select actions without requiring changes in how they learn. Within the present model class and dataset, the results indicate that switching in the action-selection policy is sufficient to capture the observed dynamics. This does not imply that learning-rate switching is absent or impossible; where future data support such variation, the framework could be extended to allow learning rates to vary by regime, although this would introduce additional identifiability challenges.

These advantages come with trade-offs. MS-RL is more computationally demanding than a stationary policy because EM with multiple random restarts is applied to a multimodal

likelihood, and sufficient trials per participant are needed to estimate transition dynamics reliably. The model also gains interpretability by holding value learning fixed, but this assumption becomes restrictive when learning and action-selection strategies change simultaneously. Finally, experimental design determines which aspects of the model can be identified. The present bandit separates value sensitivity relatively well but leaves the information weight at its estimation bound. Paradigms that manipulate information value parametrically could sharpen identification of directed exploration at the cost of greater experimental complexity. These boundaries suggest that MS-RL is most informative when the dataset contains a sufficient number of trials per participant, the research question concerns action-selection strategy, and the experimental design provides enough variation in uncertainty to distinguish directed from random exploration.

6. Conclusions

This study examined whether exploration in sequential choice is better represented as a fixed policy or as a strategy that changes over time. The proposed Markov-Switching Reinforcement-Learning (MS-RL) framework addresses this question by allowing the action-selection policy to switch among latent regimes while keeping value learning shared. The validation study showed that the model can recover its generating parameters and latent regime paths, while the empirical analysis showed that the three-regime formulation provides a better account of human choices than stationary, memoryless-mixture, time-varying, and alternative-regime models.

The results further show that the latent regimes provide an interpretable representation of within-task strategy change. Directed exploration dominates the early stage of a game and progressively gives way to exploitation as information accumulates, while random exploration remains a smaller component. The transition matrix additionally captures how persistently participants remain in each strategy and how they move between strategies. These results support the central objective of the study: within-task variation in exploration can be modeled as structured changes in action-selection strategy rather than being absorbed into stationary decision noise.

The findings also clarify the scope of the contribution. Within the present model class and dataset, switching in the action-selection policy is sufficient to account for the observed dynamics without requiring the learning rate itself to change. At the same time, the interpretation of the latent regimes remains behavioral rather than mechanistically established, and the limited identification of the information weight and the two-armed bandit setting constrain generalization. Future work should therefore validate the decoded regimes against independent neural or physiological measures, evaluate the framework in richer sequential decision-making environments such as restless or n-armed bandits and Markov decision processes, and investigate regime-dependent learning rates when sufficient data are available. Overall, the results support treating exploration as a dynamic state that can be identified, tracked, and analyzed across sequential decisions rather than as a fixed setting of a stationary policy.

Author Contributions: F.A.; Methodology: F.A.; Software: F.A. and N.R.; Validation: F.A., F.F.K., and N.R.; Formal analysis: F.A. and N.R.; Investigation: F.A.; Data curation: N.R.; Writing—original draft preparation: F.A.; Writing—review and editing: F.F.K. and N.R.; Visualization: F.A. and N.R.; Supervision: F.A. In detail, F.A. developed the mathematical model and led the drafting; F.F.K. contributed to the psychological interpretation and literature review and carried out draft revision and editing; N.R. led the analytics and contributed to revision and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: This study analyzed an openly available dataset and generated no new data. The two-armed bandit dataset from Gershman (2019), comprising 46 adult participants who completed 30 games of 10 trials across four uncertainty conditions (13,800 choices in total), is publicly available from the original author's repository at https://github.com/sjgershm/exploration_uncertainty. The model code, Expectation-

Maximization estimation routine, parameter- and state-recovery pipeline, and analysis scripts used to reproduce all figures are available from the corresponding author upon reasonable request.

Acknowledgments: The authors thank colleagues at the School of Computer Science, Bina Nusantara University, for the continuous support and helpful discussions. The authors used a large language model to assist with language editing and formatting of the manuscript; all scientific content, analyses, and conclusions are the authors' own, and the authors take full responsibility for the integrity of the work.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] N. D. Daw, J. P. O'Doherty, P. Dayan, B. Seymour, and R. J. Dolan, "Cortical substrates for exploratory decisions in humans," *Nature*, vol. 441, no. 7095, pp. 876–879, Jun. 2006, doi: 10.1038/nature04766.

[2] R. C. Wilson, A. Geana, J. M. White, E. A. Ludvig, and J. D. Cohen, "Humans use directed and random exploration to solve the explore–exploit dilemma.," *J. Exp. Psychol. Gen.*, vol. 143, no. 6, pp. 2074–2081, 2014, doi: 10.1037/a0038199.

[3] E. Schulz and S. J. Gershman, "The algorithmic architecture of exploration in the human brain," *Curr. Opin. Neurobiol.*, vol. 55, pp. 7–14, Apr. 2019, doi: 10.1016/j.conb.2018.11.003.

[4] M. M. Botvinick, T. S. Braver, D. M. Barch, C. S. Carter, and J. D. Cohen, "Conflict monitoring and cognitive control.," *Psychol. Rev.*, vol. 108, no. 3, pp. 624–652, 2001, doi: 10.1037/0033-295X.108.3.624.

[5] R. Ligneul, "Sequential exploration in the Iowa gambling task: Validation of a new computational model in a large dataset of young and old healthy participants," *PLOS Comput. Biol.*, vol. 15, no. 6, p. e1006989, Jun. 2019, doi: 10.1371/journal.pcbi.1006989.

[6] D. Tuzsus, A. Brands, I. Pappas, and J. Peters, "Exploration–Exploitation Mechanisms in Recurrent Neural Networks and Human Learners in Restless Bandit Problems," *Comput. Brain Behav.*, vol. 7, no. 3, pp. 314–356, Sep. 2024, doi: 10.1007/s42113-024-00202-y.

[7] S. J. Gershman, "Deconstructing the human algorithms for exploration," *Cognition*, vol. 173, pp. 34–42, Apr. 2018, doi: 10.1016/j.cognition.2017.12.014.

[8] M. Speekenbrink and E. Konstantinidis, "Uncertainty and Exploration in a Restless Bandit Problem," *Top. Cogn. Sci.*, vol. 7, no. 2, pp. 351–367, Apr. 2015, doi: 10.1111/tops.12145.

[9] N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, and R. J. Dolan, "Model-Based Influences on Humans' Choices and Striatal Prediction Errors," *Neuron*, vol. 69, no. 6, pp. 1204–1215, Mar. 2011, doi: 10.1016/j.neuron.2011.02.027.

[10] Y. Ger, E. Nachmani, L. Wolf, and N. Shahar, "Harnessing the flexibility of neural networks to predict dynamic theoretical parameters underlying human choice behavior," *PLOS Comput. Biol.*, vol. 20, no. 1, p. e1011678, Jan. 2024, doi: 10.1371/journal.pcbi.1011678.

[11] R. Schurr, D. Reznik, H. Hillman, R. Bhui, and S. J. Gershman, "Dynamic computational phenotyping of human cognition," *Nat. Hum. Behav.*, vol. 8, no. 5, pp. 917–931, Feb. 2024, doi: 10.1038/s41562-024-01814-x.

[12] W. K. Zajkowski, M. Kossut, and R. C. Wilson, "A causal role for right frontopolar cortex in directed, but not random, exploration," *Elife*, vol. 6, p. e27430, Sep. 2017, doi: 10.7554/eLife.27430.

[13] G. Aston-Jones and J. D. Cohen, "An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance," *Annu. Rev. Neurosci.*, vol. 28, no. 1, pp. 403–450, Jul. 2005, doi: 10.1146/annurev.neuro.28.061604.135709.

[14] M. Jepma and S. Nieuwenhuis, "Pupil Diameter Predicts Changes in the Exploration–Exploitation Trade-off: Evidence for the Adaptive Gain Theory," *J. Cogn. Neurosci.*, vol. 23, no. 7, pp. 1587–1596, Jul. 2011, doi: 10.1162/jocn.2010.21548.

[15] C. S. Chen, D. Mueller, E. Knep, R. B. Ebitz, and N. M. Grissom, "Dopamine and Norepinephrine Differentially Mediate the Exploration–Exploitation Tradeoff," *J. Neurosci.*, vol. 44, no. 44, p. e1194232024, Oct. 2024, doi: 10.1523/JNEUROSCI.1194-23.2024.

[16] J. S. B. T. Evans and K. E. Stanovich, "Dual-process theories of higher cognition: Advancing the debate," *Perspect. Psychol. Sci.*, vol. 8, no. 3, pp. 223–241, May 2013, doi: 10.1177/1745691612460685.

[17] J. B. Hirsh, R. A. Mar, and J. B. Peterson, "Psychological entropy: A framework for understanding uncertainty-related anxiety.," *Psychol. Rev.*, vol. 119, no. 2, pp. 304–320, Apr. 2012, doi: 10.1037/a0026767.

[18] A. Modirshanechi, W.-H. Lin, H. A. Xu, M. H. Herzog, and W. Gerstner, "Novelty as a drive of human exploration in complex stochastic environments," *Proc. Natl. Acad. Sci.*, vol. 122, no. 39, Sep. 2025, doi: 10.1073/pnas.2502193122.

[19] M. S. Tomov, V. Q. Truong, R. A. Hundia, and S. J. Gershman, "Dissociable neural correlates of uncertainty underlie different exploration strategies," *Nat. Commun.*, vol. 11, no. 1, p. 2371, May 2020, doi: 10.1038/s41467-020-15766-z.

[20] Z. C. Ashwood *et al.*, "Mice alternate between discrete strategies during perceptual decision-making," *Nat. Neurosci.*, vol. 25, no. 2, pp. 201–212, Feb. 2022, doi: 10.1038/s41593-021-01007-z.

[21] Š. Kucharský, N.-H. Tran, K. Veldkamp, M. Raijmakers, and I. Visser, "Hidden Markov Models of Evidence Accumulation in Speeded Decision Tasks," *Comput. Brain Behav.*, vol. 4, no. 4, pp. 416–441, Dec. 2021, doi: 10.1007/s42113-021-00115-0.

[22] M. Jepma, J. V Schaaf, I. Visser, and H. M. Huizenga, "Uncertainty-driven regulation of learning and exploration in adolescents: A computational account," *PLOS Comput. Biol.*, vol. 16, no. 9, p. e1008276, Sep. 2020, doi: 10.1371/journal.pcbi.1008276.

[23] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html

[24] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989, doi: 10.1109/5.18626.

[25] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer New York, NY, 2006. [Online]. Available: https://link.springer.com/book/9780387310732
[26] R. C. Wilson and A. G. Collins, "Ten simple rules for the computational modeling of behavioral data," *Elife*, vol. 8, p. e49547, Nov. 2019, doi: 10.7554/eLife.49547.
[27] P. L. Lockwood and M. C. Klein-Flügge, "Computational modelling of social cognition and behaviour—a reinforcement learning primer," *Soc. Cogn. Affect. Neurosci.*, vol. 16, no. 8, pp. 761–771, Mar. 2020, doi: 10.1093/scan/nsaa040.
[28] S. J. Gershman, "Uncertainty and exploration.," *Decision*, vol. 6, no. 3, pp. 277–286, Jul. 2019, doi: 10.1037/dec0000101.
[29] M. Binz and E. Schulz, "Modeling Human Exploration Through Resource-Rational Reinforcement Learning," in *Advances in Neural Information Processing Systems 35*, 2022, vol. 35, pp. 31755–31768. doi: 10.52202/068431-2302.
[30] T. S. Braver, "The variable nature of cognitive control: a dual mechanisms framework," *Trends Cogn. Sci.*, vol. 16, no. 2, pp. 106–113, Feb. 2012, doi: 10.1016/j.tics.2011.12.010.
[31] A. Diamond, "Executive Functions," *Annu. Rev. Psychol.*, vol. 64, no. 1, pp. 135–168, Jan. 2013, doi: 10.1146/annurev-psych-113011-143750.
[32] M. Mittner, G. E. Hawkins, W. Boekel, and B. U. Forstmann, "A Neural Model of Mind Wandering," *Trends Cogn. Sci.*, vol. 20, no. 8, pp. 570–578, Aug. 2016, doi: 10.1016/j.tics.2016.06.004.
[33] H. Fan, T. Burke, D. C. Sambrano, E. Dial, E. A. Phelps, and S. J. Gershman, "Pupil Size Encodes Uncertainty during Exploration," *J. Cogn. Neurosci.*, vol. 35, no. 9, pp. 1508–1520, Sep. 2023, doi: 10.1162/jocn_a_02025.