Integrating Local Learning into the Two-Stage Markov Task to Separate Model-Based from Model-Free Learning

Peizhe Li[†], Jimmy Vineyard[‡], Seungyeon Oh[§], and Jack Maloney[†]

Project advisors: Haley Colgate Kottler[¶]and Amy Cochran[‖]

Abstract. The two-stage Markov task, widely-used for measuring model-based relative to model-free learning in humans, has faced skepticism regarding its effectiveness. We suggest a modification to better distinguish the two learning approaches. Our revised task incorporates an additional phase for learning local contingencies, mirroring a strategy from machine learning for separating model-free from model-based algorithms. We evaluated the effectiveness of our revised task through simulations, employing model-free and model-based strategies.

Key words: model-free learning, model-based learning, two-stage Markov task, human decision-making, machine learning

MSC: 91E40, 68Q32, 68T05, 90C40

1. Introduction.

Reinforcement learning and psychology communities are interested in distinguishing between model-free (MF) and model-based (MB) learning [3, 26]. With decision making tasks, MB learning refers to methods that require a model for the transition probabilities of the environment, while MF learning methods do not need such a model [25]. Note that these definitions make the classification of MB or MF mutually exclusive and exhaustive. [3] proposed a two-stage Markov task to evaluate individual's inclination for MF vs. MB learning. It has been widely used [1, 4, 5, 6, 7, 9, 13, 16, 17, 18, 19, 20, 21, 23, 24, 27, 30]. However, it has recently been criticized for not performing its intended function.

In the two-stage task, participants initially choose an action that transitions them to a second stage via common or rare transitions. They make another choice that results in a probabilistic reward. By analyzing participants' likelihood of repeating their first-stage choice based on reward outcomes and transition types, researchers aim to measure individual's inclination for MF and MB learning [3]. Daw et al propose that for MF learning, the chance of repeating a choice will be based solely on reward, while for MB learners the chance of repeating a choice will be based solely on if the transition was rare or common. However, a purely MF learner can demonstrate either pattern when risk-seeking and risk-sensitivity are incorporated into their exploration strategy [8]. The behavior pattern also varies with how much participants misconstrue the task based on the directions they are given [11, 12]. Moreover, minor task modifications can alter the characteristic behavior of established MB and MF algorithms, making results difficult to interpret [10]. These observations prompt us to determine if simple modifications to the task could improve its function.

[†] University of Wisconsin - Madison, Madison, WI

[‡] corresponding author, University of Wisconsin - Madison, Madison, WI (jvineyard@wisc.edu),

[§] University of Wisconsin - Madison, Madison, WI (oh.653@osu.edu),

[¶] Department of Mathematics, University of Wisconsin - Madison, Madison, WI (haley.kottler@wisc.edu)

[‖] Department of Mathematics, Department of Population Health Sciences, University of Wisconsin - Madison, Madison, WI (cochran4@wisc.edu)

Previous approaches to refining the two-stage Markov task, such as simulation-based validations [2, 14, 28] or task complexity adjustments[15], have attempted to address these issues but often lack a clear mechanism for separating learning strategies. Inspired by machine learning's progress towards separating MB from MF algorithms, we propose adapting local change adaptation (LoCA)[26] for use in human tasks. This approach leverages principles akin to performance metrics in machine learning, such as precision and recall, to systematically differentiate the contributions of each strategy [22].By addressing these criticisms, our framework offers a novel solution that improves the robustness and discriminative power of the task compared to prior efforts. We analyze our proposed task in simulation.

2. Background. Decision-making experiments can often be described as discrete-time Markov Decision Processes (MDP). A discrete-time MDP consists of states, actions, rewards, and a transition distribution. At each time step t, an agent observes state s_t, selects action a_t, and transitions to a new state s_{t+1} while receiving reward r_t according to the transition distribution. This process repeats over a time horizon T.

We describe the two-stage Markov task from [3] as an MDP with 3 states, $s_t \in \{1, 2, 3\}$, and two possible actions $a_t \in \{1, 2\}$. The agent starts in state 1 ($s_1 = 1$). After selecting action a_1, the agent transitions to state 2 or 3, with no immediate rewards ($r_1 = 0$). If $a_t = 1$, there is a 70% chance of transitioning to state 2, termed as a *common* transition, and a 30% chance of transitioning to state 3, termed as a *rare* transition. If $a_1 = 2$, the probabilities and transition type labels reverse. After selecting action a_2, the agent receives reward r_2. The decision-making scenario then repeats, represented as a deterministic transition back to state 1. We use the 70/30 split originally proposed by Daw et al. [2]. Modifying this split affects how distinct model-based and model-free behaviors appear; more extreme splits (e.g., 90/10) may inflate model-based behavior, while near-uniform splits (e.g., 60/40) may obscure differences (see Appendix A.1).

Recent machine learning research emphasizes that a defining trait of MB learning is its capacity to modify its policy across all states when encountering a local change [26, 29]. This inspired the LoCA task, where agents navigate a finite rectangular grid. Each grid cell represents a state in an MDP with deterministic transitions, and available moves correspond to actions. There are two additional states, T_1 and T_2, located outside the grid at the left and right boundaries respectively. When the agent is near the leftmost edge of the grid, known as the *event horizon*, it is forced to navigate only toward T_1 and cannot move away from it. Rewards are given only when the agent transitions to states T_1 or T_2, after which they transition back to a state in the grid.

The task comprises three learning phases. In Phase I, the agent is initialized at any grid state and transitions back to any grid state after receiving a reward. In this phase, rewards are greater at T_1 compared to T_2. In Phase II rewards are smaller at T_1, and the agent is initialized and transitioned back to within the event horizon, restricting it to navigate near T_1. In the final phase, Phase III, the rewards remain the same as in Phase II. However, the agent is initialized at any grid state and transitioned to any grid state after receiving a reward.

Performance is assessed by the fraction of instances it reaches state T_2 instead of T_1 during Phase III. Effective use of MB learning allows the agent to adapt to the altered rewards in Phase II and integrate this locally obtained knowledge into its policy across all states, facilitating successful navigation towards the higher rewards at T_2 at the start of Phase III. Without using MB learning, the agent needs to re-explore the state space in Phase III, just as it does in Phase I, in order to effectively learn to navigate to T_2 rather than T_1.

3. Methods.

We propose a task that captures the core elements of the LoCA task within the general structure of the two-stage Markov task (see Figure 1). This new task can be described by an MDP with identical states, actions, and transition types as the two-stage Markov task. However, we also divide the task into three phases following the ideas of the LoCA task.

In Phase I, the agent starts in state 1, receives a reward after taking an action in state 2 or 3, and transitions back to state 1. In Phase II, the agent starts in state 2, receives a reward after taking an action, and then transitions back to state 2. In Phase III, transitions revert to those of Phase I. Each phase continues until a given number of visits to states 2 and 3. During Phase I, average rewards are designed to be higher in state 2 than in state 3, establishing an initial preference for actions leading to state 2. We then modify the reward distribution for Phases II and III so average rewards are higher in state 3 than in state 2. As in the LoCA task, we can evaluate an individual's inclination for MB and MF learning by measuring their performance at the start of Phase III.

We developed a simulation of our task, accessible at https://github.com/jwvineyard/mxm_sp23-Learning. We fixed the number of visits to states 2 or 3 at 50 for each phase. For simplicity, every choice in state 2 would lead to a reward drawn from a normal distribution with mean 4 and variance 1 and every choice in state 3 would lead to a reward drawn from a normal distribution with mean 2 and variance 1, regardless of the action chosen. In Phases II and III, choices in state 2 would always result in a reward drawn from a normal distribution with mean 1 and variance 1, while the rewards for state 3 remained the same.

We ran this simulation 10,000 times, applying MF and MB algorithms. For our MF algorithm, we used Q-learning with an ε-greedy action selection, as presented in [25]. For our MB algorithm, we tracked the Q-values at each step using the same update as our MF algorithm, and followed the action selection procedure used for the MB algorithm in the original two-stage Markov task paper introduced by Daw et al [3]. In states 2 and 3, we use the same ε-greedy action selection as the MF algorithm. In state 1, with probability ε we select a random action. With probability $1-\varepsilon$, the action was chosen to maximize

$$Q_{MB}(a) = P(s_{t+1} = 2 | s_t = 1, a) \max_{a'} Q(2, a') + P(s_{t+1} = 3 | s_t = 1, a) \max_{a'} Q(3, a') \tag{3.1}$$

where $Q(s,a)$ is the Q value for state s and action a, and $P(s_{t+1} = S' | s_t = S, a)$ is the probability of transitioning from state S to state S' with action a. For both algorithms we used a learning rate of 0.5, a discount rate of 0.1, and $\varepsilon = 0.1$. We then calculated the average reward obtained from each choice in state 2 or 3 in both algorithms. We included 95% Wald confidence intervals for these averages.

4. Results.

Results are shown in Figure 2. In Phase I, the MB algorithm improved slightly quicker than the MF algorithm. However, adjusting parameters could potentially reverse this difference in early performance. What is more striking is the initial performance of MB during Phase III, which occurs as a direct consequence of incorporating locally learned

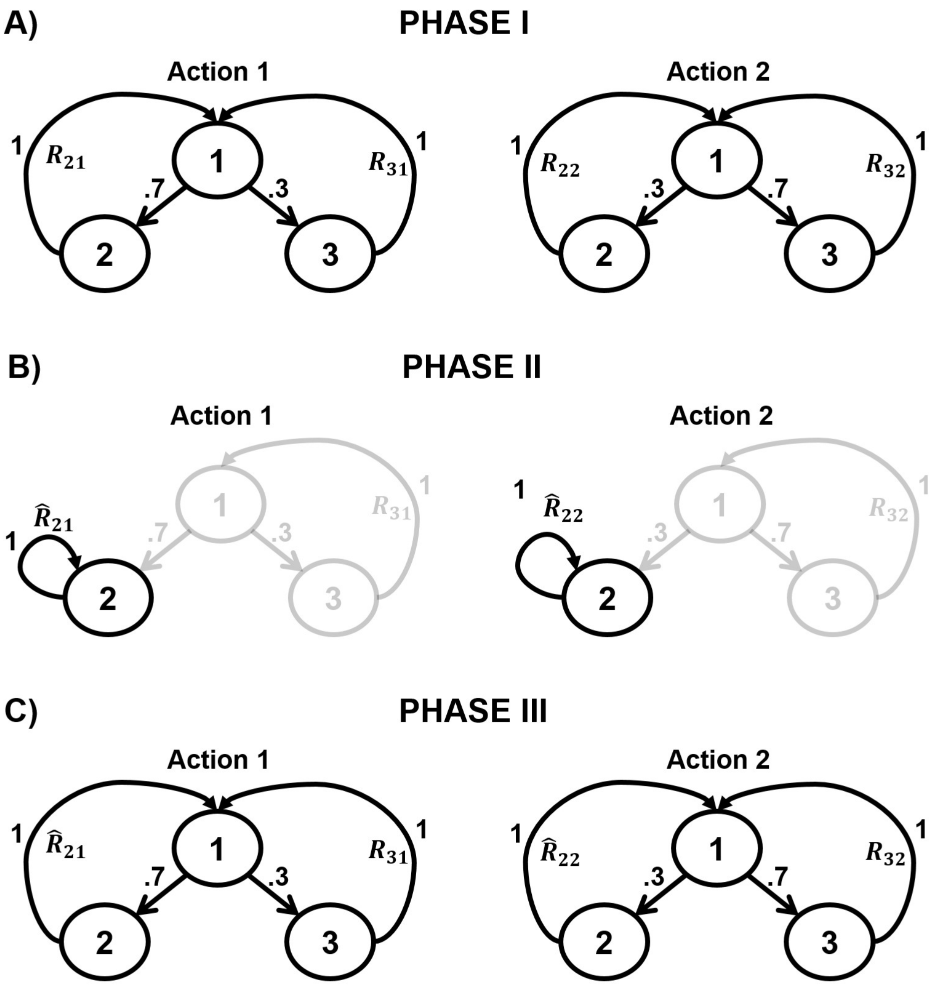

Figure 1. *The proposed task's three phases are illustrated as follows:* ***A)*** *The first phase mirrors the two-stage decision-making scenario of the two-stage Markov task.* ***B)*** *In the second phase, the emphasis shifts to locally learning the modified reward distributions associated with actions taken in state 2 ($\hat{R}_{21}$ and $\hat{R}_{22}$).* ***C)*** *The third phase repeats the two-stage decision-making scenario but uses the modified reward distribution from Phase II. Reward distributions are designed such that state 2 can yield higher average rewards than state 3 in Phase I, but lower average rewards in Phases II and III. Effective MB learning involves swiftly integrating the local knowledge gained from Phase II into a policy for Phase III, enabling navigation towards the higher average rewards. States are represented by circles, and state transitions for each action are depicted by directed arrows, with the probability of each transition shown next to the arrow. The corresponding reward distribution is also displayed next to each arrow and labeled with an uppercase R.*

reward information from Phase II and also an effect not seen in the MF learner, which must re-explore. The MB algorithm starts Phase III close to optimally, extrapolating insights gained from learning local rewards in Phase II to enhance its policy in Phase III, while the MF algorithm must first explore the new space.

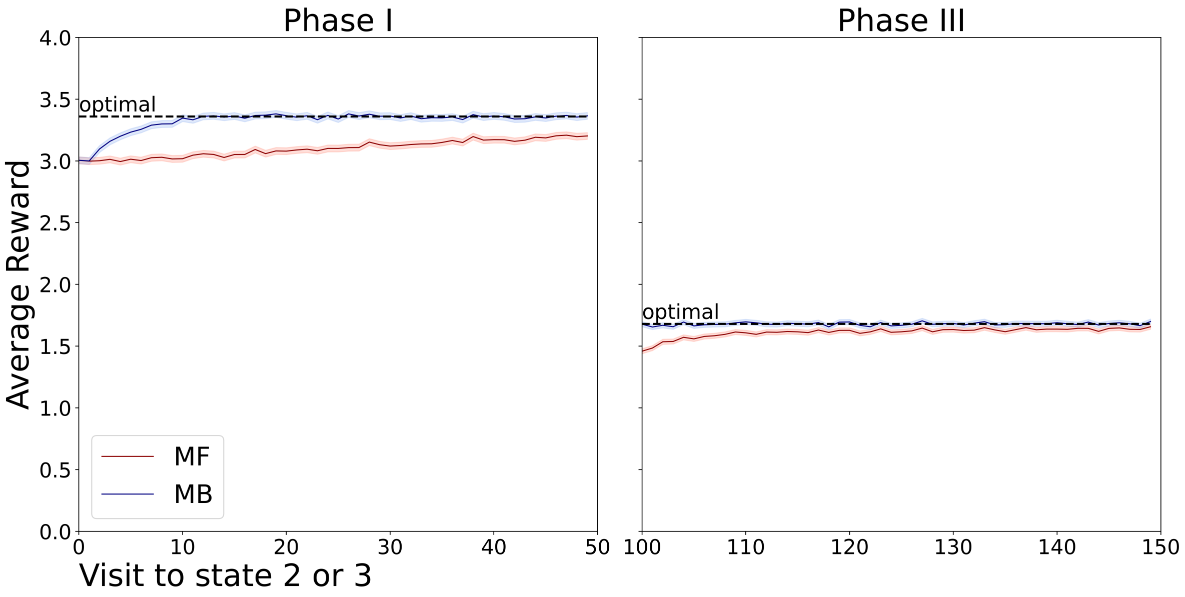

Figure 2. *Average reward, along with 95% confidence i ntervals, a cross 1 0,000 t rials p lotted a gainst visit number to state 2 or 3 during Phases I and III. At the beginning of Phase III, the MB learning algorithm shows a noticeable improvement in performance compared to both its initial phase performance and that of the MF learning algorithm.*

To assess the robustness of our results, we expanded our simulation experiments to include a variety of parameter settings. We varied the learning rate (0.1, 0.3, 0.5), discount factor (0.85, 0.95, 0.99), and exploration rate (0.1, 0.2, 0.3) in a factorial design (see Appendix A.2).

5. Discussion. Distinguishing between model-free and model-based learning is an essential component to many psychology and reinforcement learning studies. We presented an adaptation of local learning within the two-stage Markov task to better distinguish between the two. Unlike the original two-stage Markov task, our task incorporates the LoCA framework which has been shown to identify model-based learning for varied parameters and task representations [26].

There are several limitations to consider. Primarily, a study with human participants is needed to demonstrate that the task is acceptable, reliable, and learnable by humans. We made the simplifying assumption that rewards following state 2 or 3 would be independent from the decision made in state 2 or 3. This sped up learning, but could lead to disengagement with human participants. We also assumed the MB algorithm has perfect knowledge of the transition probabilities. Needing to learn the transition probabilities could also slow down learning.

6. Acknowledgements. This material is based upon work supported by the Madison Experimental Mathematics Lab (MXM) at University of Wisconsin-Madison. Haley Kottler received research funding from the Institute for Foundations of Data Science at University of Wisconsin - Madison.

REFERENCES

[1] F. Cushman and A. Morris, *Habitual control of goal selection in humans*, Proceedings of the National Academy of Sciences, 112 (2015), pp. 13817–13822.

[2] C. F. da Silva and T. A. Hare, *A note on the analysis of two-stage task results: How changes in task structure affect what model-free and model-based strategies predict about the effects of reward and transition on the stay probability*, PLoS ONE, 13 (2018), p. e0195328, https://doi.org/10.1371/journal.pone.0195328, https://doi.org/10.1371/journal.pone.0195328.

[3] 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, 69 (2011), pp. 1204–1215, https://doi.org/10.1016/j.neuron.2011.02.027, https://linkinghub.elsevier.com/retrieve/pii/S0896627311001255 (accessed 2023-01-24).

[4] J. H. Decker, A. R. Otto, N. D. Daw, and C. A. Hartley, *From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning*, Psychological science, 27 (2016), pp. 848–858.

[5] L. Deserno, Q. J. Huys, R. Boehme, R. Buchert, H.-J. Heinze, A. A. Grace, R. J. Dolan, A. Heinz, and F. Schlagenhauf, *Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making*, Proceedings of the National Academy of Sciences, 112 (2015), pp. 1595–1600.

[6] B. B. Doll, K. D. Duncan, D. A. Simon, D. Shohamy, and N. D. Daw, *Model-based choices involve prospective neural activity*, Nature neuroscience, 18 (2015), pp. 767–772.

[7] B. B. Doll, D. Shohamy, and N. D. Daw, *Multiple memory systems as substrates for multiple decision systems*, Neurobiology of learning and memory, 117 (2015), pp. 4–13.

[8] E. Enkhtaivan, J. Nishimura, C. Ly, and A. L. Cochran, *A competition of critics in human decision-making*, Computational Psychiatry, 5 (2021), p. 81, https://doi.org/10.5334/cpsy.64.

[9] B. Eppinger, M. Walter, H. R. Heekeren, and S.-C. Li, *Of goals and habits: age-related and individual differences in goal-directed decision-making*, Frontiers in neuroscience, 7 (2013), p. 253.

[10] C. Feher da Silva and T. A. Hare, *A note on the analysis of two-stage task results: How changes in task structure affect what model-free and model-based strategies predict about the effects of reward and transition on the stay probability*, PloS one, 13 (2018), p. e0195328.

[11] C. Feher da Silva and T. A. Hare, *Humans primarily use model-based inference in the two-stage task*, Nature Human Behaviour, 4 (2020), pp. 1053–1066.

[12] C. Feher da Silva, G. Lombardi, M. Edelson, and T. A. Hare, *Rethinking model-based and model-free influences on mental effort and striatal prediction errors*, Nature Human Behaviour, (2023), pp. 1–14.

[13] C. M. Gillan, A. R. Otto, E. A. Phelps, and N. D. Daw, *Model-based learning protects against forming habits*, Cognitive, Affective, & Behavioral Neuroscience, 15 (2015), pp. 523–536.

[14] Q. W. Khan, *Comprehensive survey of optimization applications and techniques*, Machine Learning Robotics, (2024), https://doi.org/10.61927/igmin210, https://doi.org/10.61927/igmin210. Received 24 Jun 2024, Accepted 03 Jul 2024, Published online 04 Jul 2024.

[15] D. Kim, G. Y. Park, J. P. O'Doherty, and S. W. Lee, *Task complexity interacts with state-space uncertainty in the arbitration between model-based and model-free learning*, Nature Communications, 10 (2019), https://doi.org/10.1038/s41467-019-13653-5, https://doi.org/10.1038/s41467-019-13653-5.

[16] W. Kool, F. A. Cushman, and S. J. Gershman, *When does model-based control pay off?*, PLoS computational biology, 12 (2016), p. e1005090.

[17] W. Kool, S. J. Gershman, and F. A. Cushman, *Planning complexity registers as a cost in metacontrol*, Journal of cognitive neuroscience, 30 (2018), pp. 1391–1404.

[18] K. J. Miller, M. M. Botvinick, and C. D. Brody, *Dorsal hippocampus contributes to model-based planning*, Nature neuroscience, 20 (2017), pp. 1269–1276.

[19] A. R. Otto, S. J. Gershman, A. B. Markman, and N. D. Daw, *The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive*, Psychological science, 24 (2013), pp. 751–761.

[20] A. R. Otto, C. M. Raio, A. Chiang, E. A. Phelps, and N. D. Daw, *Working-memory capacity protects model-based learning from stress*, Proceedings of the National Academy of Sciences,

110 (2013), pp. 20941–20946, https://doi.org/10.1073/pnas.1312011110, https://www.pnas.org/doi/abs/10.1073/pnas.1312011110, https://arxiv.org/abs/https://www.pnas.org/doi/pdf/10.1073/pnas.1312011110.

[21] A. R. Otto, A. Skatova, S. Madlon-Kay, and N. D. Daw, *Cognitive control predicts use of model-based reinforcement learning*, Journal of cognitive neuroscience, 27 (2015), pp. 319–333.

[22] O. Rainio, J. Teuho, and R. Klén, *Evaluation metrics and statistical tests for machine learning*, Scientific Reports, 14 (2024), https://doi.org/10.1038/s41598-024-56543-y, https://doi.org/10.1038/s41598-024-56543-y.

[23] M. Sebold, L. Deserno, S. Nebe, D. J. Schad, M. Garbusow, C. Hägele, J. Keller, E. Jünger, N. Kathmann, M. Smolka, et al., *Model-based and model-free decisions in alcohol dependence*, Neuropsychobiology, 70 (2014), pp. 122–131.

[24] P. Smittenaar, T. H. FitzGerald, V. Romei, N. D. Wright, and R. J. Dolan, *Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans*, Neuron, 80 (2013), pp. 914–919.

[25] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, MIT press, 2018.

[26] H. Van Seijen, H. Nekoei, E. Racah, and S. Chandar, *The loca regret: A consistent metric to evaluate model-based behavior in reinforcement learning*, 33 (2020), pp. 6562–6572, https://proceedings.neurips.cc/paper_files/paper/2020/file/48db71587df6c7c442e5b76cc723169a-Paper.pdf.

[27] V. Voon, K. Derbyshire, C. Rück, M. A. Irvine, Y. Worbe, J. Enander, L. R. Schreiber, C. Gillan, N. A. Fineberg, B. J. Sahakian, et al., *Disorders of compulsivity: a common bias towards learning habits*, Molecular psychiatry, 20 (2015), pp. 345–352.

[28] A. Wagenmaker, K. Huang, L. Ke, B. Boots, K. Jamieson, and A. Gupta, *Overcoming the sim-to-real gap: Leveraging simulation to learn to explore for real-world rl*, 2024, https://arxiv.org/abs/2410.20254, https://arxiv.org/abs/2410.20254.

[29] Y. Wan, A. Rahimi-Kalahroudi, J. Rajendran, I. Momennejad, S. Chandar, and H. H. Van Seijen, *Towards evaluating adaptivity of model-based reinforcement learning methods*, in International Conference on Machine Learning, PMLR, 2022, pp. 22536–22561.

[30] K. Wunderlich, P. Smittenaar, and R. J. Dolan, *Dopamine enhances model-based over model-free choice behavior*, Neuron, 75 (2012), pp. 418–424.

7. Appendix.

7.1. A.1 Transition Probability Sensitivity. To evaluate how transition probabilities affect task performance, we simulated two alternative versions of the task using 60/40 and 80/20 splits for common/rare transitions. All other parameters were held constant. Results indicate that while model-based learners still outperform model-free learners after the Phase II reward change, the degree of separation varies. Under 60/40, the difference narrows due to increased ambiguity in transition type; under 80/20, model-based advantages become more pronounced. These results are visualized in Figure 3 and support the robustness of our task while highlighting trade-offs in parameter tuning.

7.2. A.2 Parameter Sensitivity. To assess the robustness of our results, we expanded our simulation experiments to include a variety of parameter settings. While dynamic exploration schedules (e.g., decaying epsilon or uncertainty-based exploration) are an important direction for future research, we chose a fixed set of exploration values to simplify comparison and focus on the structural impact of task modifications. We varied the learning rate (0.1, 0.3, 0.5), discount factor (0.85, 0.95, 0.99), and exploration rate (0.1, 0.2, 0.3) in a factorial design. For each parameter combination, we ran simulations with both model-free and model-based agents and analyzed their Phase III performance. Despite differences in learning speed and variability, model-based agents consistently adapted more quickly to the contingency change introduced in Phase II, confirming that the task structure robustly differentiates learning

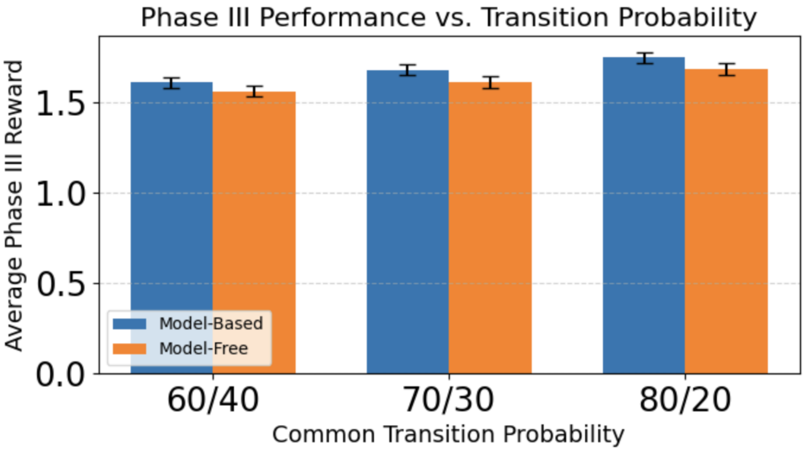

Figure 3. *Alternative Transition Prob Tests :Phase III performance of model-based and model-free agents under different common/rare transition probabilities (60/40 and 80/20). As transition probabilities become more deterministic (80/20), model-based advantages become more pronounced. With increased stochasticity (60/40), the performance gap narrows due to ambiguity in transition structure.*

strategies across a broad parameter space.

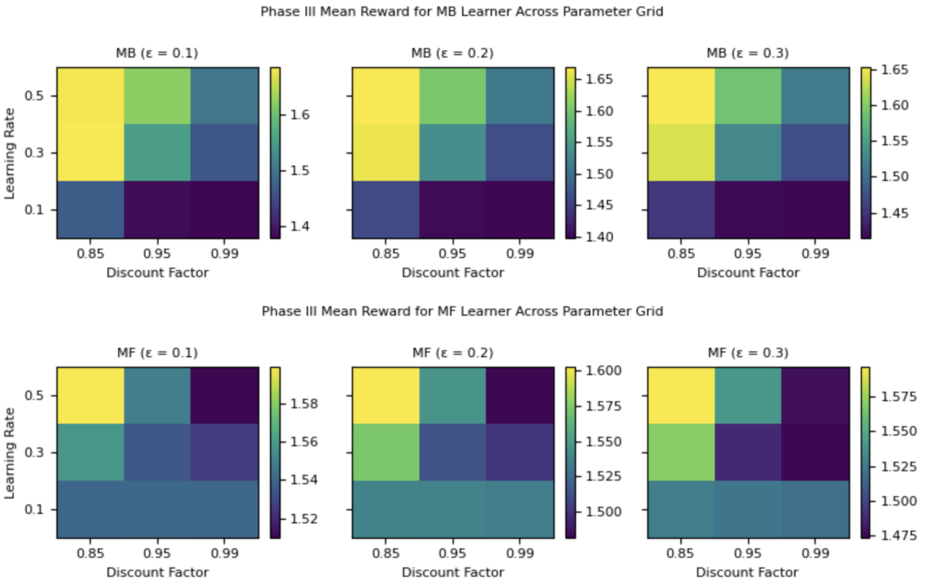

Figure 4. *Phase III accuracy of model-free and model-based agents across varying learning rates, discount factors, and exploration rates: Model-based agents consistently outperform model-free agents across all parameter combinations, indicating the robustness of our task in distinguishing learning strategies.*