Network computational mechanism underlying uncertainty-driven arbitration between model-based and model-free reinforcement learning

Siyu Wang (siyu.wang@nih.gov)
Laboratory of Neuropsychology, NIMH/NIH, 49 Convent Drive, Room 1B80
Bethesda, MD 20892 USA

Bruno Averbeck (averbeckbb@mail.nih.gov)
Laboratory of Neuropsychology, NIMH/NIH, 49 Convent Drive, Room 1B80
Bethesda, MD 20892 USA

Abstract:

In reinforcement learning, humans and animals rely on both a deliberative model-based system which builds internal models of the environment to facilitate learning and action planning, and a habitual model-free system which reinforces actions directly from rewards. How do animals arbitrate between the two systems? Previous studies based on behavioral modeling have provided evidence that the reliability of the predictions from these two systems determine how animals arbitrate between them. However, it is unknown how such computation is implemented by populations of neurons. In this work, we investigate the computational motif in networks that underlie the arbitration between a model-based and a model-free system in reinforcement learning. We trained recurrent neural network models that can flexibly switch between model-based and model-free strategies based on the task environment. By analyzing latent network activity during the arbitration process, we show how attractor population dynamics in networks underlie the model-based vs model-free arbitration. Our results suggested a general computational motif in networks on uncertainty-driven arbitration between abstract choices.

Keywords: reinforcement learning; model-based; model-free; population dynamics

Introduction

In reinforcement learning, humans and animals use both a model-based (MB) system which builds models of the external world to facilitate goal-directed behavior, and a model-free (MF) system which reinforces habitual behavior. The MB system utilizes internal predictions of the environment (e.g., estimate of state transitions) to plan actions, whereas the MF system simply reinforces actions that led to past rewards and avoids actions that led to punishments(Drummond & Niv, 2020).

Since animals use both MB and MF strategies in reinforcement learning, a natural question that arises is how the neural system arbitrates between MB vs MF strategies. Behavioral modeling work has suggested that reliability in predictions by MB and MF systems determine which system dominates behavior(Lee et al., 2014). However, it is unclear how such arbitration occurs in a neural system comprised of networks of neurons.

Neural computations in networks are studied through the lens of dynamical systems(Vyas et al., 2020). In value-based learning and decision-making, it has been shown that attractor dynamics in networks account for the decision processes in motor choices(Genkin et al., 2023; Wang et al., 2023; Wong & Wang, 2006). Although we understand how networks solve simple choices between motor actions, it remains unclear how networks resolve more abstract choices like arbitrating between MB vs MF strategies.

In this work, we investigate network computational mechanisms underlying the arbitration between MB vs MF system by studying recurrent neural network models that are trained to perform a modified version of the well-known two-step task(Daw et al., 2011). In the original two-step task and most existing variants of the task(Akam et al., 2015), a MB agent is always optimal, and a MF agent is suboptimal. However, to properly study the arbitration process, we need the same animal or network model to sometimes behave in a model-based way while other times behave in a model-free way. In the modified task, we trained the network to switch between MB and MF strategies based on the uncertainty in the environment. By analyzing latent network dynamics using dynamical systems methods, we revealed an attractor-based computational motif that underlies the arbitration between MB and MF learning strategies.

Methods

Behavioral task

In this task (Fig. 1A), agents choose between two possible actions A_1 and A_2 at starting state S0 to travel to either state S1 or S2 and collect rewards. The task is organized into blocks of 300 trials.

In each block, one action A_i will more likely transit the agent to S1 with probability $p(S1|Ai) = pT > 0.5$, whereas the other action will more likely transit the agent to S2 also with probability pT. There are two possible worlds: in world $W = W_1$, $p(S1|A_1) = p(S2|A_2) = pT > 0.5$, whereas in world $W = W_2$, $p(S1|A_1) = p(S2|A_2) = 1 - pT < 0.5$.

State S1 and S2 have complimentary probabilities of giving a reward (e.g., $p(r|S1) = 0.8, p(r|S2) = 0.2$). The reward probability reverses at a hazard rate of 0.02. Thus, agents must constantly monitor which state is more rewarding at any given time.

Crucially, the agent's observation of the state does not always match the true state. The state observation is accurate at probability $1 - pS$. Here pS reflects the stochasticity in the state observation. When $pS = 0$ and in world W1 (ipsilateral transition rule), i.e., $p(S1|A_1) > 0.5$, our modified task becomes the standard reduced version of the two-step task(Akam et al., 2015).

The parameters W, pT, and pS are randomly sampled at the beginning of each block and fixed within a block.

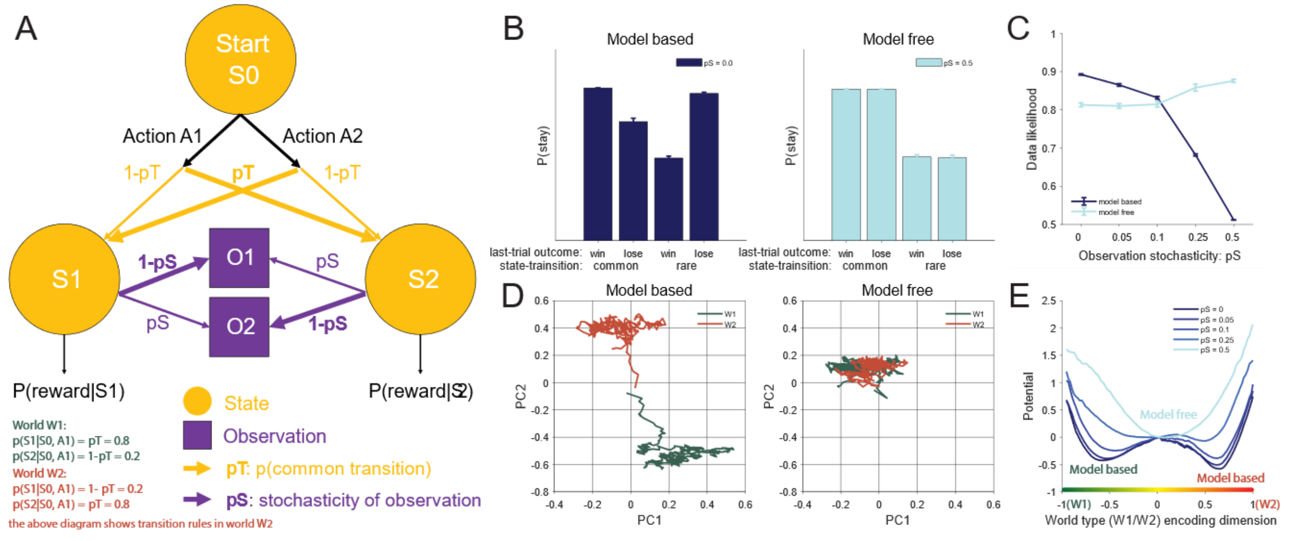

Figure 1. (A) Two-step task with stochastic state observations. (B) and (C) Trained networks show model-based behavior when pS is low, and model-free behavior when pS is high. (D) In model-based learning, network population activity enters different subregions of the state space based on world type ($W1/W2$). In model-free learning, population trajectories do not separate by world type. (E) Attractor dynamics in networks account for the arbitration between model-based and model-free reinforcement learning.

Network training

We trained fully connected gated recurrent networks (LSTM with 128 units) to perform the task. Network training was carried out using Advantage Actor-Critic(Wang et al., 2018). Network was first trained with $pS = 0$ until convergence, and then trained with $pS \in \{0, 0.25, 0.5\}$. Trained networks were tested with all possible combinations of $W \in \{W_1, W_2\}$, $pT \in \{0.6,0.7,0.8,0.9,1.0\}$, $pS \in \{0, 0.05, 0.1, 0.25, 0.5\}$.

Behavioral modeling

Cognitive MB and MF models were fitted to the network behavior. For a MB agent, values of the two actions are calculated based on state-transition and state-value: $V_a = p(S1|a)V(S1) + p(S2|a)V(S2)$, reward prediction errors $\delta = \mathrm{r} - \mathrm{V(S)}$ are used to update $V(S) = V(S) + \alpha\delta$.For a MF agent, values of the two actions are directly updated by reward prediction error: $V(a) = V(a) + \alpha\delta$.

Analysis of population dynamics in networks

We recorded the population activity from all 128 units in each trained network during the task. Using a SVM decoder, we first identified a 1-D subspace that best predicts which world W the agent is in. Our hypothesis is that when the network is certain about the world type $W = W1$ or $W2$, it behaves like a model-based agent, otherwise the network adopts a model-free strategy.

We performed principal component analysis to reduce the population activity dimensionality for visualization purposes (Fig. 1D). To analyze how population activity evolves along the encoding dimension for world type W, we adopted methods in Wang et al. (2023) and reconstructed energy landscapes in this W subspace. Briefly, we first computed the average time derivative of population activity X_t at binned locations $[x, x + \Delta x]$ and time t in the 1-D space $\frac{\mathbf{d}^t X_t}{\mathbf{dt}}|_{X_t = x} = \boldsymbol{E}_{\ X_t \in [x, x+\Delta x]} \frac{X_{t+\Delta t} - X_t}{\Delta t}$. Then we took the spatial integral over these time derivatives to get the energy potential $\Phi_{x,t}$ at location x and time t, $\Phi_{x,t} = -\int_{-\infty}^{x} \frac{\mathbf{d}^t X_t}{\mathbf{dt}}|_{X_t = x} \mathrm{d}^X X$.

Results

In our modified task, observations of the states are stochastic. By design, a MF agent would outperform a MB agent when stochasticity of observation pS is high, and a MB agent would be superior when pS is low. Our trained networks can indeed flexibly switch between MB and MF behavior for different pS (Fig. 1B, C)

When examining latent trajectories of networks in the principal component space, we observed that latent network activity goes to different subspaces depending on the world type W when the agent behaves MB and does not separate when the agent is MF (Fig. 1D).

A reconstruction of the energy landscape along the W encoding dimension reveals the attractor population dynamics underlying the arbitration between MB and MF systems. Positive and negative sides of the W subspace encodes how likely the environment is in world $W1$ or $W2$. The further away population activity is from 0, the more MB the agent becomes. When the stochasticity in observation pS is low, the world is very learnable, and we observe deep attractor basins which respectively attract the population activity into the $W1$ zone or the $W2$ zone, whereas when pS is high, population activity stays in the center and behaves in a MF fashion. Together, our results suggest an attractor-based computational motif that underlies the arbitration between MB and MF learning strategies.

Acknowledgements

We thank Jiachen Liu for his assistance in data analysis during the early phase of the project.

References

Akam, T., Costa, R., & Dayan, P. (2015). Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task. *PLoS Comput Biol*, *11*(12), e1004648. https://doi.org/10.1371/journal.pcbi.1004648

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204-1215. https://doi.org/10.1016/j.neuron.2011.02.027

Drummond, N., & Niv, Y. (2020). Model-based decision making and model-free learning. *Curr Biol*, *30*(15), R860-R865. https://doi.org/10.1016/j.cub.2020.06.051

Genkin, M., Shenoy, K. V., Chandrasekaran, C., & Engel, T. A. (2023). The dynamics and geometry of choice in premotor cortex. *bioRxiv*. https://doi.org/10.1101/2023.07.22.550183

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, *81*(3), 687-699. https://doi.org/10.1016/j.neuron.2013.11.028

Vyas, S., Golub, M. D., Sussillo, D., & Shenoy, K. V. (2020). Computation Through Neural Population Dynamics. *Annu Rev Neurosci*, *43*, 249-275. https://doi.org/10.1146/annurev-neuro-092619-094115

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nat Neurosci*, *21*(6), 860-868. https://doi.org/10.1038/s41593-018-0147-8

Wang, S., Falcone, R., Richmond, B., & Averbeck, B. B. (2023). Attractor dynamics reflect decision confidence in macaque prefrontal cortex. *Nat Neurosci*, *26*(11), 1970-1980. https://doi.org/10.1038/s41593-023-01445-x

Wong, K. F., & Wang, X. J. (2006). A recurrent network mechanism of time integration in perceptual decisions. *J Neurosci*, *26*(4), 1314-1328. https://doi.org/10.1523/JNEUROSCI.3733-05.2006