Fictive Learning Augments Model-Based Reinforcement Learning in the Two-Step Task

Jianning Chen
Neural Computation Unit
Okinawa Institute of Science and Technology Graduate University
Okinawa, Japan
jianning.chen@oist.jp

Masakazu Taira
Department of Psychology
University of Sydney
Camperdown NSW Australia
masakazu.taira@sydney.edu.au

Kenji Doya
Neural Computation Unit
Okinawa Institute of Science and Technology Graduate University
Okinawa, Japan
doya@oist.jp

Abstract

Reinforcement learning (RL) is a normative computational framework to account for reward-based learning. Classical RL algorithms are based on experienced outcomes, whereas humans and animals may generalize learning to unexperienced events based on internal world models, so-called fictive learning. We propose a simple, brain-inspired fictive learning rule to augment model-based RL and use the rodent two-step task to examine whether fictive learning can better explain the observed behavior. This learning rule uses the same reward prediction error (RPE) to update both experienced and unexperienced states and actions, with scaling by the event correlation inferred from the internal model for fictive update. Through simulations, we show that this model better reproduces key behavioral traits observed in the two-step task. Model fitting validates its superior fit over existing alternatives. Furthermore, the model replicates the striatal dopaminergic dynamics observed in the same task, suggesting that the brain might operate through fictive learning for reward-based learning. The fictive learning observed here is conceptually analogous to, and partially inspired by, existing counterfactual approaches in machine learning. This convergence illustrates how machine learning offers candidate mechanisms for biology, how experiments reveal new theoretical principles, and how paradigms like the two-step task can serve as shared testbeds to evaluate both robotic and biological agents. Fictive learning exemplifies the broader opportunities for deeper collaboration between theorists and experimentalists in understanding intelligence.

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

1 Introduction

Learning from history to improve future decisions is the key to adaptation. Reinforcement learning (RL) [19] is the canonical theory to describe reward-based learning. An RL agent learns to predict the future outcome from experience and takes the difference between the prediction and the actual outcome, the reward prediction error (RPE), to update the prediction. There have been great successes in applying RL to explain the reward-based behavior in animals and humans. Especially, the distinction between the model-based and model-free RL has been intensively studied using the two-step task [2, 4, 8, 14].

We performed a two-step task experiment in mice [4], and found that the mice's behaviors were difficult to reproduce by standard model-free, model-based, or hybrid RL algorithms. We hypothesize that the capability of fictive learning, one of the topics where cognitive science, neuroscience, and machine learning converge, might underlie this mismatch. Humans and animals often learns by asking, "If I did something different, what outcome would I have gotten?" [6, 9, 15]. Specifically, they learn about non-encountered events by imagining their potential returns based on the information from experience, which requires the correlation between experienced and non-encountered events informed by an adequate world model [6, 15]. Payoffs between options are often anticorrelated in two-step tasks in rodents [2, 4], which encourages fictive learning. The commonly observed win-stay-lose-shift strategy might result from the fictive learning from the option's anticorrelation.

Fictive reward-related signals were found in the regions that are responsible for factual RPE computation, including striatum [12, 7], and orbital frontal cortex [1], where neurons encoding different actions and states overlap [17, 11]. Factual RPE might be generalized as fictive RPE by the inferred event correlation determined by the co-activation (or mutual inhibition) or overlapping of neurons encoding multiple actions and states in those regions.

Similar fictive learning principles have also been explored in machine learning for causal inference [20], efficient policy learning [5], and credit assignment [10]. This convergence suggests that fictive learning could be a compelling example of how biological and artificial intelligence may develop similar computational principles through evolution or design. Motivated by this view, we study whether fictive learning can resolve the mismatch between current theories and experimental observation in the two-step task, which would contribute to the understanding of flexible reward-based learning. More generally, evaluating the two-step task as a shared experimental paradigm offers an opportunity to compare artificial and biological agents.

We implement fictive learning in model-based RL by the generalized RPE and conduct simulation and animal experiments in a two-step task. Our model exploits the factual RPE computed in factual learning, scaled by a variable event correlation to ensure its flexibility. Previous studies often either explicitly instructed the options anticorrelation (buying and selling in the stock market [12]) or had no correlation [3]. In our experiment, animals were not instructed of anticorrelation, allowing us to examine whether fictive learning would naturally arise in reward-based learning.

In the following sections, we first describe the experiment design and model. We then simulated existing models without fictive learning to show that they fail to explain the experimental result. Next, we show that a fictive model-based RL fits the observation by simulation, followed by the explanation of action updating. Then, the model fitting confirmed that the fictive model-based model fit the behavior better than others. Finally, we conclude by discussing the model and highlighting how this study exemplifies the mutual inspiration between theory and experiment.

2 Methodology

2.1 Experiment design

We trained 10 mice (C47/BL background) with the two-step task (Fig. 1a) [2, 4]. The mice freely chose between the left and right options in ~ 75 % trials (mice were forced to choose left or right otherwise), which led to either an up or a down state with either common (80%) or rare (20%) probability. The transition probability matrix was fixed between subjects and counterbalanced across subjects. That is, the left option commonly leads to the up state in some animals, and vice versa in others. Reward is delivered probabilistically at each state, and the reward probabilities of up and down states are different in three block types (Table. 1). The reward settings changed block-wise.

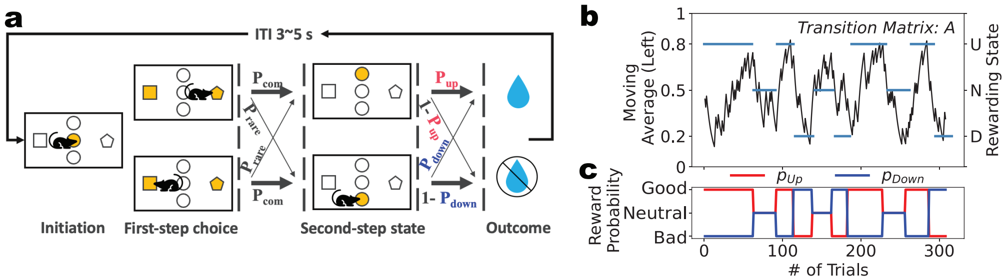

Figure 1: **a**, Task structure. After initiation, a first-step choice between left and right is presented, followed by up or down state with either common (80%) or rare (20%) probability. Two states are rewarded with different probabilities. **b**, Example behavior. The exponential moving average of animal choice (black line) traces the reward setting (blue bar). **c**, reward settings. Reward probabilities change in blocks anticorrelated.

Table 1: Reward Probability

Block	Up		Down		Neutral	
Outcome	Reward	Omission	Reward	Omission	Reward	Omission
Up state	0.8	0.2	0.2	0.8	0.5	0.5
Down state	0.2	0.8	0.8	0.2	0.5	0.5

In a non-neutral (up or down) block, block changes after 5 - 15 trials once the exponential moving average of the correct response (i.e., the option that commonly leads to the state with higher reward probability) in the last eight free-choice trials reached 0.75. The reward settings changed after 20 - 30 trials in the neutral block.

2.2 Model description

We included some canonical models from relevant literature as the baseline for model comparison, including three model-free, two model-based, and six mixture models. We then integrate our fictive learning with those baseline models to demonstrate how the fictive learning component could affect the simulation behavior and fit to the real data. We simulated 7 agents in Fig.2 and conducted a model fit with 19 models (7 fictive learning models) (A.2).

2.2.1 Baseline models

In the task, the agent chooses an action $a \in (left, right)$, which leads to the second-step state $s \in (up, down)$, where the outcome $r \in (0, 1)$ is delivered. Agents learn the action value $Q(a)$ differently, yet the action selection follows the softmax function.

$$P(a) = \frac{e^{\beta Q(a)}}{\sum_{i \in Left, Right} e^{\beta Q(i)}} \tag{1}$$

The model-free models have the same learning rule but different eligibility trace parameter, λ. The MF(lambda) agent updates its action value of chosen options $Q_{mf}(a)$ and state value of experienced state $V(s)$ by the RPEs as follow,

$$V(s) \leftarrow V(s) + \alpha\delta_s \tag{2}$$

$$\delta_s = r - V(s) \tag{3}$$

$$Q_{mf}(a) \leftarrow Q_{mf}(a) + \alpha(\delta_a + \lambda\delta_s) \tag{4}$$

$$\delta_a = V(s) - Q_{mf}(a) \tag{5}$$

The model is termed MF and MF(memory) when the eligibility trace λ is 1 or 0, respectively.

The model-based, MB, and the Bayesian hidden state model, hidden state, exploit the learned model, the transition matrix between actions and state ($P(s|a)$). The MB agent learns the state value by Equation (2) and then computes the action value by,

$$Q_{mb}(a) \leftarrow \sum_{s} P(s|a)V(s) \tag{6}$$

The hidden state agent (A.1) assumes that there are two hidden states in which either of the two second-step states is better and updates the beliefs of being one hidden state ($h \in h_{up}, h_{down}$) using Bayesian inference [4]. Specifically, the agent estimates the $P(h_{up})$ by Bayesian inference with likelihood $P(r|s, h_{up})$ being the reward probability in the experiment (Table.1).

Therefore, the state values are updated as,

$$V(s) = P(r|s, h_{up})P(h_{up}) + P(r|s, h_{down})P(h_{down}) \tag{7}$$

And the action is updated as in the MB model (6).

We also included the asymmetric hidden state model for comparison [4]. In this model, the agent treats the omission in up and down states as the same observation by using the likelihood table (A.1), while other components remain the same.

A hybrid model consists of both model-free and model-based models by,

$$Q_{hybrid} = \epsilon Q_{mf} + (1 - \epsilon)Q_{mb} \tag{8}$$

2.2.2 Fictive learning

The fictive learning is implemented by updating the state value of the unvisited state ($V(s_-)$) and the action value of the unchosen option ($Q(a_-)$), by the RPE from visited state ($V(s_+)$) and chosen option ($Q(a_+)$) in Equation (2) and (4). The proportion of updating depends on the inferred event correlation of reward probability η_s between states and η_a between options by,

$$Q(a_-) \leftarrow Q(a_-) + \alpha(\eta_a \delta_a + \lambda \eta_s \delta_s) \tag{9}$$

$$V(s_-) \leftarrow V(s_-) + \alpha \eta_s \delta_s \tag{10}$$

The event correlation factors η are zero when the agent believes the reward probability of two actions and states change independently (as in baseline models), negative if anticorrelated, and positive if changing in the same direction.

In model simulation and fitting, since the transition matrix was fixed and well-instructed, we assumed that the agent believes the correlations between states and actions are the same (i.e., $\eta_s = \eta_a = \eta$). Note that our model is different from [3, 16]. Specifically, we did not assume a separate learning rate for fictive updating. And, event correlation is a free meta-parameter learned and developed over sessions. Besides, the agent infers the fictive RPE instead of the fictive reward.

2.3 Analysis method

Analyses used custom Python, R, and Matlab scripts. For normally distributed data, we use the paired *t*-tests when within-subject comparison with equal sample size and unpaired *t*-tests otherwise. Otherwise, we used Wilcoxon signed-rank tests and Mann-Whitney U-tests, respectively.

We built the generalized linear mixed model (GLMM) using *fitglme* (Matlab 2023b) to predict the stay/switch behavior in free-choice trials with the logit link function. The random effects with subjects as grouping factors were included for all variables and the intercept. The model structure is,

stay/switch $\sim$ intercept + trials + choice + Δvalue + trans. + out $\times$ trans. + (variables | subject)

- stay/switch: 1 if the animal stayed at the same choice as the last trial and 0 otherwise.
- trials: number of trials experienced in the session.
- choice: previous action, 0.5 if the previous choice was left, -0.5 otherwise.
- ΔValue: inferred value difference. The estimated difference in reward probabilities between the chosen and unchosen options prior to the current trial (for calculation, see A.3).
- Trans.: 0.5 if the previous transition is common, -0.5 if rare.
- Out $\times$ Trans.: 0.5 if the previous trial was a common/reward transition or a rare/omission transition, -0.5 otherwise.

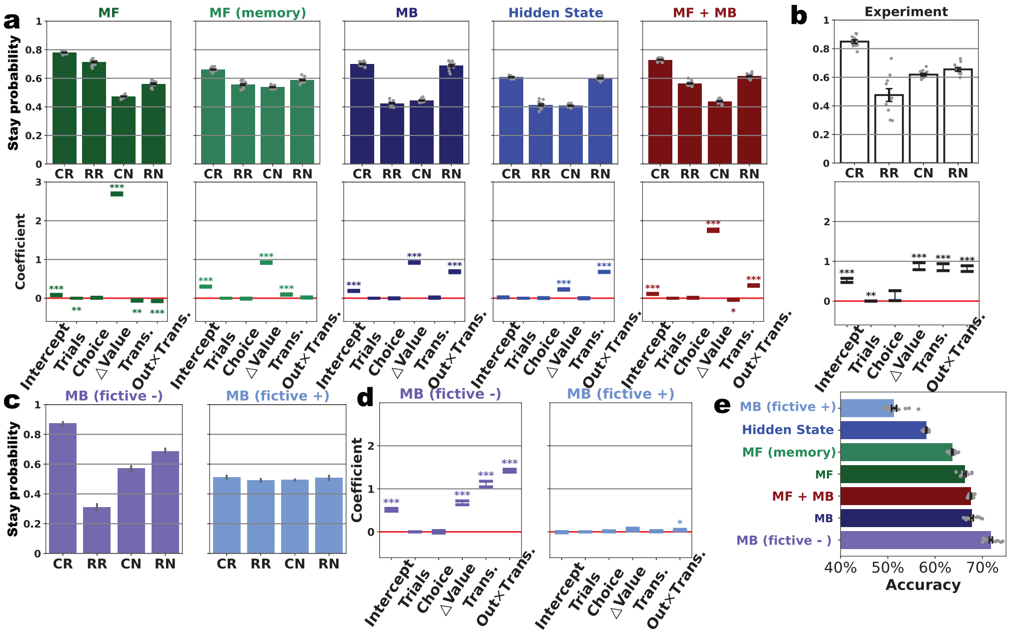

Figure 2: **a, b**, Behavior in simulation (a) and experiment (b). **Top**, the stay probability after trials with different outcome-transition pairs. Dots show each subject, and the error bar shows the between-subject mean $\pm$ s.e.m.. **Bottom**, GLMM result. The error bar shows the estimated coefficient $\pm$SE, and the star represents the significance. *, $0.01 \leq p \leq 0.05$; **, $0.001 \leq p \leq 0.01$, and *** ≤ 0.001.**c, d**, stay probability (c) and GLMM result (d) of fictive learning agent. **e**, The accuracy of included agents.

3 Result

3.1 Existing model fails to explain the experimental result

10 mice were tested to perform 17.600 ± 3.720 sessions, and were able to perform 425.500 ± 57.642 trials and completed 8.290 ± 1.645 of non-neutral blocks per session. Animals learned to optimize the choice (Fig. 1b) with 64.95% of correct choices.

In simulation, models from the MB and MF classes show distinct stay probabilities and GLMM coefficients (Fig. 2a). MB and hidden state model switch frequently after a rare/reward and common/no-reward trials, leading to the positive coefficient of interaction of outcome and transition in GLMM (Out $\times$ Trans.. MB: β = 0.684, SE = 0.026, t = 25.925, p < 0.001; hidden state: β = 0.678, SE = 0.016, t = 43.616, p < 0.001). Such a tendency reverses or disappears in the stay probability and GLMM for MF (β = -0.072, SE = 0.017, t = -4.183, p < 0.001) and MF (memory) (β = 0.018, SE = 0.015, t = 1.184, p = 0.236). The stay behavior depends heavily on the reward prediction of the chosen one over the unchosen ones based on the reward history (i.e., ΔValue) in model-free models (MF: β = 2.691, SE = 0.026, t = 102.09, p < 0.001; MF(memory): β = 0.920, SE = 0.020, t = 45.136, p < 0.001) than model-based models (MB: β = 0.922, SE = 0.020, t = 45.738, p < 0.001; hidden state: β = 0.226, SE = 0.018, t = 12.679, p < 0.001), as it essentially captures the direct reinforcement of the outcome on the action. Besides, all models show a similar tendency to repeat actions (i.e., intercept) and no or a marginal effect of transition type (Trans. MF(memory): β = 0.095, SE = 0.018, t = 5.370, p < 0.001).

Animal behavior is different from the above agents (Fig. 2b). Animals show the highest stay probability after the common/reward trial ($84.923 \pm 3.910\%$), resulting from the effect of value history (ΔValue: β = 0.623, SE = 0.086, t = 7.265, p < 0.001). However, animals were likely to switch following the rare/reward trial, and the stay probability is marginally lower after common/no-reward trials ($61.863 \pm 2.763\%$) than rare/no-reward trials ($65.513 \pm 3.973\%$)($stat$ = 5.00, p = 0.020, Wilcoxon test), suggesting the effect of the interaction of outcome and transition type (Out $\times$Trans.: β = 0.613, SE = 0.061, t = 10.080, p < 0.001) and the involvement of model-based learning. Besides, transition type strongly modulates the stay probability, mainly after a rewarded trial, whereas it only has a subtle effect after the unrewarded trials, leading to a significant positive coefficient of transition type in GLMM (Trans.: $\beta = 0.666$, SE = 0.062, $t = 10.696$, $p < 0.001$).

The existing models cannot replicate the observation. Animals' stay probability after common/reward trials is higher than all model predicts. Animals are more likely to switch after rare/reward trials than after common/no-reward trials. By contrast, the model-based models showed the equal stay probability in two cases, and the model-free and hybrid models showed the opposite pattern. Therefore, none of those models shows a strong positive coefficient of transition type.

3.2 Fictive learning agent fits the experiment result

Including fictive learning with anticorrelation (i.e., η = -1), MB(fictive -) generates behavior similar to animal behavior in the stay probability (Fig. 2c) and the GLMM coefficients (Fig. 2d). This model also achieves the highest accuracy (71.793 $\pm$1.252 %)(Fig. 2e). By contrast, MB(fictive +) (i.e., η = 1) shows seemingly random behavior, yet GLMM reveals a significant coefficient of outcome transition type interaction ($\beta = 0.041$, SE = 0.017, $t = 2.489$, $p = 0.013$).

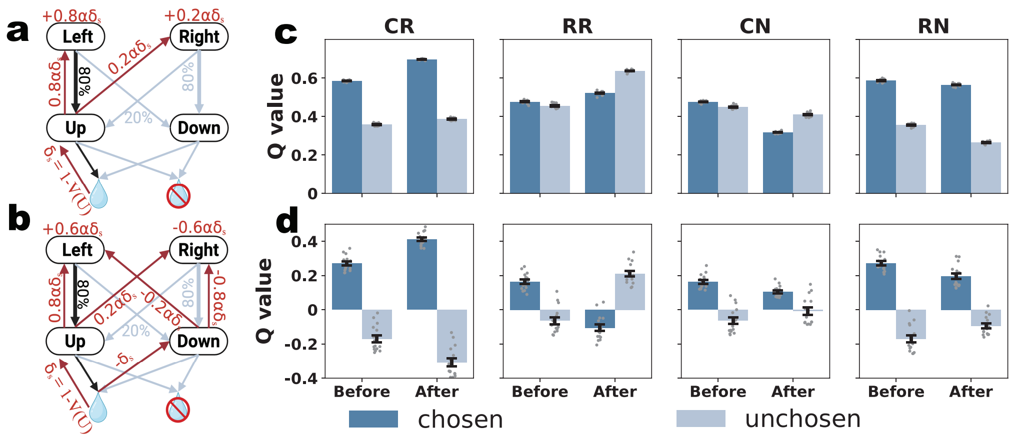

Figure 3: **a,b**, The action value update in common/reward trials. When the left option was chosen, followed by a common transition to the up state, and a reward. Action values of chosen and unchosen options get updated by $0.8\alpha\delta_s$ and $0.2\alpha\delta_s$ in MB (a) and $0.6\alpha\delta_s$ and -$0.6\alpha\delta_s$ in MB(fictive-) (b)(Created in BioRender. https://BioRender.com/e8qv5x0). **c,d**, Action updating in MB (a), and MB(fictive -) (b) in four transition/outcome pairs.

Table 2: Action updating table

	MB				MB (fictive –)			
	CR	RR	CN	RN	CR	RR	CN	RN
$\Delta Q(a_+)$	$0.8\alpha\delta_s$	$0.2\alpha\delta_s$	$0.8\alpha\delta_s$	$0.2\alpha\delta_s$	$0.6\alpha\delta_s$	-$0.6\alpha\delta_s$	$0.6\alpha\delta_s$	-$0.6\alpha\delta_s$
$\Delta Q(a_-)$	$0.2\alpha\delta_s$	$0.8\alpha\delta_s$	$0.2\alpha\delta_s$	$0.8\alpha\delta_s$	-$0.6\alpha\delta_s$	$0.6\alpha\delta_s$	-$0.6\alpha\delta_s$	$0.6\alpha\delta_s$

We examine why MB(fictive -) behaves differently from MB by analyzing how action value is updated (for complete derivation, see A.4). As an example, we domesticate the value updating after common/reward trials, assuming the agent chose the left (L) and visited the up state (U) (Fig. 3a,b). The action value of chosen ($Q(a_+)$) and unchosen options ($Q(a_-)$) are updated via the transition matrix in MB and MB(fictive -) by different magnitudes (Table. 2). The MB model updates the $Q(a_+)$ with δ_s scaled by 80% common probability and $Q(a_-)$ by 20% rare probability(Fig. 3c). In MB(fictive -), action value updates by two opposite δ_s, resulting in the simultaneous reinforcing of $Q(a_+)$ and fictive punishing $Q(a_-)$(Fig. 3d).

After a rare/reward trial, preference reversal happens with distinct rationales in the two models. The MB model (Fig. 3c) learns to increase the action value of both options, but with a larger magnitude for the $Q(a_-)$, leading to the takeover in action value and a switch choice. MB(fictive -) (Fig. 3d) decreases the $Q(a_+)$ but increases the $Q(a_-)$. Thus, the takeover in action value is more substantial in MB (fictive -) and leads to a lower stay probability than MB (Fig. 2a,c).

After common/no-reward trials, two models update the action value by the negative RPE differently. In the MB model, $Q(a_+)$ decreases dramatically, yet $Q(a_-)$ drops modestly, leading to the preference reversal (Fig. 3c). In MB(fictive -) model (Fig. 3d), $Q(a_+)$ decreases and $Q(a_-)$ increases. Since the agent and animals performed the task well, getting a reward omission after a common transition, an incorrect choice, is not due to insufficient learning, but likely happened because omission happened with 20% probability, or block change. In either case, $Q(a_+)$ should be substantially higher than $Q(a_-)$, and hence one-shot updating is not enough to trigger the preference reversal but only attenuates the difference between action values.

After rare/no-reward trials, in MB model (Fig. 3c), an omission causes a negative RPE, leading to a notable decrease in $Q(a_+)$ and a slight decrease in $Q(a_-)$ without preference reversal. However, in the MB(fictive -) model (Fig. 3d), a rare transition usually directs the agent to an infavored state with negative state value by fictive punishment (i.e., given the negative correlation between state, since the visited state gives fascinate outcome, being unvisited state might give a punishment), implying the RPE can be positive in some cases. Hence, the action value difference was equalized, but no preference reversal happened.

3.3 Fictive Model-based agent fits the observation better

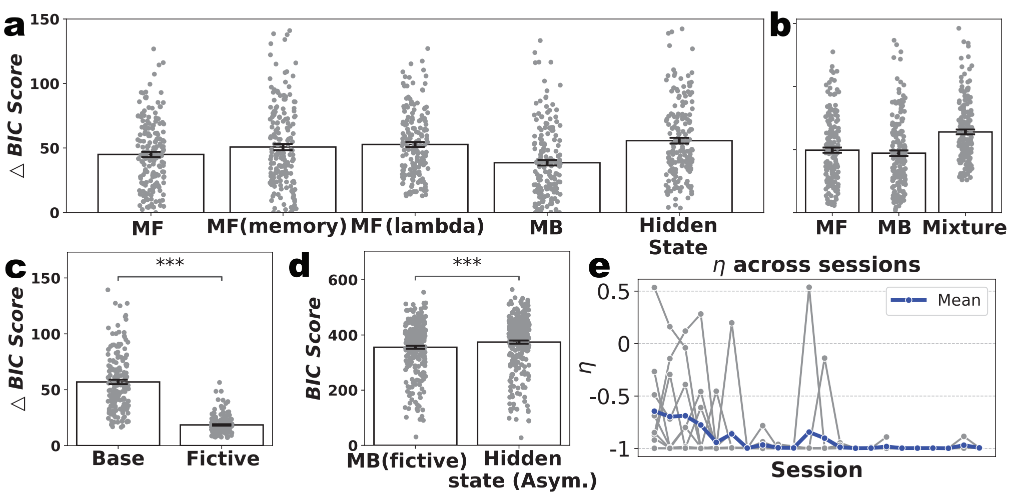

Figure 4: **a**, Model comparison of single baseline models. The ΔBIC score is the Bayesian Information Criterion (BIC) score normalized by the BIC score of the winning model per subject/session. The MB model achieves the lowest ΔBIC score. **b**, Model comparison of MF, MB, and mixture classes. MB classes fit the data better. **c**, Model comparison between all baseline models and their variants with extra fictive learning. Extra fictive learning improves model fit. **d**, Model comparison between MB (fictive) and asymmetric hidden state model. MB (fictive) fits the data better in general. **e**, estimated event correlation parameter η in each mouse (gray) and its group mean (blue). η tends to decrease to around -1 with considerable individual variance.

The Bayesian modeling and model comparison suggest observed behavior is likely to be model-based and captured well by fictive learning. Amoing baseline model, MB model provides better fits in general (Δ BIC: 38.587 $\pm$ 28.149) (Fig. 4a,b), suggesting that model-based learning is indeed dominant. Adding fictive learning causes a significant decrease in BIC score (Fig. 4c) (baseline model: 56.970 $\pm$25.994; fictive agent: 18.545 $\pm$8.201; $stat$ = 0, p < 0.001, Wilcoxon test). A hidden state model that learns differently from reward and omission could also predict a similar behavior pattern [4]. Yet, MB(fictive) shows a better fit than it suggested by absolute BIC score (Fig. 4d) (asymmetric hidden state model: 374.183 $\pm$97.681; MB (fictive): 355.407 $\pm$90.124; $stat$ = 9521.000, p < 0.001, Wilcoxon test), despite having the same number of hyperparameters. Consistent with simulation and task setting, estimated η is negative overall (-0.906 $\pm$0.248) (Fig. 4e) and shows a decreasing pattern over sessions, implying that animals learned the event anticorrelation by experience, with notable individual differences that might result from the learning speed or prior knowledge.

3.4 RPE explains the striatal dopaminergic activity

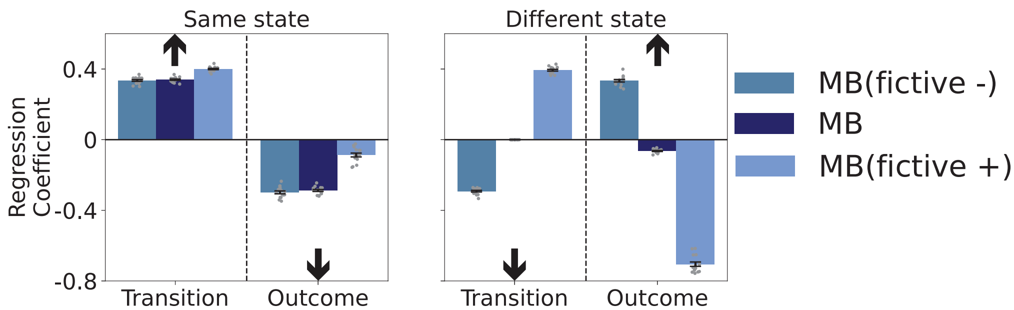

Figure 5: **a, b**, The coefficient of last outcome predicting the hypothetical dopamine signal derived from RPE in MB (fictive -), MB and MB (fictive +), when the same (a) or different (b) second-step state is presented. All three models predict the same pattern when the same state is presented. However, only MB (fictive -) predicts the same pattern as observed in dopamine release when the different state is presented. Arrows show the direction of the coefficient predicting the dopamine signal at the nucleus accumbens recorded in [4]. Transition RPE is the difference between the state value before and after the updating in the last trial of the current presented state, $V(s_+, t) - V(s_+, t-1)$, and the outcome RPE is the difference between the outcome and the state value, $r(t) - V(s_+, t)$.

The RPE from MB(fictive -) is consistent with the dopaminergic (DA) activity in the nucleus accumbens in mice performing the same task [4]. Strital DA dynamics are believed to signal the RPE modulated by the last outcome. Blanco-Pozo et al. [4] reported the reversal in the coefficient of the last outcome predicting the DA activity in the current trial. It was negative when the second-step state was revealed, but positive during the outcome period, if the presented state was the same as in the last trial. Yet, it shows a negative-to-positive reversal when experiencing the state that was not visited before, for which the classical MB model fails to reproduce.

To examine whether our model can reproduce this phenomenon, we derive the RPE as a proxy for the dopaminergic signal and predict it based on the last outcome. In Fig. 5a, when experiencing the same state, all models replicate the same reversal observed in real DA activity. However, when experiencing the different state (Fig. 5b), MB(fictive -) reproduces the observed pattern. Furthermore, MB(fictive +) predicts the positive-to-negative reversal in both cases. This result suggests that the DA pathways might operate computation similar to fictive learning.

4 Discussion

This study introduces a novel model that integrates fictive learning with the model-based RL model. This model outperforms others in the two-step task. It learn the task efficiently by exploiting the internal model. The simultaneous reinforcing and punishing mechanism facilitates learning, leading to superior accuracy. After understanding the task, the fictive learning agent could stay at the optimal choice more deterministically by enlarging the contrast between choices two-fold compared to other agents, and avoid exploration by explicitly updating the unvisited state and unchosen choice. It also fits the observed behavior in the two-step task better. The presented model provides a simpler, more integrated interpretation than the conventional view of model-based, model-free tradeoff [8, 2]. This study is also different from the previous literature in model design. Instead of deriving the fictive error as the outcome difference between the chosen and optimal action, which is often unknown [12, 7], or assumes the fixed and absolute anticorrelation [3, 16], our model exploits the factual RPE with scaling by an inferred correlation from learning.

Model comparison and possible biological mechanism Our model resembles the animal behavior and neural activities in the two-step task. The asymmetric hidden state model [4], one Bayesian inference model, shows similar performance, yet we argue that the presented model might be favored.

Both models reproduce the observed behavior. However, the significant difference in stay probability after common/no-reward and rare/no-reward trials is not consistent with the assumption that agents treat the reward omission in up and down states as the same observation in the asymmetric hidden state model. Hence, our model provides a lower BIC in model comparison.

More broadly, the fictive model-based model reproduces the RPE that fits the dynamics of striatal DA, which were argued to only be explained by the Bayesian inference model [4]. The previous unvisited state is not involved in factual learning. So when the previously unvisited state is presented, the dopamine activity can only be replicated by fictive learning or, Bayesian inference which implicitly implements the update of unvisited states and unchosen options by assuming that one state (and hence option) is better than the other. Similarly, after switching the choice in a rodent two-arm bandit task, NAc DA activity is more intense if more rewards were obtained by the previously chosen option, implying the action value of the current chosen option (i.e., the previously unchosen one) decreased before [16]. They [16] show only the Bayesian inference model, or the fictive learning-based model replicates this activity pattern. Together, that evidence highlights the parallel factual and fictive updating, which does not specifically favor the Bayesian inference model.

The Bayesian inference model only allows for a negative fictive update, whereas our model provides a more general rule, allowing for other variants. The reward likelihood is essential for Bayesian inference, which becomes intractable in more realistic settings. However, our model only requires the factual RPE to be broadcast and a rough estimate of event correlation, which is computationally and biologically plausible. The co-activation or mutual inhibition between neurons that represent experienced and non-encountered events develops from learning and allows for fictive learning. Furthermore, the observation that some neurons represent different options or states [17] might be related to fictive learning.

Hypotheses for future validation Our model generates hypotheses to investigate whether the brain engages in fictive learning. When correlation changes from negative to independent to positive, our model predicts the stay probability changes, from the pattern observed here, to an inverted-U shape, to seemingly random. The GLMM result would also change as the coefficient of transition type and value history diminishes. In neural activity, the fictive learning predicts that the coefficient of the last outcome on dopamine release is modulated by whether the same state is presented. This modulation disappears when event correlation is positive. Additionally, the co-activation between neurons or the proportion of neurons representing multiple options or states decreases when correlation is weakened.

Limitation The proposed model has its limitations. Animals appeared to learn this event correlation over sessions, with an unknown mechanism. A model-free RL or Bayesian updating rule might track this correlation online on a slow time scale. This approach imposes low cognitive and computational demands, but it is vulnerable to rapid environmental change. A computationally demanding Dyna-style architecture [18] can capture the correlation more robustly by mentally replaying past events. A more plausible compromise might be to maintain slow online tracking and trigger Dyna-like sampling selectively when predictions become unreliable (e.g., after multiple large RPEs). Fictive learning can also backfire if the estimate of the correlation is biased. Fictive learning's involvement might be modulated by the confidence of the estimate, which essentially depends on the learning of the world model. A hypothesis for further examination is that the transition from model-free to model-based systems would accompany or even drive the use of fictive learning.

The implication for machine learning and neuroscience Fictive learning is an example of a shared computational principle that links neuroscience, RL theory, and robotics. The parallel updating of factual and fictive information can improve efficiency, offering biological rules that may inform the design of adaptive artificial agents. The multi-step task can serve as a shared experimental platform for directly comparing the capacity of artificial and biological agents for fictive learning. In this direction, future work should extend the proposed learning rule into a non-tabular, generalizable form and benchmark it against related frameworks [5, 10, 13] in multi-step, multi-arm environments. Moreover, particular attention should be devoted to how fictive signals propagate through large action spaces and over extended temporal horizons. On the other side, a variety of developed counterfactual approaches, such as hindsight experience replay [10] and counterfactual regret minimization [21], provide algorithmically sophisticated candidate mechanisms for how the brain might implement fictive updates.

More broadly, advances in adaptive intelligence require the integration of theory, neuroscience, and robotics. Neuroscientists extract learning rules from the brain, while machine learning theorists formalize those rules into rigorous and efficient algorithms. Those algorithms are then evaluated in biological experiments for their plausibility, and also in robotic experiments for their strengths and limitations. The established behavioral experiment paradigm and analysis pipeline from neuroscience also contributes to the understanding of robotics. This iterative loop would advance our progress toward more general and adaptive forms of intelligence.

5 Conclusion

In conclusion, we integrate model-based RL with fictive learning and conduct in-silico and animal experiments to examine its ability to learn faster and explain animal behavior. We found that fictive learning facilitates learning while being algorithmically simple and biologically plausible. Model simulation and fitting show that it describes the behavior and dopamine dynamics in the two-step task better than the existing model. The presented result contributes to filling the gap between biological learning and the current RL theories, and highlights fictive learning as an example of how machine learning theories and experiments converge and inspire each other.

Acknowledgments and Disclosure of Funding

We acknowledge Thomas Akam and Mark Walton for their helpful advice on setting up the behavioral task. This study was funded by JSPS KAKENHI Grants 16H06563 and 23H04975, and Research Support of Okinawa Institute of Science and Technology Graduate University to K. D.

References

[1] Hiroshi Abe and Daeyeol Lee. Distributed coding of actual and hypothetical outcomes in the orbital and dorsolateral prefrontal cortex. *Neuron*, 70(4):731–741, 2011.

[2] Thomas Akam, Rui Costa, and Peter Dayan. Simple plans or sophisticated habits? state, transition and learning interactions in the two-step task. *PLoS Computational Biology*, 11(12): e1004648, 2015.

[3] Ido Ben-Artzi, Yoav Kessler, Bruno Nicenboim, and Nitzan Shahar. Computational mechanisms underlying latent value updating of unchosen actions. *Science Advances*, 9(42):eadi2704, 2023. doi: 10.1126/sciadv.adi2704. URL `https://www.science.org/doi/abs/10.1126/sciadv.adi2704`.

[4] Marta Blanco-Pozo, Thomas Akam, and Mark E. Walton. Dopamine-independent effect of rewards on choices through hidden-state inference. *Nature Neuroscience*, 27(2): 286–297, 2024. doi: 10.1038/s41593-023-01542-x. URL `https://doi.org/10.1038/s41593-023-01542-x`.

[5] Lars Buesing, Théophane Weber, Yori Zwols, Sébastien Racanière, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Manfred Otto Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. *ArXiv*, abs/1811.06272, 2018. URL `https://api.semanticscholar.org/CorpusID:53438249`.

[6] Ruth MJ Byrne. Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Sciences*, 6(10):426–431, 2002.

[7] Pearl H Chiu, Terry M Lohrenz, and P Read Montague. Smokers' brains compute, but ignore, a fictive error signal in a sequential investment task. *Nature Neuroscience*, 11(4):514–520, 2008.

[8] Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6): 1204–1215, 2011.

[9] Kai Epstude and Neal J Roese. The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, 12(2):168–192, 2008.

[10] Anna Harutyunyan, Will Dabney, Thomas Mesnard, Nicolas Heess, Mohammad G. Azar, Bilal Piot, Hado van Hasselt, Satinder Singh, Greg Wayne, Doina Precup, and Rémi Munos. Hindsight credit assignment. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

[11] Makoto Ito and Kenji Doya. Distinct neural representation in the dorsolateral, dorsomedial, and ventral parts of the striatum during fixed-and free-choice tasks. *Journal of Neuroscience*, 35(8): 3499–3514, 2015.

[12] Terry Lohrenz, Kevin McCabe, Colin F Camerer, and P Read Montague. Neural signature of fictive learning signals in a sequential investment task. *Proceedings of the National Academy of Sciences*, 104(22):9493–9498, 2007.

[13] Alexander Meulemans, Simon Schug, Seijin Kobayashi, Nathaniel D. Daw, and Gregory Wayne. Would i have gotten that reward? long-term credit assignment by counterfactual contribution analysis. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[14] Kevin J Miller, Matthew M Botvinick, and Carlos D Brody. Dorsal hippocampus contributes to model-based planning. *Nature Neuroscience*, 20(9):1269–1276, 2017.

[15] P Read Montague, Brooks King-Casas, and Jonathan D Cohen. Imaging valuation models in human choice. *Annual review of neuroscience*, 29:417–448, 2006. ISSN 0147-006X (Print); 0147-006X (Linking). doi: 10.1146/annurev.neuro.29.051605.112903.

[16] Albert J. Qü, Lung-Hao Tai, Christopher D. Hall, Emilie M. Tu, Maria K. Eckstein, Karyna Mishchanchuk, Wan Chen Lin, Juliana B. Chase, Andrew F. MacAskill, Anne G. E. Collins, Samuel J. Gershman, and Linda Wilbrecht. Nucleus accumbens dopamine release reflects bayesian inference during instrumental learning. *PLOS Computational Biology*, 21(7):1–31, 07 2025. doi: 10.1371/journal.pcbi.1013226. URL https://doi.org/10.1371/journal.pcbi.1013226.

[17] Kazuyuki Samejima, Yasumasa Ueda, Kenji Doya, and Minoru Kimura. Representation of action-specific reward values in the striatum. *Science*, 310(5752):1337–1340, 2005.

[18] Richard S. Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In Bruce Porter and Raymond Mooney, editors, *Machine Learning Proceedings 1990*, pages 216–224. Morgan Kaufmann. ISBN 978-1-55860-141-3. doi: 10.1016/B978-1-55860-141-3.50030-4. URL https://www.sciencedirect.com/science/article/pii/B9781558601413500304.

[19] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[20] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data*, 15(5), May 2021. ISSN 1556-4681. doi: 10.1145/3444944. URL https://doi.org/10.1145/3444944.

[21] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'07, page 1729–1736, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520.

A Appendix / supplemental material

A.1 Hidden state model

The hidden state agent assumes that there are two hidden states in which either of the two second-step states is better and updates the beliefs of being one hidden state ($h \in h_{up}, h_{down}$) using Bayesian inference [4].

Specifically, the agent estimates the $P(h_{up})$ by,

$$P(h_{up}) \leftarrow \frac{P(r|s, h_{up})P(h_{up})}{P(r)} \tag{1}$$

where the likelihood $P(r|s, h_{up})$ is the reward probability in experiment (Table. 1).

The marginal likelihood $P(r)$ is calculated as,

$$P(r) = P(h_{up})P(r|s, h_{up}) + P(h_{down})P(r|s, h_{down}) \tag{2}$$

The agent might assume that the block type would change with a certain probability τ. Hence, the posterior is updated as,

$$P(h_{up}) \leftarrow (1 - \tau)P(h_{up}) + \tau P(h_{down}) \tag{3}$$

Therefore, the state values are updated as,

$$V(s) = P(r|s, h_{up})P(h_{up}) + P(r|s, h_{down})P(h_{down}) \tag{4}$$

And the action is updated as in the MB model (6).

Table 1: Reward Probability

Block	Up block		Down block	
Outcome	Reward	Omission	Reward	Omission
Up state	0.4	0.5	0.1	0.5
Down state	0.1		0.4	

In asymmetric hidden state model, the agent exploits the reward likelihood in table.1.

A.2 Model simulation and fitting procedure

Table 2: Model and hyperparameter settings

Model	β	α	λ	τ	ϵ	η
MF	3	0.4	1			
MF (memory)	3	0.4	0			
MB	3	0.4				
Hidden state	3			0.2		
MF+MB	3	0.4	1		0.5	
MB (fictive -)	3	0.4	1			-1
MB (fictive +)	3	0.4				1

To examine which model will behave similarly to the experimental observation, we simulated 7 agents in table.2. To cover a wide range of hyperparameters, we added a noise term $noise \sim \mathcal{N}(0, 0.05 \times |hyperparameter|)$ for each agent. For each model, we simulated 15 agents for 20 sessions, and each session contained 400 free-choice trials, which is the typical length in animal experiments.

We implemented the Bayesian fitting with Rstan 2.32.6 with 4 MCMC chains for 2000 iterations (500 warm-up runs). We included 11 baseline models, an asymmetric hidden state model, and 7 fictive learning models. The hidden state model implicitly implements fictive learning by anticorrelation, so we did not add fictive learning to the 4 baseline models in which the hidden state model is involved to prevent confounding. The model fitting was performed for each subject and session to account for high subject/session-level variability and examine how η is learned over time. Events in force-choice trials are only used in updating the action value and state value, but do not contribute to the likelihood calculation. BIC score was used for model evaluation.

A.3 Inferred value difference calculation

$$\Delta Value_t = P_{a=a_{t-1},t} - P_{a\neq a_{t-1},t} \tag{5}$$

where,

$$P_{a,t} = \frac{\alpha_{a,t}}{\alpha_{a,t} + \beta_{a,t}} \tag{6}$$

where,

$$\alpha_{a,t+1} = decay * \alpha_{a,t} + r_t \tag{7}$$

$$\beta_{a,t+1} = decay * \beta_{a,t} + (1 - r_t) \tag{8}$$

The $decay$ is set as 0.5 to ensure the results' generalizability.

A.4 Proof of value updating rule

Model-based agents In the main text, we define an MB agent that updates the state value of visited $V(s_+)$ and unvisited state $V(s_-)$ by,

$$V(s_+) \leftarrow V(s_+) + \alpha\delta_s, \tag{9}$$

$$V(s_-) \leftarrow V(s_-), \tag{10}$$

and compute the action value via the transition matrix $P(s|a)$ by,

$$Q_{mb}(a) \leftarrow \sum_s P(s|a)V(s) \tag{11}$$

By definition 11, let $Q_{mb,new}(a)$ and $Q_{mb,old}(a)$ be action value before and after the state update in 9 and 10,

$$\begin{aligned} Q_{mb,new}(a) &= \sum_s P(s|a)V_{new}(s) = \sum_{s_-} P(s_-|a)V(s_-) + P(s_+|a)V_{new}(s_+) \\ &= \sum_{s_-} P(s_-|a)V(s_-) + \sum_{s_+} P(s_+|a)\left[V(s_+) + \alpha\delta_s\right] \\ &= \underbrace{\sum_s P(s|a)V(s)}_{=Q_{mb,old}(a)} + P(s_+|a)\alpha\delta_s. \end{aligned} \tag{12}$$

Therefore,

$$Q_{mb}(a) \leftarrow Q_{mb}(a) + P(s_+|a)\alpha\delta_s \tag{13}$$

This update weight $P(s_+|a)$ is the common/rare transition probability: actions more likely to lead to the reached state receive a larger portion of the prediction error.

Model-based agent with fictive learning MB(fictive-) agent updates both states by,

$$V(s_+) \leftarrow V(s_+) + \alpha\delta_s, \tag{14}$$

$$V(s_-) \leftarrow V(s_-) + \eta\alpha\delta_s, \tag{15}$$

Similarly, action value is updated by,

$$\begin{aligned} Q_{mb,new}(a) &= P(s_+|a)\left[V(s_+) + \alpha\delta_s\right] + \sum_{s_-} P(s_-|a)\left[V(s_-) + \eta\alpha\delta_s\right] \\ &= \underbrace{\left[P(s_+|a)\,V(s_+) + \sum_{s_-} P(s_-|a)\,V(s_-)\right]}_{=Q_{mb,old}(a)} + \alpha\delta_s\left[P(s_+|a) + \eta\sum_{s_-} P(s_-|a)\right]. \end{aligned} \tag{16}$$

Using $\sum_s P(s|a) = 1$, we have $\sum_{s_-} P(s_-|a) = 1 - P(s_+|a)$, so (16) becomes,

$$\begin{aligned} Q_{mb,new}(a) &= Q_{mb,old}(a) + \alpha\delta_s\Big[P(s_+|a) + \eta\big(1 - P(s_+|a)\big)\Big] \\ &= Q_{mb.old}(a) + \alpha\delta_s\Big[\eta + (1-\eta)P(s_+|a)\Big]. \end{aligned} \tag{17}$$

Therefore,

$$Q(a) \leftarrow Q(a) + \Big[\eta + (1-\eta)P(s_+|a)\Big]\alpha\delta_s \tag{18}$$