Multi-timescale reinforcement learning in the brain

Anonymous Author(s)
Affiliation
Address
email

Abstract

To thrive in complex environments, animals and artificial agents must learn to act adaptively to maximize fitness and rewards. Such adaptive behavior can be learned through reinforcement learning1, a class of algorithms that has been successful at training artificial agents and at characterizing the firing of dopamine neurons in the midbrain. In classical reinforcement learning, agents discount future rewards exponentially according to a single time scale, known as the discount factor. This strategy is at the odds with the empirical observation that humans and animals use non-exponential discounts in many situations. Here, we explore the presence of multiple timescales in biological reinforcement learning. We first show that reinforcement agents learning at a multitude of timescales possess distinct computational benefits. Next, we report that dopamine neurons in mice performing two behavioral tasks encode reward prediction error with a diversity of discount time constants. Our model explains the heterogeneity of temporal discounting in both cue-evoked transient responses and slower timescale fluctuations known as dopamine ramps. Crucially, the measured discount factor of individual neurons is correlated across the two tasks suggesting that it is a cell-specific property. Together, our results provide a new paradigm to understand functional heterogeneity in dopamine neurons, and open new avenues for the design of more efficient reinforcement learning algorithms.

1 Computational advantages of multi-timescale learning.

In traditional reinforcement learning (RL), value estimates $V(s)$ encode the sum of discounted future rewards expected from the current state s (Eq. 1, left). The exponential temporal discount is not an arbitrary choice but a consequence of using Temporal Difference (TD) learning: after transitioning from s to s' and receiving reward r, a TD-error $\delta_k = r + \gamma_k V^{s'} - V^s$ is used to update $V^s \leftarrow V^s + \alpha\delta$. This discount factor can be interpreted as capturing uncertainty about the evolution of future states [1–3]. Labelling values by their discount and taking the expectation inside the sum reveals a very useful property[4]: $V_\gamma(s)$ is the Z-transform of $E[r_\tau|s]$ (i.e. the discrete version of the Laplace transform). Since the Z-transform is invertible, multi-timescale values encode not only the expected sum of discounted rewards, as in traditional RL, but also the *expected reward at all future timesteps* (Eq. 1, right).

$$V_\gamma(s) = E\Big[\sum\nolimits_{\tau=0}^{\infty} \gamma^\tau r_\tau \big| s\Big] = \sum\nolimits_{\tau=0}^{\infty} \gamma^\tau E[r_\tau|s], \qquad \mathcal{Z}^{-1}\{V_\gamma(s)\}_{\gamma\in(0,1)} = \{E[r_\tau|s]\}_{\tau=0}^{\infty} \quad (1)$$

In single-timescale value learning, the value of a cue (at t = 0) predicting future rewards (Fig. 1a, first panel) is evaluated by discounting these rewards with a single exponential discounting function (Fig. 1a, second panel). The expected reward size and timing are encoded, but confounded, in the value of the cue (Fig. 1a, third panel). In multi-timescale value learning, the same reward delays are evaluated

Submitted to the Workshop on Unifying Representations in Neural Models (UniReps) at NeurIPS 2023. Do not distribute.

with multiple discounting functions (Fig. 1b, second panel). The relative value of a cue as a function of the discount depends on the reward delay (Fig. 1b, third panel). A simple linear decoder based on the Laplace transform can thus reconstruct both the expected timing and magnitude of rewards (Fig. 1b, fourth panel).

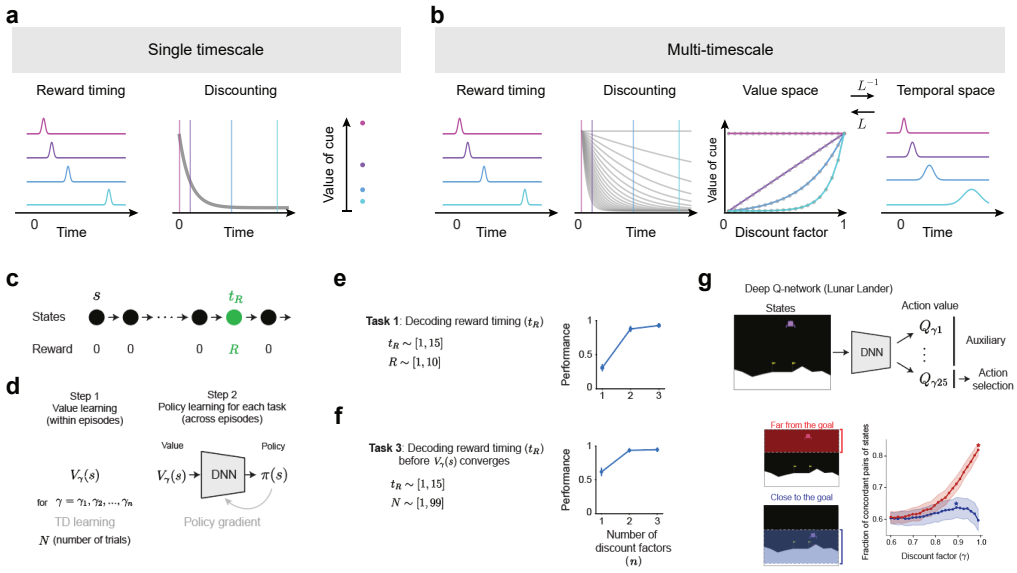

Figure 1: **Computational benefits of multi-timescale reinforcement learning.**

To illustrate the computational advantages of Laplace-transform multi-timescale agents , we consider several simple example tasks. The agent navigates through a linear track (a sequence of 15 states), where it encounters a reward of a certain magnitude (R) at a specific time point (t_R, Fig. 1c). The value of R and t_R changes across episodes and remains constant within episodes. Each episode is initiated by a cue presented at the initial state (s). Within each episode, the agent first learns the expected future rewards. Using the learned value $V_\gamma(s)$ associated with the cue, the agent then performs various tasks, using a deep neural network (DNN, using a policy gradient [PG] method trained across episodes (Fig. 1d). Performance is reported after 1,000 training episodes. Error bars are the standard deviations (s.d.) across 100 test episodes and 3 trained policy gradient (PG) networks.

We first asked whether an agent can correctly discern the magnitude (R) and the timing (t_R) of reward separately (Fig. 1e). We vary R and t_R across episodes. In each episode, the agent learns the values of states using 1, 2 or 3 discount factors. We then train the DNN across episodes to decode the timing of the reward (t_R) with the vector of values associated with the cue $V_\gamma(s)$ as its input. The pattern of values across discount factors (third panel in Fig. 1b) is invariant to reward magnitude and allows multi-timescale agents to decode the timing of reward.

We further hypothesized that, multi-timescale agents can leverage this advantage of extracting timing information even before value learning has fully converged (Fig. 1f). Consider an agent that has encountered a reward only a limited number of times (N). For single-timescale systems, a high value of the cue could be due to a short delay (t_R) or simply because the value estimate has undergone more positive updates from an initial value of 0. In contrast, the shape of values encoded across discount factors is invariant to the number of reward encounters (N), to the extent that all value estimates depart from similar baselines and share similar learning parameters. As a result, multi-timescale agents can decode the time of reward (t_R) even in situations where learning is incomplete (Fig. 1f).

An alternative way to leverage multi-timescale learning benefits is to employ them as auxiliary tasks (Fig. 1g, top). These networks only act according to the value of a single behavioral timescale, but concurrently learn about multiple other timescales as auxiliary tasks to enhance the representation in the hidden layers, which allows them to obtain superior performance in complex RL environments [3]. Multi-timescale systems could preferentially adjust between myopic and farsighted perspectives based on context and the accuracy (measured as fraction of concordant state pairs between the empirical value function and the discount specific Q-value) with which $Q_\gamma(s, a_{beh})$ captures the true empirical $V^{true}_{\gamma_{beh}}(s)$ across states depends on whether the agent is close to the goal (blue) or far from

the goal (orange) (Fig. 1g, bottom, Error bars are s.e.m across 20 trained networks, maximums are highlighted with stars.).

To summarize, in multi-timescale value systems the vectorized learning signal robustly contains temporal information independently of the information about reward magnitude. This property empowers agents to selectively focus on either myopic or far-sighted estimates depending on the current situation.

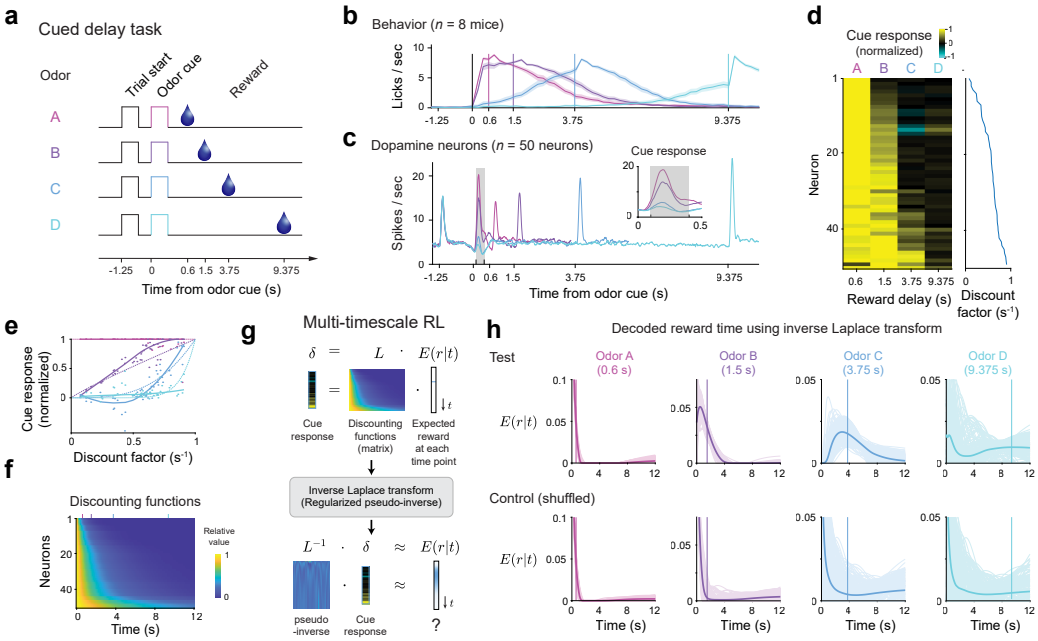

Figure 2: **Dopamine neurons exhibit a diversity of discount factors that enables decoding of reward delays.**

2 Dopamine-based multi-scale reinforcement learning

Considering these computational advantages, we wondered whether the TD-error conveyed by dopaminergic neurons [5] carried signatures of a multi-timescale computation. We recorded the activity of optogenetically identified dopaminergic (DA) neurons in mice performing two behavioural tasks. In a cued delayed reward task, in each trial, one of 4 possible cues predicted the reward delay (Fig. 2a). The mice exhibit anticipatory licking prior to reward delivery for all 4 reward delays indicating that they have learned task contingencies (Fig. 2b, mean across behavior for all recorded neurons, shaded error bar indicates 95% confidence interval using bootstrap). The DA neurons' (n=50) responses to odour cues decreased as a function of increasing delays (Fig. 2c, Inset shows the firing rate in the 0.5s following the cue predicting reward delay. The firing rate in the shaded grey box ($0.1s < t < 0.4s$) was used as the cue response in subsequent analysis). These neural discount functions were diverse and well-fit by an exponential discount function, allowing us to estimate a distinct discount factor γ for each DA neuron (Fig. 2d). The dopaminergic cue responses for each reward delay exhibited unique shapes as a function of discount factors, suggesting that reward timing information is embedded in the dopaminergic population responses (Fig. 2e, compare with Fig. 1b, third panel, Thick lines, smoothed fit, dotted lines, theory, dots, responses of individual neurons.). For each neuron we plot the relative value of future events given its inferred discount factor, resulting in the discount matrix (Fig. 2f), which we can invert using a parameter-free regularized inverse Z-transform (we compute the singular value decomposition (SVD) of the discount matrix L. Then, we use the SVD to compute a regularized pseudo-inverse L^{-1}. Finally, we normalize the resulting prediction into a probability distribution, Fig. 2g and see also ref [4]). The subjective expected timing of future reward $E(r|t)$ can be decoded from the population responses to the cue predicting reward delay. Decoding based on mean cue responses for test data (Fig. 2h, top row). The ability to decode the timing of expected future reward is not due to a general property of the discounting matrix and collapses if we randomize the identity of the cue responses (Fig. 2, bottom row). This suggests that

the dopaminergic signal also represents temporal evolution of the expected reward via a Z-transform and downstream areas can exploit the computational advantages highlighted above.

In a navigation task, mice experienced a 1-D linear track in virtual reality at the end of which they obtained a reward (Fig. 3a). Average activity of single neurons (n=90) exhibited an upward ramp as mice approached the reward, as has been found in bulk dopaminergic signal across several tasks [6, 7], but individual DA neurons showed diverse shapes of ramping activity, including upward, downward, and non-monotonic ramps (Fig. 3c). We hypothesized that ramping activity occurs due to mismatch between the increase in the value function that each DA neuron experiences and the discount factor that each DA neuron use to compute the TD-error (Fig. 3e-h). For agents experiencing an exponential value function (Fig. 3e-f) there is no TD error for an agent with the same discount factor as the parameter of the value function (red line). The TD error ramps upwards (downwards) if the discount factor is larger (smaller), dark red and light red lines respectively. In the case a cubic value function (Fig. 3g-h), Agents with large (small) discount factor experience a monotonic positive (negative) ramp in their TD error (dark red and light red lines respectively). Agents with intermediate discount factors experience non-monotonic ramps (red line). Unlike in the exponential value function case, no agent matches its discount to the value function at all the time steps (Fig. 3h). We found the qualitatively different ramping activities of single neurons can be quantitatively explained by this model (Fig. 3d) in which neurons have different discount factors (Fig. 3j, 0.42 ± 0.23, mean ± s.d.) and experience a common value function (Fig. 3k, grey line, individual bootstrap estimates. blue line, mean estimate, a similar formulation can be derived for a common reward timing expectation). Finally, a subset of neurons (n=43 neurons) was recorded across the two behavioural tasks. The inferred discount factors in the VR task and in the cued delay task were correlated (Fig. 3l, r = 0.45, P = 0.0013), and we were able to decode reward timing in the cued delayed reward task using discount factors inferred in the VR task. These results suggest that individual DA neurons have their own characteristic discount factor that dictates the parameter they use to compute the TD-error and regulate learning.

These results show that diversity in slow changes in activity across single neuron (known as dopamine ramps) in environments with gradual changes in value can be explained by a diversity of discount factors and is a signature of multi-timescale reinforcement learning. They also suggest that the discount factor (or its ranking) is a cell-specific property and strongly constrains the biological implementation of multi-timescale reinforcement learning in the brain.

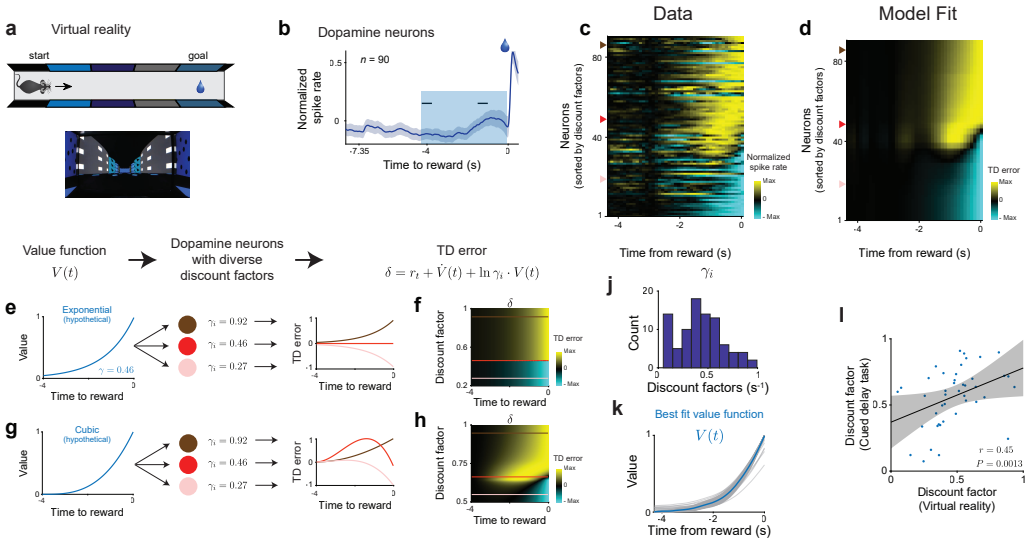

Figure 3: **The diversity of discount factors across dopamine neurons explains qualitatively different ramping activity.**

To conclude, our study investigates the computational advantages of multi-timescale reinforcement learning and establishes a new paradigm to understand the functional role of prediction error computation in dopaminergic neurons. It opens new avenues to develop mechanistic explanations for deficits in intertemporal choice in disease and inspire the design of new algorithms.

References

[1] P D Sozou. On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265, 1998.

[2] Zeb Kurth-Nelson and A David Redish. Temporal-difference reinforcement learning with distributed representations. *PLoS One*, 4(10):e7362, 2009.

[3] William Fedus, Carles Gelada, Yoshua Bengio, Marc G. Bellemare, and Hugo Larochelle. Hyperbolic Discounting and Learning over Multiple Horizons. *arXiv*, 2019.

[4] Pablo Tano, Peter Dayan, and Alexandre Pouget. A local temporal difference code for distributional reinforcement learning. *NeurIPS*, 33, 2020.

[5] Wolfram Schultz, Peter Dayan, and P Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

[6] HyungGoo R Kim, Athar N Malik, John G Mikhael, Pol Bech, Iku Tsutsui-Kimura, Fangmiao Sun, Yajun Zhang, Yulong Li, Mitsuko Watabe-Uchida, Samuel J Gershman, et al. A unified framework for dopamine signals across timescales. *Cell*, 183(6):1600–1616, 2020.

[7] Akash Guru, Changwoo Seo, Ryan J Post, Durga S Kullakanda, Julia A Schaffer, and Melissa R Warden. Ramping activity in midbrain dopamine neurons signifies the use of a cognitive map. *BioRxiv*, 2020.