Distributional reinforcement learning explains ensemble dopamine responses in habenula lesioned mice

In traditional reinforcement learning (RL), an agent predicts the average future rewards to guide actions. A novel algorithm called distributional RL (DRL), predicts the distribution of rewards and improves performance in artificial agents. A recent study (Dabney et al., 2020) provided initial evidence that the dopamine system may be working under DRL, accounting for the reported diversity in reward prediction errors (RPE). A previous study (Tian and Uchida, 2015) showed that habenula lesions impaired specific aspects of average dopamine responses while largely preserving RPEs. Our re-examination of this dataset, however, revealed a change in ensemble responses not explained by traditional RL. Here, we tested whether these changes conform to DRL predictions.

Mice were trained in a task in which odors predicted reward with different probabilities (10, 50 and 90%). In this task, traditional RL predicts that cue responses are proportional to reward probability. This was the case in controls but not in lesioned animals. We hypothesized that DRL provides additional constraints on the ensemble response pattern of cue-evoked responses. We categorized them into levels of optimism based on whether their amplitude in 50% reward trials was above or below the interpolated response from 10% and 90% reward trials. While we found similar proportion of optimistic and pessimistic responses in controls, the distribution was skewed toward optimistic responses after lesions. We tested whether this can be explained by dopamine neurons' asymmetries in reward responses to positive and negative prediction errors. As opposed to traditional RL, a DRL model incorporating reward response asymmetries reproduced the optimistic bias in cue responses. Accordingly, the decoded distribution associated to the 50%-reward cue based on an expectile code was biased to optimistic values after lesions. Together, these results demonstrate that a change in ensemble dopamine responses, not explained by traditional RL, could be readily explained by DRL.

Additional detail

Distributional RL (DRL). In RL, learning of value is driven by reward prediction errors (RPE, δ). The value prediction (V) is updated with a learning rule, $V \leftarrow V + \alpha \cdot \delta$, where α is the learning rate. With this update rule, V converges to a single quantity, the *expected value* of the reward distribution. DRL, on the other hand, learns the entire reward probability distribution. This can be achieved by an assortment of value predictors (V_i) with diverse learning rates for positive and negative RPEs ($\alpha_i^{+\text{¿},\alpha_i^{-\text{¿¿}}\text{¿}}$). Each V_i is characterized by a particular asymmetric scaling factor, τ_i equivalent to $\alpha_i^{+\text{¿}}/\text{¿¿¿¿}$. Value predictors trained using learning rates with $\tau_i > 0.5$ converge to quantities above the expected value (considered "optimistic") whereas value predictors trained using $\tau_i < 0.5$ converge to quantities below the mean (considered "pessimistic"). More precisely, each V_i converges to a quantity equivalent to the τ_ith-expectile of the distribution, which is an extension of the mean as a quantile is an extension of the median. A previous study[1] proposed that the learning rate parameters, $\alpha_i^{+\text{¿¿}}$ and $\alpha_i^{-\text{¿¿}}$, correspond to the slope of dopamine responses as a function of positive and negative RPEs. Furthermore, RPEs are computed based on the predicted value V_i ($\delta = R - V_i$). As a result, the amount of reward (R) at which dopamine reward response reverses its sign (reversal point Z_{τ_i}) should be directly related to V_i[1]. An extension of these models with temporal difference (TD) learning, conceptualizes responses to cues predictive of reward as the error computed over the discounted sum of future rewards. Thus, cue responses should be driven by V_i, and reflect the diversity of optimistic and pessimistic $V_i s$.

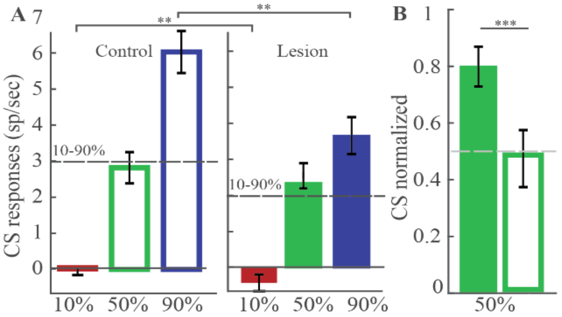

Fig. 1A. Raw CS dopamine responses for lesion and control groups. **B**. Normalized 50% CS responses to the 10-90% CS responses.

Data set and nomenclature. The analysis was made on optogenetically-identified dopamine neurons from the lesion and control groups (n = 44 and 45 neurons, respectively; [2]). We refer to responses to the cues predictive of X% probability of rewards as "X%-cue responses". The term 'CS' in figures refer to 'conditioned stimulus'.

Changes in dopamine cue responses after lesions are not explained by traditional RL. The original study[2] reported a reduction in positive raw responses to cues and rewards, and of negative responses to reward omissions (Fig.1A, P=0.005, P=0.007, two-sample KS-test). A closer examination revealed that cue responses were linearly increased by reward probability in control animals, with the 50%-cue responses closely matching the interpolated value between the 10 and 90%-cue responses (P=0.86, two-tailed Mann-Whitney test). In lesioned animals, however, the 50%-cue response was greater than the interpolated one ($P < 10^{-5}$, left-tailed Mann-Whitney test), exhibiting a non-linear relationship between cue response and reward probability. Accordingly, the normalized 50%-cue response was larger in lesioned animals than controls (Fig. 1B, P=0.0007, KS-test). Traditional RL does not predict the non-linear relationship in lesioned animals. Here we examined whether DRL explains the lesion effects.

The distribution of 50%-cue responses is shifted to optimistic values after lesions. To understand the cue responses in the framework of DRL, we first categorized neurons into optimistic, neutral and pessimistic by examining whether each neuron's 50%-cue response was significantly different from the 10-90% interpolated value (P<0.05, single-tailed Mann-Whitney tests). During the tests, we obtained the t-statistic of the differences. The proportion of pessimistic cells was lower in the lesion group (3/44) than controls (16/45) (P = 0.013, χ^2 test), with the latter presenting a uniform distribution of categories (Fig. 2A). In addition, there was a positive shift in the t-statistic distribution on lesioned animals compared to controls (Fig. 2B, P = 0.0059, KS-test). Computing this statistic on separate halves of data yielded consistent metrics (r = 0.93, r = 0.90, $P < 10^{-17}$).

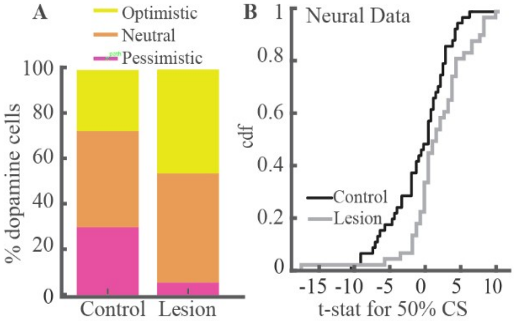

Fig.2A. Categorization of cue responses. **B**. Distribution of t-statistics for the deviation of the 50%-cue responses from the 10-90% interpolated response

Lesions increased the variance of the distribution of learning rate asymmetries. We derived the Z_{τ_i} and learning rate asymmetries (τ_i) from reward responses as previously proposed[1]. This revealed that the variance of the τ_i distribution was larger after lesions (bootstrapping test for difference of variance P < 0.001), but not the mean.

DRL TD learning models disclose an effect of the τ_i distribution variance in cue responses. Next, we examined how changes in the distribution of τ_i affects cue responses of dopamine neurons in a DRL model. Through DRL TD learning simulations, we systematically varied the variance (σ^2) from which artificial τ_i distributions are sampled, keeping all other parameters fixed. An increase in variance of the τ_i distribution resulted in shifts of cue responses toward optimistic levels (Fig. 3A), together with graded increases in the t-statistics as a function of σ^2 for units with a given value of τ_i (Fig. 3B, $P < 10^{-12}$ for effect of τ_i and σ^2 on t- statistics, P = 0.0002 for interaction between τ_i and σ^2, 2-way ANOVA). Indeed, DRL models trained using the τ_i obtained from data were able to generate the pattern of cue responses that qualitatively match the data in lesioned animals (Fig.3C, top. P=0.006, KS-test). Importantly, this effect was not observed with a model that used the data-derived τ_is but with an RL update rule that relied on the expected value (Fig.3C, bottom. P=0.56, KS-test).

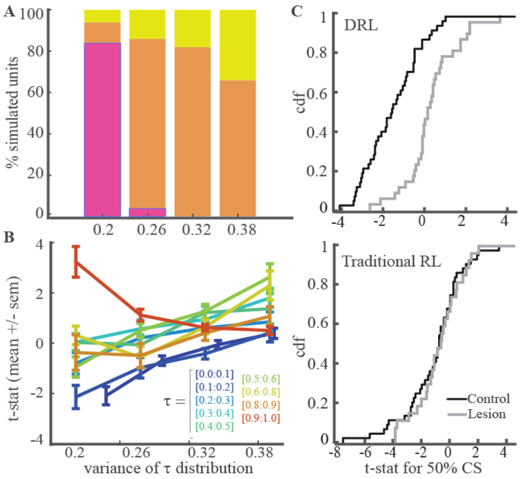

Fig. 3A Categorization of cue responses from simulated data when varying the variance of distribution. **B**.T-statistics of cue responses from simulations as a function of variance of distributions and single unit . **C**. Distribution of t-statistics from models' cue responses when trained with a DRL (top) or traditional RL rule (bottom)

Additional evidence for DRL. Both in control and lesioned groups we found ~~(1)~~ a positive correlation between Z_{τ_i} and τ_i as in the previous study[1] (r = 0.46, r = 0.53, $P < 10^{-4}$ for control and lesion across separate halves of data), and ~~(2) an ability~~ were able to decode the return distribution from reward responses, with a positive bias in the lesion group 50%-cue distribution ($P < 10^{-12}$, KS-test). Together, dopamine responses in lesioned animals are consistent with DRL, and the DRL framework further accounts for the ensemble response patterns to reward predictive cues.

1.Dabney, W. et al. Nature ***577****, 671–675 (2020).*
2.Tian, J. & Uchida, N. Neuron ***87****, 1304–1316 (2015).*