frontiers | Frontiers in Computational Neuroscience

TYPE Original Research
PUBLISHED 27 March 2025
DOI 10.3389/fncom.2025.1559915

OPEN ACCESS

EDITED BY
Hyunsu Lee,
Pusan National University, Republic of Korea

REVIEWED BY
Anil Yaman,
VU Amsterdam, Netherlands
Jieun Kim,
Ewha Womans University, Republic of Korea

*CORRESPONDENCE
Jee Hang Lee
✉ jeehang@smu.ac.kr;
✉ jeehang.iais@gmail.com

RECEIVED 13 January 2025
ACCEPTED 10 March 2025
PUBLISHED 27 March 2025

CITATION
Kim J and Lee JH (2025) Prefrontal meta-control incorporating mental simulation enhances the adaptivity of reinforcement learning agents in dynamic environments. *Front. Comput. Neurosci.* 19:1559915. doi: 10.3389/fncom.2025.1559915

Prefrontal meta-control incorporating mental simulation enhances the adaptivity of reinforcement learning agents in dynamic environments

JiHun Kim[1] and Jee Hang Lee[1,2,3]*

[1]Graduate School of AI and Informatics, Sangmyung University, Seoul, Republic of Korea, [2]Department of Human-Centered AI, Sangmyung University, Seoul, Republic of Korea, [3]Center for Neuroscience-Inspired AI, Institute for Advanced Intelligence Study, Daejeon, Republic of Korea

Introduction: Recent advances in computational neuroscience highlight the significance of prefrontal cortical meta-control mechanisms in facilitating flexible and adaptive human behavior. In addition, hippocampal function, particularly mental simulation capacity, proves essential in this adaptive process. Rooted from these neuroscientific insights, we present *Meta-Dyna*, a novel neuroscience-inspired reinforcement learning architecture that demonstrates rapid adaptation to environmental dynamics whilst managing variable goal states and state-transition uncertainties.

Methods: This architectural framework implements prefrontal meta-control mechanisms integrated with hippocampal replay function, which in turn optimized task performance with limited experiences. We evaluated this approach through comprehensive experimental simulations across three distinct paradigms: the two-stage Markov decision task, which frequently serves in human learning and decision-making research; *stochastic GridWorldLoCA*, an established benchmark suite for model-based reinforcement learning; and a *stochastic Atari Pong* variant incorporating multiple goals under uncertainty.

Results: Experimental results demonstrate *Meta-Dyna*'s superior performance compared with baseline reinforcement learning algorithms across multiple metrics: average reward, choice optimality, and a number of trials for success.

Discussions: These findings advance our understanding of computational reinforcement learning whilst contributing to the development of brain-inspired learning agents capable of flexible, goal-directed behavior within dynamic environments.

KEYWORDS

prefrontal meta-control, mental simulation, model-free learning strategy, model-based learning strategy, neuroscience of reinforcement learning, reinforcement learning agents

1 Introduction

The integration of reinforcement learning (RL) with deep learning architectures, called Deep RL, has accomplished unprecedented performance across numerous domains. Whilst the majority of RL implementations have relied upon model-free (MF) principles, there exists an expanding collection of model-based (MB) algorithms that aim to leverage enhanced sample efficiency and adaptive capacity. However, recent benchmark analyses by

Wan et al. (2022) challenge the assumption of MB supremacy over MF approaches. Recent findings in decision neuroscience suggest compelling evidence that the fundamental principle underlying human RL resides in meta-control mechanisms—specifically, the arbitration between MB and MF learning strategies, which dynamically adjusts based on environmental complexity and cost-benefit trade-offs (Kool et al., 2017; Kim et al., 2023), highlighting its crucial role in learning efficiency (Daw and Dayan, 2014).

Neuroscientific research in RL indicates a dual-process system consisting of MF and MB learning strategies: MF learning facilitates habitual behavior acquisition through reward prediction error (RPE), whilst MB learning enables goal-directed behavior through state prediction error (SPE), as established by Daw et al. (2005). Although state-of-the-art RL implementations predominantly adopt MF principles, MB approaches gain increasing attention due to their enhanced sample efficiency and adaptive capacity. Moreover, Dayan and Berridge (2014) demonstrated that MB learning augments MF approaches through cognitive prediction via environmental representation, thereby optimizing reward maximization with respect to computational efficiency and cognitive resource allocation (Dayan and Berridge, 2014).

However, despite scant research focus on MB approaches, their purported advantages in sample efficiency and adaptive capacity relative to MF implementations remain contentious (Wan et al., 2022). This observation precipitates critiques that MB RL does not consistently outperform MF RL, particularly in tasks that humans successfully achieve with relative ease. When environmental models fail to acquire complete state transition probabilities, performance becomes sub-optimal due to prolonged training requirements, as demonstrated by Bansal et al. (2017). Additionally, in experimental paradigms prevalent in neuroscience and cognitive psychology, such as the two-stage Markov decision task (MDT), purely MB strategies exhibit partial alignment with human behavior but fail to achieve complete correspondence, as established by Daw et al. (2011). These findings suggest inherent limitations in purely MB learning strategies, even within simple task environments that humans readily master.

In effect, research indicates that the fundamental principle underlying RL in the human brain centers on meta-control—specifically, the arbitration between MB and MF RL strategies (Lee et al., 2014; Daw et al., 2005; Abbott and Dayan, 2001). Neural correlates of State Prediction Error and Reward Prediction Error manifest in the dorsal prefrontal cortex (dlPFC) and ventral striatum, respectively. The inferior lateral prefrontal cortex (ilPFC) evaluates the relative reliability of competing learning strategies, whilst the ventromedial prefrontal cortex (vmPFC) functions as an arbitrator, implementing parallel control of MB and MF valuations (Lee et al., 2014; Dolan and Dayan, 2013). This neural architecture enables robust adaptation to environmental dynamics whilst optimizing the trade-off between performance, efficiency, and processing speed (Lee et al., 2019, 2022).

Complementary to meta-control mechanisms, mental simulation processes play a pivotal role in facilitating rapid and adaptive behavior during MB learning. Mental simulation constitutes a core mechanism through which the brain evaluates potential actions using internal environmental representations (Tolman, 1948; Daw et al., 2005). This cognitive capacity manifests even in rodent behavior, enabling novel route discovery and flexible planning under changing goal conditions (Dickinson, 1985; Tolman, 1948). Whilst step-wise mental simulation incurs substantial computational demands, it provides efficient adaptation to environmental dynamics through comprehensive state-space evaluation for optimal policy execution (Daw and Dayan, 2014).

The functional relationship between mental simulation and hippocampal replay elucidates neural mechanisms for managing these computational complexity. Hippocampal replay serves multiple functions: (i) facilitating sequential activation of hippocampal cells during rest periods (Karlsson and Frank, 2009), (ii) encoding topological structures of novel environments (Wu and Foster, 2014), and (iii) enabling goal-directed path simulation (Pfeiffer and Foster, 2013). These processes operate synergistically within the successor representation (SR) framework.

The SR framework extends predictive abilities through replay mechanisms, facilitating offline training via simulated experiences (Russek et al., 2017; Momennejad et al., 2017). This integration enables rapid action evaluation whilst maintaining behavioral flexibility through offline learning processes (Sutton, 1990; Mattar and Daw, 2018). Past research demonstrates that replay transcends mere experiential recapitulation, enabling novel trajectory construction (Gupta et al., 2010). Moreover, human studies indicate that replay events manifest abstract structural knowledge of acquired tasks (Liu et al., 2019). Notably, replay disruption impairs learning in contexts requiring history-dependent inference (Jadhav et al., 2012).

Based upon these neuroscientific insights, we present *Meta-Dyna*, RL architecture. Extending the *Dyna-Q* framework (Sutton and Barto, 2018), *Meta-Dyna* implements prefrontal meta-control mechanisms, providing an algorithmic model of arbitration between MF and MB learning strategies (Lee et al., 2014). We in addition enhance the MB component through integration of deep learning-based environmental modeling, enabling replay capacity via roll-out methodology. This architecture synthesizes meta-control mechanisms with mental simulation capabilities, thereby aiming to enhance RL agents' performance, environmental adaptation, and behavioral flexibility.

2 Preliminaries

2.1 Mental simulation and *Dyna* architecture

RL constitutes a framework where an agent acquires optimal action selection through environmental interaction to maximize future rewards. The environment, which is modeled as a Markov Decision Process (MDP), comprises a tuple $< \mathbf{S}, \mathbf{A}, R, T, \gamma, \mathbf{S'} >$. This tuple includes a set of states S, actions A, rewards R, state-action-state transition probability T, discount factor γ and transitions to the next states $\mathbf{S'}$. Within an MDP, the probability of transition to the subsequent state relies solely upon the

current state and action, irrespective of antecedent history (by Markov property). The agent endeavors to ascertain a policy π that stipulates actions for each state to maximize cumulative rewards.

Dyna-Q (Sutton, 1990), which amalgamates direct experiential learning with simulated experience planning, employs a planning component that enables the agent to augment its knowledge via a learnt environmental model rather than exclusively through actual experiences. The *Dyna-Q* planning process involves training an environmental model that predicts subsequent states and rewards, based on current states and actions. These simulated data facilitate Q-learning implementation, expediting optimal policy convergence. The acceleration of convergence correlates positively with increased planning steps, thus demonstrating the efficacy of incorporating simulation-based planning mechanisms into RL (Sutton, 1990; Sutton and Barto, 2018).

The brain exhibits analogous mechanisms to *Dyna*, utilizing both direct experience and simulated trajectories in choice evaluation (Momennejad et al., 2017). Through this simulation capacity, humans and animals successfully identify and circumvent suboptimal outcome pathways (Allen et al., 2020; Miller et al., 2017).

In this sense, mental simulation in the brain can be computationally implemented through various approaches, including the *Dyna* model architecture. Sutton (1990) introduced *Dyna* as an integrated architecture for learning, planning, and reacting, which simulates experience offline to update predictions. This approach mirrors how the brain might use mental simulation and replay to enhance learning. Nonetheless, whilst *Dyna* reflects how the brain improves learning via replay, recent studies have revealed its limitations, including decreased learning efficiency in real-world environments (Barkley and Fridovich-Keil, 2024). To that end, researchers have developed several improvements. These include data-driven inventory management using *Dyna-Q* (Qu et al., 2025) and out-of-distribution (OOD) data filtering, which enhances model reliability (Li et al., 2024). These optimisation efforts have expanded into various fields, such as industrial automation (Dong et al., 2020; Liu and Wang, 2021; Budiyanto and Matsunaga, 2023; Samaylal, 2024) and energy management (Saeed et al., 2024; Ghode and Digalwar, 2024; Liu et al., 2024, 2025).

The relationship between mental simulation and *Dyna* manifests particularly within the successor representation (SR) framework. The SR, which accommodates various hybrid implementations, includes "SR-*Dyna*" that employs either simulation or replay for offline SR updating (Russek et al., 2017; Momennejad et al., 2017). This offline updating process within *Dyna* mirrors the brain's adoption of hippocampal replay for goal-directed path simulation and construction.

A fundamental characteristic that bridges mental simulation and *Dyna* lies in their offline planning functionality. The brain's utilization of mental simulation during rest periods for enhanced learning and decision-making parallels *Dyna*'s deployment of simulated experience for offline value estimation updates. This correspondence suggests that *Dyna* captures essential aspects of the brain's flexible planning implementation through mental simulation (Mattar and Daw, 2018).

2.2 *Dyna* architecture and prefrontal meta-control in the human brain

Nevertheless, a fundamental question persists with regards to the harmonization of the dual systems within *Dyna*—namely, MB and MF components—toward bringing about optimal policy. Recent advances in decision neuroscience, which illuminate the arbitration control mechanisms that govern multiple learning strategies, inspire us. These advances proffer a resolution to this conundrum. These mechanisms, which specifically contain MB and MF RL paradigms, demonstrate remarkable efficacy in strategy reconciliation (Lee et al., 2014; Daw et al., 2005; Abbott and Dayan, 2001).

The arbitration control mechanism proves integral to decision-making optimisation, particularly in contexts where the relative appropriateness of MB vs. MF strategies exhibits variability. Lee et al. (2014), which presents neural evidence for an arbitration mechanism, demonstrates that the degree of control exerted by these dual strategies depends upon their respective prediction error (PE) reliability.

The Reward Prediction Error (RPE), which serves to compute the reliability of the MF strategy (Rel_{MF}), is calculated through Temporal Difference error (TD-Error). The mathematical formulation for this error is expressed as $RPE = r_t + \gamma Q_{MF}(s', a'; \theta_{MF}) - Q_{MF}(s, a; \theta_{MF})$. Conversely, the State Prediction Error (SPE), which determines the reliability of the MB strategy (Rel_{MB}), is formulated as $SPE = 1 - T(s, a, s')$, where T denotes the state-action-state transition probabilities defined in Equation 1.

$$T(s,a,s') = \begin{cases} P_s + \gamma(1 - P_s) & \text{if } P_s == P_{s'} \\ P_s \times (1 - \gamma) & \text{otherwise.} \end{cases} \tag{1}$$

The probability of MB control (P_{MB}) is derived from Rel_{MB} and Rel_{MF}, which are computed through Bayesian and non-Bayesian approaches, respectively (Li et al., 2011; Le Pelley, 2004; Sutton, 1992; Krugel et al., 2009; Pearce and Hall, 1980). The pseudo-level formulations are presented in Equation 2.

$$\begin{aligned} Rel_{MF} &= Pearce\ Hall(RPE|PE), \\ Rel_{MB} &= Bayesian(SPE|PE). \end{aligned} \tag{2}$$

For the MB reliability (Rel_{MB}), the Dirichlet distribution parameters—mean $E(Dir_i)$ and variance $V(Dir_i)$—are employed, where the subscript i indicates three distinct State Prediction Error (SPE) categories: negative, positive, and zero prediction errors. These categorical boundaries are established through a tolerance threshold ω, such that $PE < \omega$ constitutes negative error, $PE > \omega$ represents positive error, and values within this interval denote zero error.

The probability of utilizing MB learning strategies (P_{MB}) is determined via the arbitration mechanism, which adopts reliability (Equation 3).

$$\begin{aligned} \alpha &= \frac{A_\alpha}{(1+exp(B_\alpha * Rel_{MF}))}, \\ \beta &= \frac{A_\beta}{(1+exp(B_\beta * Rel_{MB})}, \\ P_{MB} &= P_{MB} + \alpha * (1 - P_{MB}) - \beta * P_{MB}. \end{aligned} \tag{3}$$

Through this computational framework, MB and MF values facilitate meta-control implementation. Finally they are concurrently applied in action selection through the weighting factor P_{MB}.

The theoretical framework described above exhibits strong correspondence with the neural substrates that underlie these computational processes. The bilateral inferior lateral prefrontal cortex regions, which encode MB and MF signal reliability, function in concert with the anterior cingulate cortex, which integrates reliability differentials to mediate arbitration (Lee et al., 2014). This reliability-driven arbitration mechanism determines strategic dominance, thereby facilitating dynamic environmental adaptation. Furthermore, the arbitration process receives support from additional neural structures: the ventral and dorsal striatum, which encode Reward Prediction Error (RPE), and the temporo-parietal cortex, which encodes State Prediction Error (SPE).

This neurobiological evidence substantiates that the fundamental question regarding dual-process amalgamation can be resolved, which has precipitated the development of neuroscience-inspired algorithmic frameworks that extend beyond the classical *Dyna* architecture.

3 *Meta-Dyna*: a neuroscience-inspired RL architecture

Within the context of recent RL, we propose *Meta-Dyna*, which constitutes a neuroscience-inspired algorithmic framework that extends the foundational *Dyna-Q* architecture (Figure 1). This novel approach implements dual learning processes – MB planning and MF Q-learning—whilst incorporating an arbitration mechanism for implementing the prefrontal meta-control. Moreover, owing to the inherent properties of the *Dyna* architecture, it facilitates mental simulation through the MB learning system. The framework, which embodies current neuroscientific understanding of cognitive processes, demonstrates how the brain synthesizes diverse learning strategies through mental simulation and experiential replay.

3.1 Overview

As described, *Dyna-Q* integrates Q-learning with planning mechanisms, which update Q-values through MB processes. The architecture employs planning through an environmental model that agents acquire via direct experiential interaction. *Meta-Dyna*, which extends this foundational architecture, implements dual Q-value systems: MB planning (Q_{MB}) and MF Q-learning (Q_{MF}).

Meta-Dyna distributes received state information to both real and model experience buffers. During the Q_{MF} update process, the world model incorporates training using a superset, which comprises both simulated and real experiences. The trained model subsequently generates simulated experiences through a recursive process. These experiences facilitate mental simulation, which ultimately updates Q_{MB}.

The prefrontal meta-control arbitrator computes the MB probability (P_{MB}) using reliability, which derive from Prediction Errors of each learning system. This meta-control component computes an integrated Q-value through weighted summation: P_{MB} for Q_{MB} and $1 - P_{MB}$ for Q_{MF} (Figure 1A). The complete processes are detailed in Algorithm 1.

3.2 Dual Q-value system for MB and MF

Meta-Dyna incorporates a dual Q-value system that implements MB and MF learning strategies. The *Meta-Dyna* architecture comprises one main inference network and three component networks: two Q-networks for MB and MF learning strategies, and a world model for simulating environmental dynamics (Figure 1A).

The main inference network combines outputs from the component networks, which guide decision-making processes. This main network operates under the governance of the prefrontal meta-control framework. The two Q-networks independently learn distinct behavioral patterns: habitual behaviors (MF) and goal-directed behaviors (MB). The MF Q-network learns habitual behaviors through reward signals from real experience, which utilizes standard Q-learning methods. Meanwhile, the MB Q-network learns from the world model to perform goal-directed planning. The world model learns to predict the environment's structure, which includes future states and rewards, thereby enabling the generation of simulated data for planning.

Through the separation of Q-values into MF (Q_{MF}) and MB (Q_{MB}) components, *Meta-Dyna* enables independent acquisition of habitual and goal-directed strategies. This architectural distinction implements a framework, which enables the agent to adaptively favor between MB and MF strategies based on their respective reliability, particularly in response to environmental dynamics (Figure 1B).

We note that two distinct implementations of *Meta-Dyna* were presented: a tabular version and a neural network (NN) variant. This architectural flexibility derives from the framework's generalisable principles, which enable deployment across diverse environmental contexts. We here employ MDN-RNN as the NN variant, specifically for its ability to handle uncertainty and model probabilistic environmental dynamics, aligning with *Meta-Dyna's* goal of generalization (Ha and Schmidhuber, 2018). MDN-RNN combines Mixture Density Networks with recurrent neural networks, enabling it to output probability distributions rather than deterministic predictions. This approach is particularly valuable in environments that are stochastic in nature, as it allows the model to capture inherent randomness and discrete random events.

As demonstrated in Ha and Schmidhuber (2018), using MDN-RNN helps prevent the agent from exploiting imperfections in the world model by introducing controlled uncertainty through temperature parameters. This makes it more difficult for agents to find adversarial policies that might work in the model but fail in actual environments. In addition, due to its generative model characteristics, it serves as an excellent candidate for mental simulation, which can foresee future states through the roll-out functionality. It showed how their world model (using MDN-RNN) can generate hypothetical scenarios and environments that the agent can interact with. They specifically created virtual environments generated by the MDN-RNN where agents

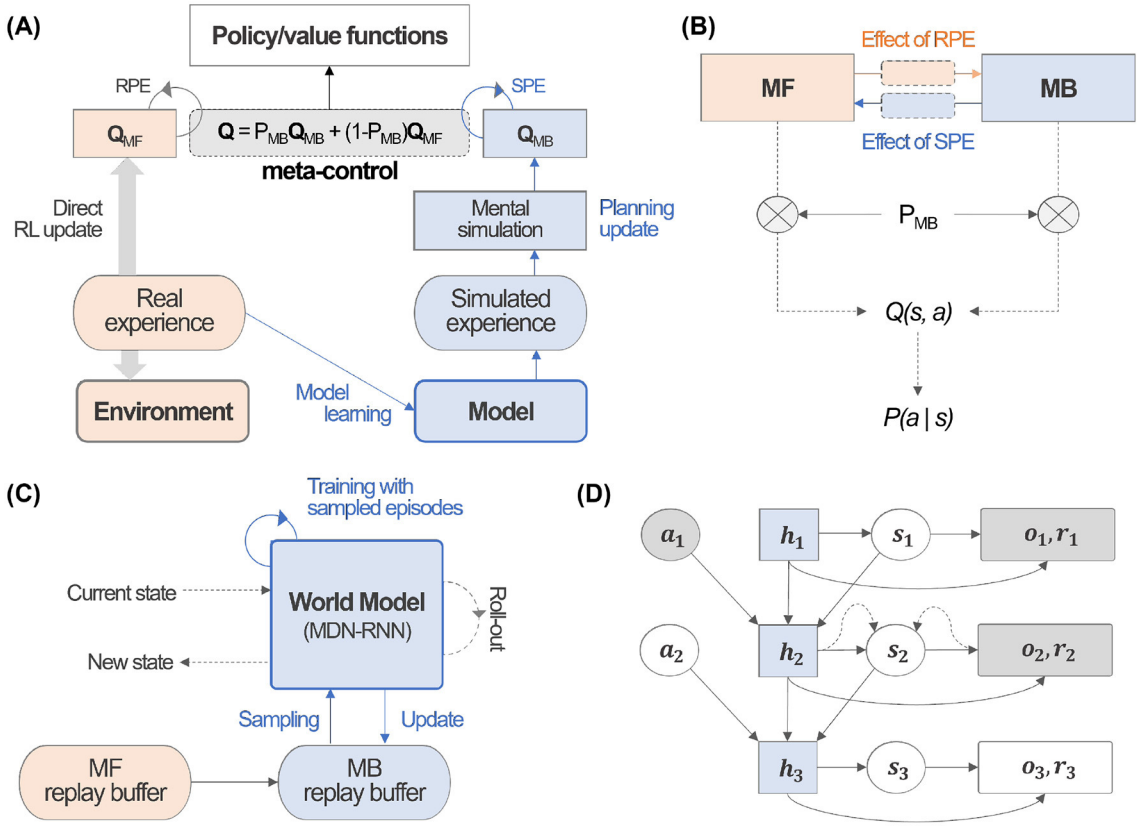

FIGURE 1
Meta-Dyna overview. **(A)** Schematic architecture of *Meta-Dyna*. **(B)** An illustration of prefrontal meta-control, which arbitrates between MB and MF learning strategies. **(C)** A schematic representation of the world model that enables mental simulation. **(D)** A demonstration of the roll-out process, which implements mental simulation for given state-action pairs.

could train entirely inside these simulated environments before transferring policies to actual environments. The concept of using the model to "*foresee future states*" through roll-outs is central to their approach, as they demonstrate in both the CarRacing and VizDoom experiments where the agent learns policies by simulating potential futures within the generated environment. The complete architectural framework is presented in Figure 1A.

3.3 Prefrontal meta-control in *Meta-Dyna*

3.3.1 Prediction errors and reliability

A key ingredient of prefrontal meta-control lies in the computation of reliability for both MB and MF strategies, which derive from their respective Prediction Errors: SPE and RPE. As previously established, RPE was computed as the temporal difference (TD) error of the Q_{MF} network, aligned with its canonical formulation (Lee et al., 2014). The computation of SPE, however, exhibited distinct methodologies contingent upon the state space characteristics—discrete or continuous.

In discrete state spaces, SPE computation employed the same computational model for learning state transition probabilities as that established in Lee et al. (2014). Within continuous state spaces, which was known to be intractable for traditional state transition probability methods, we implemented the Mixture Density Network-Recurrent Neural Network (MDN-RNN). We adopted its Gaussian Mixture Model (GMM) outputs as components for computing the SPE. The GMM generated multiple Gaussian distributions, which were characterized by parameters μ_k and σ_k, representing the distribution of potential subsequent states.

The standard deviation σ_k quantifies the predictive uncertainty of subsequent states. An increased magnitude of σ_k indicates elevated uncertainty, which indicates diminished predictive accuracy. For continuous state spaces, we formulate SPE using the distributional parameters μ_k and σ_k as follows:

$$SPE = \frac{\mu}{\sigma}, \text{ where } \mu, \sigma \text{ from GMM}. \quad (4)$$

These Prediction Errors facilitated the computation of reliability measures for MB and MF strategies. The arbitration mechanism utilizes these reliability values, which modulate the strategic balance dynamically, in accordance with the prefrontal meta-control framework.

3.3.2 Meta-control leading to decision making

As described, *Meta-Dyna* enables independent acquisition of Q-values through MB and MF Q-networks. This separation facilitates distinct learning trajectories, which characterize

habitual and goal-directed behaviors within each Q-network. The integration of MB and MF Q-networks proceeds through a weighted mechanism, which derives from their computed reliability indices.

The reliability determines the probability (P_{MB}) of selecting the MB Q-network. The Q-values from Q_{MB} and Q_{MF} are integrated through weighted summation, which employs P_{MB} and $(1 - P_{MB})$ as their respective coefficients. In the tabular implementation, this integration process comprises multiplicative operations between the Q-table values of MB and MF Q-networks with their corresponding probabilities, P_{MB} and $(1 - P_{MB})$, to update the main Q-table for inference:

$$Q(s, a) = P_{MB} \times Q_{MB}(s, a) + (1 - P_{MB}) \times Q_{MF}(s, a). \quad (5)$$

In the neural network implementation, this integration process applies to the Q-network parameters. The main network, which incorporates the integrated Q-values, guides action selection through an ϵ-greedy policy.

Through dynamic modulation of MB and MF Q-network integration based on their respective reliability, *Meta-Dyna* implements adaptive behavior in response to environmental dynamics, which harnesses the complementary strengths of both Q-networks.

3.4 Mental simulation in *Meta-Dyna*

3.4.1 World model learning environmental dynamics

The world model which learns environmental structure employs a Mixture Density Network-Recurrent Neural Network (MDN-RNN) architecture for sequential modeling (Ha and Schmidhuber, 2018). The MDN-RNN architecture learns sequential state-action pairs, which predict probability distributions of subsequent states and rewards. The integrated approach captures temporal dependencies through recurrent neural networks, which represent future states through probabilistic distributions via mixture density networks. Specifically, the MDN-RNN generates distributions of potential subsequent states, which are characterized by distributional parameters—means (μ) and standard deviations (σ)—that quantify expected outcomes and their associated uncertainties.

In RL, the world model serves as a simulation framework, which enables agents to evaluate hypothetical scenarios without direct environmental interaction. This methodology demonstrates particular efficacy in dynamic environments, where the MDN-RNN architecture captures latent uncertainties within the environmental dynamics. The model's predictive capabilities enhance planning processes, which facilitate efficient policy optimisation through simulated state transitions.

The world model's predictive capacity, which are integrated within the *Meta-Dyna* arbitration framework, use MDN-RNN outputs for SPE computation. The SPE, which quantifies predictive uncertainty, is mathematically defined as

$$SPE = \frac{\mu}{\sigma}, \quad (6)$$

where μ and σ derive from the Gaussian distributions that characterize subsequent states.

3.4.2 Architecture

In short, the MDN-RNN learns sequences of state-action pairs and predicts the probability distribution of the next state and reward. This world model, which incorporates dynamic uncertainties into decision-making processes, enhances the agent's reasoning capacity. The architecture comprises (i) recurrent neural network layers that process input sequences to capture temporal dependencies and (ii) mixture density network layers that predict distributional parameters (μ and σ) for subsequent states and rewards. The model receives PEs as inputs, which implements an error-based learning mechanism that emulates neural computation. These errors, which *Meta-Dyna* computes during the learning process, emerge naturally without external augmentation.

3.4.3 Training

The training protocol involves minimizing the PEs between predicted distributions and observed subsequent states and rewards. The training data comprise state-action sequences that derive from environmental interactions. An experience replay buffer stores model-generated experiences, which enhances learning efficiency and stability.

Once trained, the world model generates simulated scenarios for training the MB policy (Q_{MB}), which is analogous to Q-learning with simulated experiences. This process enables policy-model interaction through simulated state-action pair outcomes, thus facilitating efficient strategy implementation (Figures 1C, D).

3.5 Algorithm

The implementation of *Meta-Dyna* is contingent upon the task complexity of the environment. For tasks with manageable state-action spaces, a tabular implementation is sufficient, where Q_{MB} and Q_{MF} are represented as tables that contain values for all possible state-action pairs. The world model maintains SPE and RPE for each state-action pair.

High-dimensional tasks, which include image-based scenarios, necessitate a neural network implementation on the other hand. In this context, Q_{MB} and Q_{MF} are implemented as feed-forward networks. In high-dimensional environments such as Atari Pong, a convolutional neural network (CNN) is used as an encoder, specifically to extract spatiotemporal features from raw pixel inputs, enabling efficient mental simulation and policy adaptation under uncertainty. These extracted features serve as input for both the MB and MF components, enhancing learning efficiency in complex visual environments. The processing of high-dimensional inputs employs convolutional neural networks (CNNs) for preprocessing, which subsequently interface with feed-forward networks for Q-value representation. The world model in the neural network variant used the MDN-RNN architecture, which processes state, action, SPE, and RPE inputs to generate subsequent states and rewards.

```
1: Initialize Q_MF, Q_MB, World Model M, and replay
   buffer RB_MF, RB_MB
2: for each episode do
3:    for each time step do
4:       Select action a_t based on integrated Q values

5:       Execute a_t and observe s_{t+1} and r_{t+1}
6:       Store (s_t, a_t, r_{t+1}, s_{t+1}) in RB_MF
7:    end for
8:    Train Q_MF using experiences from RB_MF
9:    Train M using experiences from RB_MB
10:   Compute RPE and SPE
11:   for i = 1 to n do
12:      Generate simulated experience using M
13:      Train Q_MB using simulated experience
14:   end for
15:   Update reliability and calculate P_MB
16:   Integrate Q_MF and Q_MB using P_MB
17: end for
```

Algorithm 1. ***Meta-Dyna* learning process.**

The learning protocol included several stages shown in Algorithm 1. Initially, all components—Q_{MB}, Q_{MF}, and the world model—were initialized (line 1). At each timestep, the agent acquired sequences through actions derived from integrated Q-values (lines 4–7). These Markov Decision Process sequences facilitated Q_{MF} and environmental model training, whilst computing Prediction Errors (lines 8–10).

The training of Q_{MB} used simulated experiences generated by the environmental model. Planning proceeds for n predefined steps, where the model generated state-action pair sequences (lines 11–14). This parameter n balanced planning depth with model accuracy. Empirical testing established $n = 10$ as optimal, given that larger values result in diminishing returns when model accuracy converges.

The computed PEs update reliability, which subsequently modulates the integration of MB and MF learning strategies (lines 15–16). This iterative process enables *Meta-Dyna*'s adaptive learning within dynamic environments.

4 Experiments

To evaluate the behavioral flexibility and adaptation capacity of *Meta-Dyna*, we conducted experimental evaluations across three distinct paradigms. The first assessment employed the *Two-Stage Markov Decision Task (MDT)*, which is widely used in human decision-making research, specifically for choice behavior that is governed by MF and MB learning strategies. We subsequently developed a variant of the *GridWorldLoCA* environment to examine the rapid adaptation capacity inherent in MB strategies. Finally, to assess mental simulation coupled with prefrontal meta-control, we designed a stochastic *Atari-Pong* variant that incorporates decision-making parameters from the *Two-Stage MDT* (Figure 2).

4.1 Challenges: environmental uncertainty and goal condition dynamics

Our research addresses two fundamental challenges in decision-making processes that arise from recent neuroscience tasks: environmental uncertainty and goal-state dynamics. Within a MDP framework, environmental changes manifest through continuous modifications in reward conditions $R(s, a)$. This formulation establishes that the agent's learning objective consists of achieving specific state conditions, which remains fundamental to strategy acquisition and behavioral adaptation (Lee et al., 2014, 2019; Kim et al., 2019).

Goal conditions serve as critical metrics for assessing RL agents' flexibility within dynamic environments. These conditions manifest in two forms: specific criteria that require precise reward thresholds, and flexible criteria that accommodate varied outcomes normalized to values 0, $\frac{1}{4}$, $\frac{1}{2}$, 1, which are contingent upon received rewards (Lee et al., 2014; Kim et al., 2019, 2021). This framework necessitates continuous policy re-evaluation and adaptive response refinement through iterative learning mechanisms.

An alternative representation of environmental dynamics involves modifying state transition probabilities $P(s' \mid s, a)$ within the MDP across specified ranges at predetermined intervals. This approach provides quantitative modeling of environmental dynamics, where state-transition probability serves as a crucial metric for uncertainty representation. The framework encompasses varying uncertainty levels: from near-deterministic outcomes with 90% probability (low uncertainty) to equiprobable outcomes of 50% (high uncertainty). These probabilistic structures reflect diverse environmental scenarios, which prove essential for addressing unpredictable changes.

This stochastic framework enables optimal action selection under dynamic conditions whilst facilitating predictive planning of future scenarios. Thus, state transition probabilities constitute fundamental factors in determining an agent's adaptive capacity within changing environments. Human decision-making demonstrates robust adaptation to varying goal conditions and state transition probabilities through prefrontal meta-control. To evaluate RL agents' adaptability, we employ MDP tasks that embody these environmental dynamics, particularly the Two-Stage MDT (Figure 2A), which incorporates both goal condition and state-transition probability variations.

These experimental paradigms comprise alternating phases of varying goal criteria and uncertainty levels, which challenge adaptive capabilities across environmental shifts. To extend these stochastic frameworks into general domains, we integrated these challenges into two distinct environments: *GridWorldLoca*, a

benchmark suite for MB RL agents (Figure 2B), and a modified *Atari Pong* environment (Figure 2D). These implementations instantiate environmental uncertainty and goal-state dynamics within standardized testing paradigms.

4.2 Two-stage Markov decision task

4.2.1 Overview

The Two-Stage MDT (Lee et al., 2014) represents a widely accepted paradigm in human decision-making research, where participants execute sequential actions to obtain color-coded tokens that correspond to specific rewards. This experimental environment consists of two distinct conditions: *Specific*– and *Flexible* Goal conditions. The former necessitates MB strategies for reward maximization, whilst the latter facilitates MF strategy utilization. We adopted this paradigm for computational evaluation of agent adaptability. Our implementation maintains fidelity with the original experimental design, incorporating both *Goal Conditions* and *State-Transition Uncertainty* parameters.

As illustrated in Figure 2A, the Goal Condition bifurcates into *Specific Goal* and *Flexible Goal* variants. Under the Specific Goal condition, the agent receives a binary reward (1.0 for required token acquisition, 0.0 otherwise). Conversely, the Flexible Goal condition implements a normalized reward spectrum 0, $\frac{1}{4}$, $\frac{1}{2}$, 1 based on token acquisition, without specific token requirements. State-Transition Uncertainty emerges in two forms: *Low* and *High*. Low uncertainty conditions encompass deterministic action execution with 0.9 probability, whilst High Uncertainty conditions exhibit equiprobable action outcomes (0.5 probability for intended and opposite actions).

4.2.2 Experimental setting

The Two-Stage MDT structure, as presented in Figure 2A, comprises sequential binary actions (left/right), where the initial action results in a state representation without reward, whilst the subsequent action generates stochastic rewards based on environmental parameters before episode termination.

The simulation protocol proceeds as follows (Figure 3A). During the initial 500 episodes, the Goal Condition exhibits pseudo-random alterations (uniformly distributed) at fixed 125-episode intervals, maintaining Low State-Transition Uncertainty. This configuration facilitates situations requiring rapid adaptation. The subsequent 500 episodes exhibit high state-transition uncertainty, where goal conditions become less salient, compelling the agent to develop reward-maximizing policies focusing on high-value tokens. This 1,000-episode sequence repeats five times, resulting in 5,000 total episodes, thus enabling comprehensive evaluation of adaptive capabilities across diverse environmental conditions.

4.2.3 Evaluation metrics

The assessment of *Meta-Dyna*'s performance adopted comparative analysis against baseline algorithms: *Dyna-Q*, *Q-learning*, and *FORWARD*. The reward structure covers a normalized range [0.0, 1.0]. The evaluation metrics includes: (i) mean reward and (ii) choice optimality that quantifies the proportion of episodes where the agent achieves maximum possible reward under specific Goal Conditions (Kim et al., 2019). The choice optimality is formally defined as:

$$\text{Choice Optimality} = \frac{\text{Number of episodes with maximum reward}}{\text{Total number of episodes}}. \tag{7}$$

4.2.4 Result

Figure 3 presents the result of the experiment. Figure 3B demonstrates that *Meta-Dyna* achieved superior mean rewards to that of baseline algorithms (*Meta-Dyna*: 0.61, *Dyna-Q*: 0.55, Q-learning: 0.52, FORWARD: 0.54; $p < 0.0001$, independent t-test). The clustered distribution of data points suggests *Meta-Dyna*'s robust reward acquisition across environmental variations. The optimality analysis (Figure 3C) shows *Meta-Dyna*'s statistically significant performance on optimal decision over baseline models. This indicates enhanced decision-making capacity under changes in Goal Conditions and State-Transition Uncertainties. These results establish *Meta-Dyna*'s superior performance across all metrics thereby validating the efficacy of meta-control integration within the *Dyna-Q* framework.

We note that the results presented in Figure 3 derive from the tabular implementation of *Meta-Dyna*. Subsequent evaluation using the neural network variant exhibits comparable superiority, with *Meta-Dyna* demonstrating significantly higher mean rewards (*Meta-Dyna*: 0.71, DQN: 0.61, *Dyna-Q*: 0.66; $p < 0.0001$, independent sample t-test).

4.3 A stochastic *GridWorldLoCA*

4.3.1 Overview

The stochastic *GridWorldLoCA* environment, which originally evaluates MB agents' sample efficiency and adaptive capacity (Wan et al., 2022), exhibits dynamic environmental conditions. This framework comprises distinct phases that are characterized by variations in both initial state distribution (presented as the green region) and reward configurations (presented as black vertical bars at both edges in Figure 2B).

We parameterised the *GridWorldLoCA* environment's complexity through modification of state-transition probabilities, deriving from the Two-Stage MDT. The framework exhibits binary probability distributions—(0.9, 0.1) for the deterministic- and (0.5, 0.5) for the uncertain configuration—which introduce stochastic elements to agent locomotion (Figure 2B). When an agent attempts to move toward a target state, the deterministic configuration assigns 0.9 probability to the intended direction and 0.1 probability distributed across remaining directions. Similarly, in the uncertain configuration, both the intended direction and remaining directions receive equal probabilities of 0.5.

4.3.2 Experimental settings

The experimental procedure consists of two blocks each of which has three sequential phases. One block repeats twice

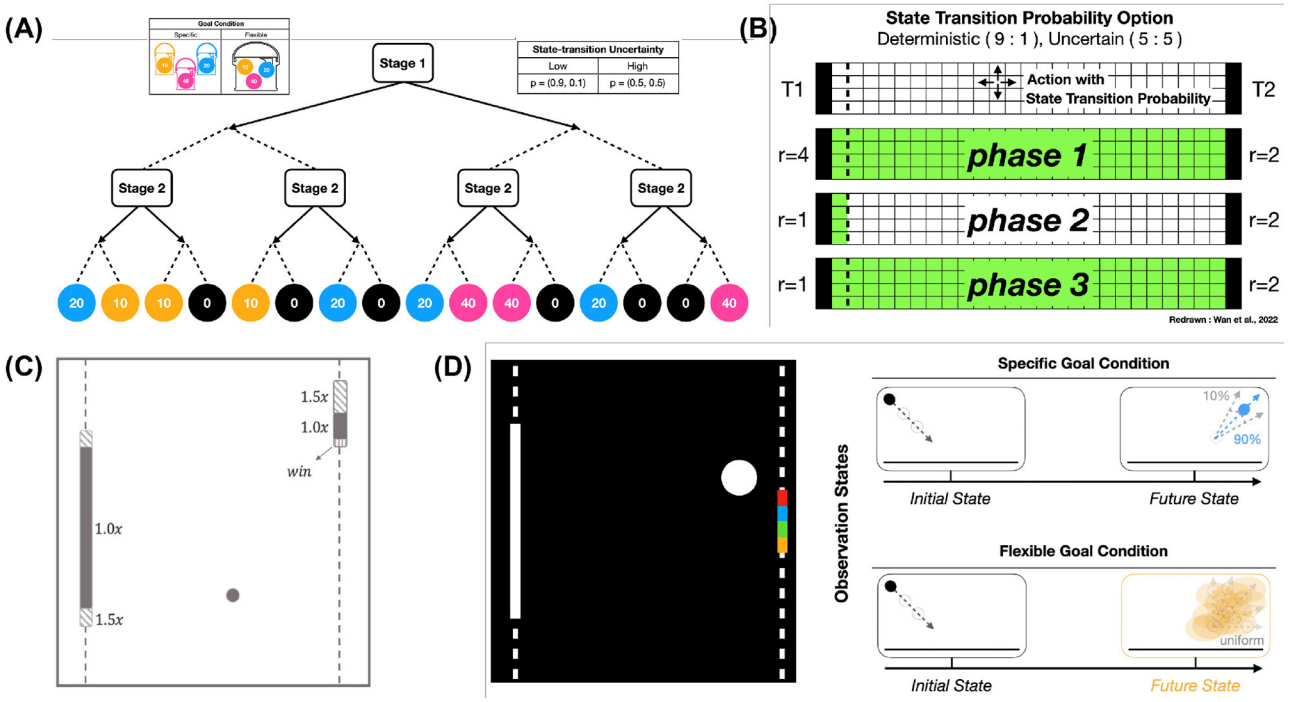

FIGURE 2
Meta-Dyna tasks for experiments. **(A)** Two-stage Markov decision task. **(B)** A variant of GridWorldLoca. **(C)** A dimension of Atari Pong variant (© copyright, Keramati et al., 2018). **(D)** A novel Atari Pong incorporating decision uncertainty and frequent goal changes.

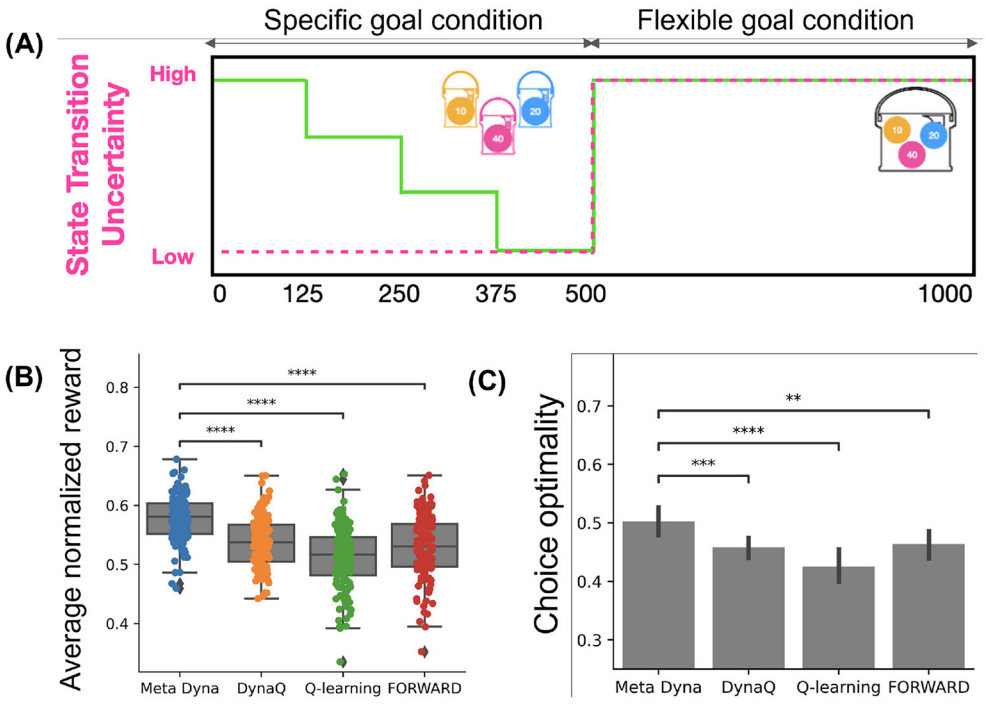

FIGURE 3
Simulation result. **(A)** Experimental setting. **(B)** Result on average normalized reward. **(C)** Results on choice optimality. ($^{*}: p < 0.05, ^{**}: p < 0.01, ^{***}: p < 0.001, ^{****}: p < 0.0001$; independent t-test).

during training with distinct state-action-state transition probabilities (please refer to the sequence of phases at the bottom of Figure 4). Each phase introduces novel environmental parameters, characterized by modifications to both the initial state distribution (shown in green) and reward structure (shown in black vertical bars), thus challenging the agent's adaptive capacities.

To rigorously assess adaptive capacity within rapidly changing environments, we implemented a significant reduction in episodic duration per phase—to 3% of the baseline configuration. This constraint intensifies task complexity whilst necessitating accelerated adaptation with limited sampling.

Our experimental evaluation entails comparative analyses across multiple agent architectures: pure MB and MF implementations (FORWARD, *Dyna-Q*, and SARSA, respectively), the proposed *Meta-Dyna*, and hybrid architectures that integrate FORWARD with SARSA or Q-learning. The experimental protocol consists of dual traversals through Phases 1 to 3, with ten complete iterations that establish statistically significant reliability ($p < 0.05$) of performance metrics, including cumulative rewards and policy adaptation speeds.

4.3.3 Result

As illustrated in Figure 4, standalone agents (MB and MF) exhibited suboptimal performance (with the exception of *Dyna-Q*), indicating insufficient adaptability to environmental dynamics. These agents demonstrated particular difficulty in policy adjustment relative to varying initial states and reward structures across experimental phases. Conversely, a family of meta-control agents exhibited rapid adaptation whilst maintaining robust performance throughout all phases.

Meta-Dyna demonstrated superior performance, achieving elevated average returns and accelerated optimal policy convergence compared to those of *Dyna-Q* (average return—*Meta-Dyna* > *Dyna-Q*; $p < 0.05$, trial for success—*Meta-Dyna* < *Dyna-Q*; $p < 0.05$, independent t-test). These results suggest that *Meta-Dyna*'s meta-control mechanism effectively modulates the integration of MB and MF strategies, thereby facilitating enhanced adaptivity compared to that of *Dyna-Q*. In addition, *Meta-Dyna* exhibits more stable learning behavior (variance: 0.0191) than other models (variance: 0.0733), which allows it to maintain consistency across varying environmental conditions without sacrificing adaptability.

These findings demonstrate that the application of meta-control mechanisms to RL agents brings about significant enhancement in adaptive capabilities compared to standalone implementations. Through effective integration of MB and MF strategies, *Meta-Dyna* achieves rapid policy adaptation in response to environmental dynamics, even under the limited number of training episodes. This experimental evaluation validates *Meta-Dyna*'s efficacy in enhancing sample efficiency and adaptive capabilities within dynamic environments.

4.4 Stochastic *Atari Pong*: a probabilistic extension

To evaluate rapid adaptation capacity under the challenges defined in Section 4.1, we developed a *stochastic Atari Pong* environment. This novel framework, which extends the standard Gym, implements state-transition uncertainty and goal-condition dynamics that derive from the Two-Stage MDT.

4.4.1 Experimental setting

Our *stochastic Atari Pong* implementation introduces terminal conditions (episode completion) at single-point acquisition (score of 1), diverging from the traditional 21-point rally structure (Figures 2C, D left). This framework implements two fundamental modifications: state-transition uncertainty and goal condition dynamics.

The state-transition uncertainty manifests through dual mechanisms: paddle locomotion and ball reflection angles. Under deterministic conditions, these parameters maintain high predictability, enabling probabilistic trajectory anticipation (Figure 2D, upper right). Conversely, under stochastic conditions, the introduction of randomness to both parameters necessitates adaptive strategies focused on real-time ball tracking rather than trajectory prediction (Figure 2D, bottom right).

The goal condition framework entails two distinct paradigms. The specific goal condition necessitates ball contact within one of four predetermined paddle segments (uniformly divided quarters of the 20-pixel paddle length, Figure 2D), which results in immediate terminal conditions and reward allocation (+1 for target segment, 0 for others) upon successful execution. In contrast, the flexible goal condition implements standard victory conditions, where reward acquisition occurs upon opponent failure to return the ball.

These environmental parameters are subject to systematic modification (alternating between deterministic and stochastic configurations) at 1,000-timestep intervals throughout the experimental protocol, thus necessitating continuous adaptation to dynamic conditions. This framework requires the RL agent to exhibit rapid adjustment capacity (convergence within 100–200 timesteps) in response to both state-transition uncertainty and goal-condition variations. The experimental evaluation entails a comparative analysis between *Meta-Dyna* and two baseline architectures: DQN and *Dyna-Q*.

In addition, we conducted another experiments to investigate the impact of mental simulation on *Meta-Dyna*'s performance. As described in Section 3.5, *Meta-Dyna* employs a default of 10 mental simulations using the world model, which was maintained across all previous experiments. In this environment, we specifically examined how the number of mental simulations influences performance with respect to average reward (pertaining to the main RL objective) and trials required for success (relating to sample efficiency). The latter metric strongly corresponds to sample efficiency when the RL agent receives a fixed number of real experiences. Specifically, it measures how quickly the agent achieves maximum rewards relative to environmental interactions. We posit that sample efficiency increases as the required number

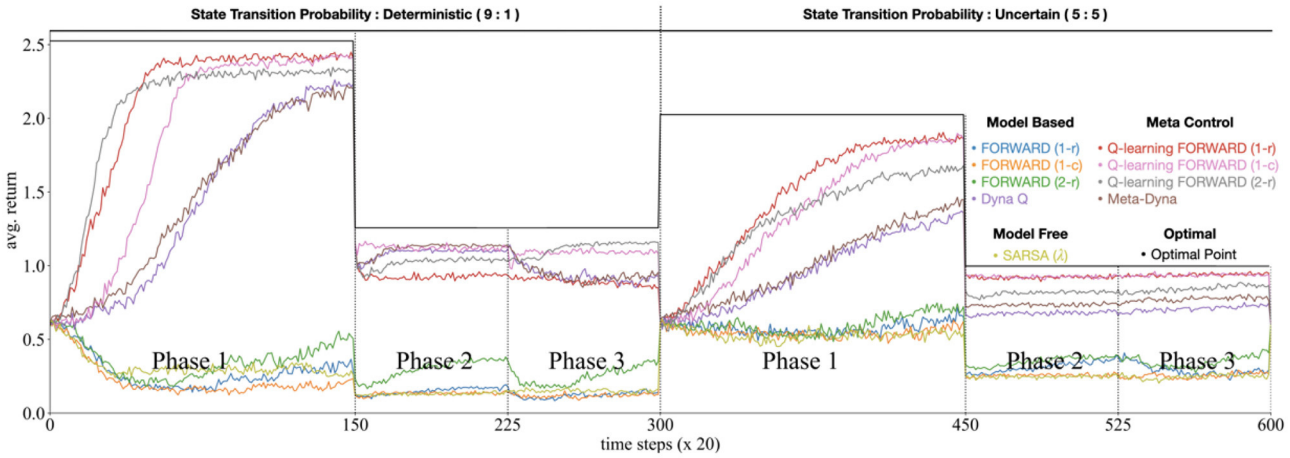

FIGURE 4
Simulation result on the stochastic *GridWorldLoCA*.

of environmental interactions decreases for achieving maximum reward.

To evaluate this relationship, we parameterised the number of mental simulations and conducted four experimental conditions (n = 10, 20, 50, 100) with a fixed number of environmental interactions (i.e., real experiences). We then analyzed the data using the Number of Trials for Success (NTS) metric, defined as:

$$NTS = \frac{R_{max}}{N_{env-interaction}}, \tag{8}$$

where R_{max} denotes the maximum reward achieved by the RL agent, and $N_{env-interaction}$ refers to the number of environmental interactions (equivalent to the number of real experiences).

4.4.2 Results

As illustrated in Figure 5, DQN exhibits limited adaptive capabilities (mean reward $<$ 0.4 across all test conditions) within the dynamic environment. The architecture fails to establish policy stability under environmental perturbations (state-transition probability shifts from 0.9 to 0.5 and goal condition alterations), showing only gradual performance improvements after extensive training episodes. This behavior indicates DQN's inherent limitations in rapid adaptation without environmental modeling capabilities.

Both *Meta-Dyna* and *Dyna-Q* exhibit enhanced performance (average policy convergence within 150 timesteps) through their integrated environmental models, which facilitate efficient learning through dynamic environment simulation. Despite similar performance trajectories in the initial training phase (first 1,000 episodes, learning rate $\approx$ 0.15), *Meta-Dyna* achieves superior cumulative rewards compared with both *Dyna-Q* and DQN (Figure 5B). Quantitative analysis reveals the following cumulative rewards over the entire 5,000-episode evaluation period: Original Reward—*Meta-Dyna*: −0.091, *Dyna-Q*: −0.132, DQN: −0.782; Exponential Reward—*Meta-Dyna*: 0.913, *Dyna-Q*: 0.876, DQN: 0.457 ($p < 0.001$, independent samples t-test).

Meta-Dyna exhibits superior adaptive capabilities, achieving elevated rewards across all goal conditions. Through effective utilization of its meta-control mechanism, it consistently surpasses both *Dyna-Q* and DQN performance metrics. Despite implementing a more parsimonious parameter structure (single-layer neural network with 128 units compared to LSTM networks with $>$ 500K parameters) compared with sequence-learning architectures such as Meta-RL (Wang et al., 2016), which employ LSTM mechanisms, *Meta-Dyna* achieves comparable or superior cumulative rewards without additional memory architectures. These findings indicate *Meta-Dyna*'s robust generalization capabilities (maintaining $>$ 85% performance across novel environmental configurations with $<$ 200 timesteps adaptation period).

With regards to the impact of mental simulation, Figure 5D illustrates that larger numbers of mental simulations contribute to higher performance given fixed real experiences. This results in two principal conclusions: increasing mental simulations correlates with enhanced performance, and higher numbers of mental simulations improve sample efficiency.

Figure 5E demonstrates *Meta-Dyna*'s superiority in this domain. Whilst baseline models (DQN and *Dyna*) converge to sub-optimal points, all variants of *Meta-Dyna* achieve optimal convergence more rapidly. Moreover, increased mental simulations correlate with faster convergence times and higher reward acquisition. These results indicate that *Meta-Dyna*'s mental simulation mechanism shows promise for improving both environmental adaptivity and sample efficiency.

Figure 5F emphasizes the superiority of *Meta-Dyna*, which implements prefrontal meta-control incorporating mental simulation, compared with the vanilla *Dyna* architecture. As described in Section 2, models of *Dyna* (e.g., *Dyna-Q*) are also equipped with mental simulation capacity. Thus, one might expect performance benefits from increasing *Dyna-Q*'s mental simulations. However, the results reveal a different outcome: even 100 mental simulations of *Dyna-Q* fail to converge toward the optimal point, which *Meta-Dyna* achieves with merely 20

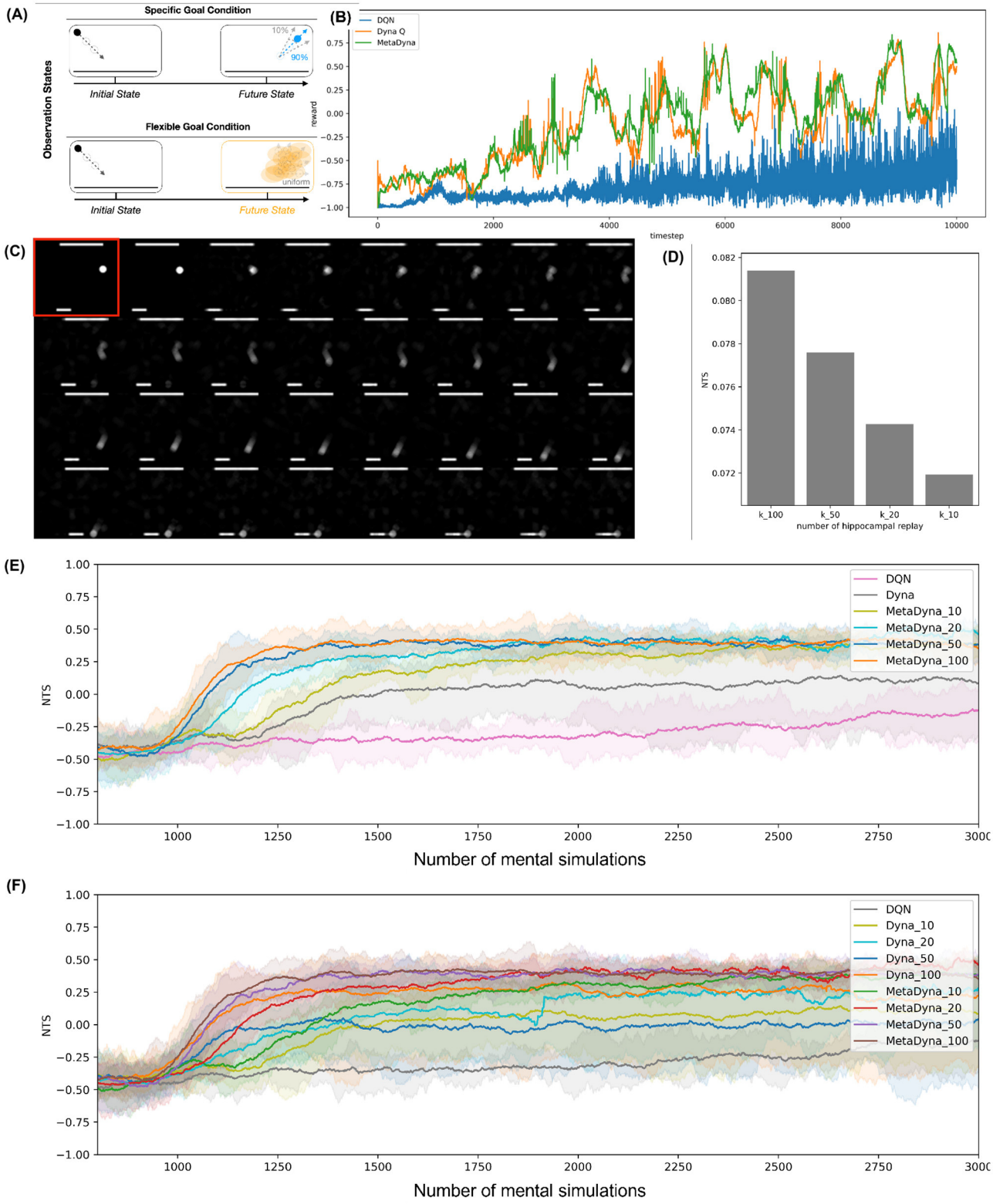

FIGURE 5
Simulation result on the stochastic *Atari Pong*. **(A)** Two goal conditions coupled with environmental uncertainty. **(B)** Result on three types of RL agents. The X-axis refers to the reward, and the Y-axis refers to the episode. **(C)** The internal representation of the world model. **(D)** Result on the mental simulation. The X-axis refers to the number of mental simulation, and the Y-axis means the number of trials for success (NTS). **(E)** The NTS plot for three types of RL agents. Here, as a baseline model, *Dyna* does not use the function of mental simulation i.e., $n = 1$. *Meta-Dyna* utilizes the function of mental simulation ($n = 10, 20, 50, 100$). **(F)** The NTS plot for three types of RL agents with enabling the function of mental simulation for both *Dyna* and *Meta-Dyna*.

simulations. Furthermore, given equivalent numbers of mental simulations, *Meta-Dyna* consistently outperforms *Dyna-Q*. This discrepancy in performance suggests that prefrontal meta-control contributes significantly to the efficiency differential, despite identical mental simulation functionality. These findings highlight *Meta-Dyna*'s superior sample efficiency and adaptivity through its brain-inspired computational approach.

Overall, the experimental results from our stochastic Atari-Pong environment validate *Meta-Dyna*'s sample efficiency as well as adaptive capacity. Through the incorporation of state-transition uncertainty and goal-condition variability, this framework provided rigorous evaluation of adaptive capabilities within complex, dynamic environments. *Meta-Dyna*'s sustained performance across these conditions substantiates the efficacy of meta-control mechanisms in enhancing RL architectures.

5 Discussions and conclusion

This research presents *Meta-Dyna*, a novel prefrontal meta-control framework incorporating mental simulation for RL agents that improves adaptivity and behavioral flexibility. Grounded in neuroscience, it emulates prefrontal cortex-mediated arbitration control mechanisms (Lee et al., 2014) and hippocampal functions (Stachenfeld et al., 2017). The architecture enables both habitual and goal-directed behaviors through integrated Q-learning and planning processes, whilst exhibiting adaptive capacity with sample efficiency through mental simulations under dynamic environmental changes.

The empirical investigation of *Meta-Dyna* entailed three experimental paradigms: the *Two-Stage MDT*, *Stochastic GridWorldLoCA*, and a *Stochastic Atari-Pong* environment. These frameworks examined cognitive adaptability and computational efficiency through systematic modulation of environmental uncertainty and goal condition dynamics, which derive from the two-stage MDT, an established human decision-making paradigm (Daw et al., 2005; Lee et al., 2014). Within the *Two-Stage MDT*, *Meta-Dyna* exhibited elevated performance metrics relative to baseline architectures in reward acquisition and choice optimisation.

The *stochastic GridWorldLoCA* evaluation, which examined rapid cognitive adaptation capacity, substantiated the computational efficacy of *Meta-Dyna*. Through its prefrontal meta-control mechanisms, the architecture manifested heightened behavioral flexibility relative to standalone implementations (FORWARD and SARSA). *Meta-Dyna* surpassed *Dyna-Q* in both reward maximization and policy optimisation speed, thus establishing its elevated adaptive capacity.

However, *Meta-Dyna* was not always dominant in the *stochastic GridWorldLoCA*. Compared to a family of *Q-learning FORWARD*, *Meta-Dyna* demonstrated the lower performance across all phases. In brief, we assume that this phenomenon was caused by differences in implementation methods—direct computation vs. approximation of state-action-state transition probabilities (state-transition probability hereinafter), that is, Tabular RL vs. approximate RL in a broad sense. Although the conceptual framework between these approaches is identical, this implementation difference brings about a discrepancy due to the approximation error which in the end affects the results of the average reward across the stages (for more details, see Supplementary Section 1.1).

The *stochastic Atari-Pong* environment incorporated multidimensional complexity to examine behavioral flexibility and computational efficiency. *Meta-Dyna* manifested proficient arbitration of state-transition uncertainty and variable goal conditions, resulting in elevated performance relative to *Dyna-Q* and DQN architectures. In particular, *Meta-Dyna* exhibited exceptional computational efficiency in experience utilization. The environmental model instantiated dynamic mental simulations, engendering heightened proficiency in both reward accumulation and trial-success speed compared with baseline architectures.

Whilst the environmental model implemented in *Meta-Dyna* enables sequence modeling through MDN-RNN architecture, rapid environmental dynamics lead to sub-optimal sequence predictions. Performance enhancement could be achieved through transformer integration (Vaswani et al., 2017), facilitating more efficient and precise predictions (Radford et al., 2019; Chen et al., 2021). Moreover, the current LSTM-based transition probability approximation within the MDN-RNN framework may benefit from direct probability computation approaches (Hafner et al., 2019). Alternative architectures, such as Recurrent State Space Models (RSSM) (Hafner et al., 2019) or Transformer State Space Models (TSSM) (Chen et al., 2022), could produce enhanced MDP prediction accuracy.

Regarding decision architectures, *Meta-Dyna*'s Q-value foundation could be extended to incorporate recent policy-based achievements, including Trust Region Policy Optimization (TRPO) (Schulman et al., 2015), Proximal Policy Optimization (PPO) (Schulman et al., 2017), and Soft Actor-Critic (SAC) (Haarnoja et al., 2018). Integration of meta-control mechanisms within these frameworks could facilitate rapid optimal action selection under environmental perturbations.

Although *Meta-Dyna* demonstrated superior efficiency metrics relative to baseline models, its single-step environmental model approximation resulted in sub-optimal absolute performance. Implementation of *n*-step approximation, similar to Imagination-Augmented Agents (I2A) (Racanière et al., 2017), could potentially enhance optimality convergence.

The computational cost of mental simulation primarily depends on the number of simulated rollouts. Our analysis shows that increasing the number of mental simulations enhances performance and sample efficiency (Figure 5D). However, increasing the number of rollouts does not guarantee a proportional performance gain relative to the additional computational cost. Considering that the only difference between real experience and mental simulation in our framework is the number of rollouts, we can optimize computational efficiency by dynamically regulating the number of simulations based on task urgency. This adaptive approach ensures that Meta-*Dyna* remains feasible even for time-critical tasks.

Another alternative under time constraint would be the use of an asynchronous approach in which mental simulation is performed by concurrent multiple threads. Like A3C architecture (Mnih et al., 2016), mental simulations in MB

RL could be carried out through the parallel execution of hypothetical experiences concurrently, which implements a constant time block to do the rollout. The asynchronous approach offers significant advantages by decorrelating the data into a more stationary process, as parallel agents experience different states simultaneously. This allows for more stable learning while achieving nearly linear speedups with the number of parallel threads employed. Thus, it would be able to achieve the maximum rewards within the time constraint by executing MF with real experiences and MB with simulated experiences. We note that the computational cost in the form of Floating Point Operations Per Second (FLOPS) was not measured in this study, which we will be able to do in the future coupled with an asynchronous approach.

In conclusion, this research presents a neuroscience-inspired RL agent that demonstrates not only rapid environmental adaptation through the parallel implementation of MB and MF strategies, but also exceptional sample efficiency achieved via mental simulation using an RNN-based world model. The agent's architecture, featuring a meta-control mechanism that parallels Acceptance and Commitment Therapy (ACT)'s emphasis on cognitive flexibility (Banerjee et al., 2021), highlights its potential relevance to computational psychiatry and suggests that adaptive regulation may offer valuable insights into cognitive impairments and therapeutic applications (Mei et al., 2025). These findings contribute to our understanding of human RL and the development of biomimetic learning agents. Furthermore, exploring its application in AI-driven cognitive training and behavioral interventions, particularly in the context of digital therapeutics, will be an important avenue for future research (Muyesser et al., 2018; Lee et al., 2024).

Moreover, future investigations will also examine whether Meta-*Dyna*'s mental simulation mechanism can be adapted to model and potentially mitigate maladaptive mental simulation—such as the negative rumination and rigid thought patterns observed in depression (Heo et al., 2021; Senta et al., 2025)—thus providing further insights into pathological cognitive rigidity. Recent work by Heo et al. (2021) demonstrated that depression disrupts arbitration control between MB and MF learning whilst undermining exploitation sensitivity. Similarly, Senta et al. (2025) identified a dual-process impairment in physiological anxiety affecting both reinforcement learning rates and working memory decay. These findings suggest several potential adaptations for Meta-*Dyna*: incorporating variable learning rates that account for altered prediction error processing; modeling impaired arbitration control between simulation strategies; representing working memory constraints; and implementing asymmetric value updating parameters. Such modifications could help simulate how depressive rumination develops when mental simulation becomes trapped in negative feedback loops with diminished capacity to integrate new information or shift between cognitive strategies.

Ultimately, with such computational models of human reinforcement learning, we could provide the means for modeling patients' minds from a computational psychiatry perspective. Once we succeed in this endeavor, we might be able to simulate various treatment regimens using Meta-*Dyna*, which creates opportunities to optimally generate behavioral therapy approaches for patients with precise medication and treatment recommendations.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

JK: Data curation, Formal analysis, Software, Visualization, Writing – original draft. JL: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This study was supported by (i) Korea Institute of Police Technology (KIPoT; Police Lab 2.0 program) grant funded by MSIT (RS-2023-00281194; Development for a remotely-operated testimony system for the children based on AI and Cloud technology) and (ii) a Research Grant (2024-0035) funded by HAII Corporation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. We used chatGPT to perform the proofread including typos and grammatical errors.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2025.1559915/full#supplementary-material

References

Abbott, L. F., and Dayan, P. (2001). Theoretical neuroscience. *Comput. Math. Model Neural.* 60, 489–495. doi: 10.1016/j.neuron.2008.10.019

Allen, K. R., Smith, K. A., and Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proc. Nat. Acad. Sci.* 117, 29302–29310. doi: 10.1073/pnas.1912341117

Banerjee, A., Rikhye, R. V., and Marblestone, A. (2021). Reinforcement-guided learning in frontal neocortex: emerging computational concepts. *Curr. Opin. Behav. Sci.* 38, 133–140. doi: 10.1016/j.cobeha.2021.02.019

Bansal, S., et al. (2017). Mbmf: Model-based priors for model-free reinforcement learning. *arXiv preprint arXiv:1709.03153*.

Barkley, B., and Fridovich-Keil, D. (2024). Stealing that free lunch: exposing the limits of dyna-style reinforcement learning. *arXiv preprint arXiv:2412.14312*.

Budiyanto, A., and Matsunaga, N. (2023). Deep dyna-q for rapid learning and improved formation achievement in cooperative transportation. *Automation* 4, 210–231. doi: 10.3390/automation4030013

Chen, C., et al. (2022). Transdreamer: reinforcement learning with transformer world models. *arXiv preprint arXiv:2202.09481*.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., et al. (2021). "Decision transformer: reinforcement learning via sequence modeling," in *Advances in Neural Information Processing Systems*, 15084–15097.

Daw, N. D., and Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philos. Trans. R. Soc.* 369:20130478. doi: 10.1098/rstb.2013.0478

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69, 1204–1215. doi: 10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* 8, 1704–1711. doi: 10.1038/nn1560

Dayan, P., and Berridge, K. C. (2014). Model-based and model-free pavlovian reward learning: revaluation, revision, and revelation. *Cogn. Affect. Behav. Neurosci.* 14, 473–492. doi: 10.3758/s13415-014-0277-8

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philos. Trans. R. Soc. London B* 308, 67–78. doi: 10.1098/rstb.1985.0010

Dolan, R. J., and Dayan, P. (2013). Goals and habits in the brain. *Neuron* 80, 312–325. doi: 10.1016/j.neuron.2013.09.007

Dong, L., Li, Y., Zhou, X., Wen, Y., and Guan, K. (2020). Intelligent trainer for dyna-style model-based deep reinforcement learning. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 2758–2771. doi: 10.1109/TNNLS.2020.3008249

Ghode, S. D., and Digalwar, M. (2024). Deep dyna reinforcement learning based energy management system for solar operated hybrid electric vehicle using load scheduling technique. *J. Energy Storage* 102:114106. doi: 10.1016/j.est.2024.114106

Gupta, A. S., Van Der Meer, M. A., Touretzky, D. S., and Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. *Neuron* 65, 695–705. doi: 10.1016/j.neuron.2010.01.034

Ha, D., and Schmidhuber, J. (2018). World models. *arXiv preprint arXiv:1803.10122*.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). "Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International Conference on Machine Learning* (PMLR), 1861–1870.

Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., et al. (2019). "Learning latent dynamics for planning from pixels," in *International Conference on Machine Learning*, 2555–2565.

Heo, S., Sung, Y., and Lee, S. W. (2021). Effects of subclinical depression on prefrontal-striatal model-based and model-free learning. *PLoS Comput. Biol.* 17:e1009003. doi: 10.1371/journal.pcbi.1009003

Jadhav, S. P., Kemere, C., German, P. W., and Frank, L. M. (2012). Awake hippocampal sharp-wave ripples support spatial memory. *Science* 336, 1454–1458. doi: 10.1126/science.1217230

Karlsson, M. P., and Frank, L. M. (2009). Awake replay of remote experiences in the hippocampus. *Nat. Neurosci.* 12, 913–918. doi: 10.1038/nn.2344

Keramati, R., Whang, J., Cho, P., and Brunskill, E. (2018). Fast exploration with simplified models and approximately optimistic planning in model based reinforcement learning. *arXiv preprint*. arXiv:1806.00175.

Kim, D., Jeong, J., and Lee, S. W. (2021). Prefrontal solution to the bias-variance tradeoff during reinforcement learning. *Cell Rep.* 37:110185. doi: 10.1016/j.celrep.2021.110185

Kim, D., Lee, J. H., Jung, W., Kim, S. H., and Lee, S. W. (2023). Long short-term prediction guides human metacognitive reinforcement learning. *Res Sq.* doi: 10.21203/rs.3.rs-3080402/v1

Kim, D., Park, G. Y., O.', Doherty, J. P., and Lee, S. W. (2019). Task complexity interacts with state-space uncertainty in the arbitration between model-based and model-free learning. *Nat. Commun.* 10:5738. doi: 10.1038/s41467-019-13632-1

Kool, W., Gershman, S. J., and Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychol. Sci.* 28, 1321–1333. doi: 10.1177/0956797617708288

Krugel, L. K., Biele, G., Mohr, P. N., Li, S.-C., and Heekeren, H. R. (2009). Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proc. Nat. Acad. Sci.* 106, 17951–17956. doi: 10.1073/pnas.0905191106

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: a selective review and a hybrid model. *Quart. J. Exper. Psychol. Section B* 57, 193–243. doi: 10.1080/02724990344000141

Lee, J. H., Heo, S. Y., and Lee, S. W. (2024). Controlling human causal inference through in silico task design. *Cell Rep.* 43:113702. doi: 10.1016/j.celrep.2024.113702

Lee, J. H., Leibo, J. Z., An, S. J., and Lee, S. W. (2022). Importance of prefrontal meta control in human-like reinforcement learning. *Front. Comput. Neurosci.* 16:1060101. doi: 10.3389/fncom.2022.1060101

Lee, J. H., Seymour, B., Leibo, J. Z., An, S. J., and Lee, S. W. (2019). Toward high-performance, memory-efficient, and fast reinforcement learning—lessons from decision neuroscience. *Sci. Robot.* 4:eaav2975. doi: 10.1126/scirobotics.aav2975

Lee, S. W., Shimojo, S., and O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron* 81, 687–699. doi: 10.1016/j.neuron.2013.11.028

Li, J., et al. (2011). Differential roles of human striatum and amygdala in associative learning. *Nat. Neurosci.* 14, 1250–1252. doi: 10.1038/nn.2904

Li, Y., Dong, Z., Luo, E., Wu, Y., Wu, S., and Han, S. (2024). When to trust your data: enhancing dyna-style model-based reinforcement learning with data filter. *arXiv preprint arXiv:2410.12160*.

Liu, H., Yao, Y., Li, T., Du, M., Wang, X., Li, H., et al. (2024). Dyna algorithm-based reinforcement learning energy management for fuel cell hybrid engineering vehicles. *J. Energy Storage* 94:112526. doi: 10.1016/j.est.2024.112526

Liu, W., Wang, Y., Xu, C., and Zheng, M. (2025). A smart grid computational offloading policy generation method for end-edge-cloud environments. *J. Reliable Intell. Environ.* 11, 1–13. doi: 10.1007/s40860-025-00243-5

Liu, X.-Y., and Wang, J.-X. (2021). Physics-informed dyna-style model-based deep reinforcement learning for dynamic control. *Proc. R. Soc. A* 477:20210618. doi: 10.1098/rspa.2021.0618

Liu, Y., Dolan, R. J., Kurth-Nelson, Z., and Behrens, T. E. (2019). Human replay spontaneously reorganizes experience. *Cell* 178, 640–652. doi: 10.1016/j.cell.2019.06.012

Mattar, M. G., and Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* 21, 1609–1617. doi: 10.1038/s41593-018-0232-z

Mei, J., Rodriguez-Garcia, A., Takeuchi, D., Wainstein, G., Hubig, N., Mohsenzadeh, Y., et al. (2025). Improving the adaptive and continuous learning capabilities of artificial neural networks: Lessons from multi-neuromodulatory dynamics. *arXiv preprint arXiv:2501.06762*.

Miller, K. J., Botvinick, M. M., and Brody, C. D. (2017). Dorsal hippocampus contributes to model-based planning. *Nat. Neurosci.* 20, 1269–1276. doi: 10.1038/nn.4613

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., et al. (2016). "Asynchronous methods for deep reinforcement learning," in *Proceedings of The 33rd International Conference on Machine Learning, volume 48 of Proceedings of Machine Learning Research*, eds. M. F. Balcan, and K. Q. Weinberger (New York, New York, USA: PMLR), 1928–1937.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., and Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nat. Hum. Behav.* 1, 680–692. doi: 10.1038/s41562-017-0180-8

Muyesser, N. A., Dunovan, K., and Verstynen, T. (2018). Learning model-based strategies in simple environments with hierarchical q-networks. *arXiv preprint arXiv:1801.06689*.

Pearce, J. M., and Hall, G. (1980). A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* 87:532. doi: 10.1037//0033-295X.87.6.532

Pfeiffer, B. E., and Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* 497, 74–79. doi: 10.1038/nature12112

Qu, X., Liu, L., and Huang, W. (2025). Data-driven inventory management for new products: a warm-start and adjusted dyna-*q* approach. *arXiv preprint arXiv:2501.08109*.

Racaniére, S., Weber, T., Reichert, D., Buesing, L., Guez, A., Jimenez Rezende, D., et al. (2017). "Imagination-augmented agents for deep reinforcement learning," in *Advances in Neural Information Processing Systems*, 30.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog* 1:9.

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., and Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Comput. Biol.* 13:e1005768. doi: 10.1371/journal.pcbi.1005768

Saeed, M. H., Kazmi, H., and Deconinck, G. (2024). Dyna-pinn: physics-informed deep dyna-q reinforcement learning for intelligent control of building heating system in low-diversity training data regimes. *Energy Build.* 324:114879. doi: 10.1016/j.enbuild.2024.114879

Samaylal, S., (2024). *Real-time digital twin with reinforcement learning for industrial manipulator applications* (Master of Science Thesis). Tampere University.

Schulman, J., et al. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). "Trust region policy optimization," in *Proceedings of the 32nd International Conference on Machine Learning, Vol. 37*, 1889–1897.

Senta, J., Bishop, S. J., and Collins, A. G. E. (2025). Dual process impairments in reinforcement learning and working memory systems underlie learning deficits in physiological anxiety. *bioRxiv*, 2025–02. doi: 10.1101/2025.02.14.638024

Stachenfeld, K. L., Botvinick, M. M., and Gershman, S. J. (2017). The hippocampus as a predictive map. *Nat. Neurosci.* 20, 1643–1653. doi: 10.1038/nn.4650

Sutton, R. S. (1990). "Integrated architectures for learning, planning, and reacting based on approximating dynamic programming," in *Proceedings of the Seventh International Conference* (Austin, Texas), 216–224. doi: 10.1016/B978-1-55860-141-3.50030-4

Sutton, R. S. (1992). "Adapting bias by gradient descent: an incremental version of delta-bar-delta," in *AAAI* (Citeseer), 171–176.

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. London: MIT press.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychol. Rev.* 55:189. doi: 10.1037/h0061626

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems*, 30.

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

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., et al. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Wu, X., and Foster, D. J. (2014). Hippocampal replay captures the unique topological structure of a novel environment. *J. Neurosci.* 34, 6459–6469. doi: 10.1523/JNEUROSCI.3414-13.2014