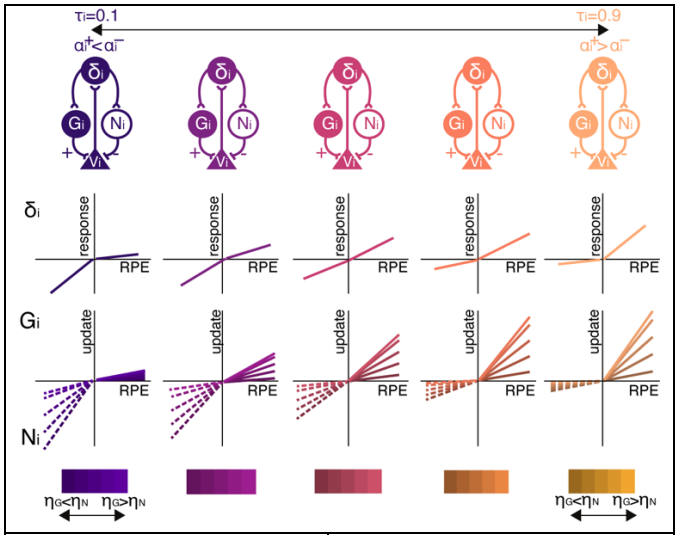
Linking tonic dopamine and biased value predictions in a biologically inspired reinforcement learning model

Some psychiatric disorders are characterized by excessively optimistic or pessimistic predictions of future events, and changes in dopamine levels. However, how such broad changes in dopamine could lead to biased value predictions is unknown. Here, we draw this link by examining the role of dopamine baseline levels in value learning.

A way in which value learning could be implemented in the brain, is through plasticity processes driven by dopamine reward prediction errors (RPEs) acting upon D1 and D2 receptors in medium spiny neurons (MSNs). At normal dopamine levels, D1 and D2 receptors are mostly unoccupied and occupied by dopamine, making them sensitive to increases and decreases of dopamine, respectively. Accordingly, studies have reported that potentiation in MSNs expressing D1 or D2 receptors is triggered by phasic increases or decreases of dopamine (Yagishita, 2018; Iino, 2020). Moreover, given the sigmoidal dose-occupancy relationship of these receptors, shifts in dopamine baseline should change their sensitivity to dopamine transients (i.e., take the baseline to a "steep" or "shallow" region of the dose-occupancy curve). Here, we show that a reinforcement learning (RL) model incorporating these plasticity rules, develops positive or negative biases in predictions of probabilistic rewards if the baseline dopamine is increased or decreased, respectively. We validate this model using data from a previous study (Tian, 2015). This study showed that lesions of habenula resulted in positive biases both in reward-seeking behavior (anticipatory licking) and in responses of dopamine neurons to cues predictive of probabilistic rewards. In our model, increases in baseline firing of dopamine neurons alone, as in the data, leads to these optimistic biases. Taken together, our biologically inspired RL model highlights a causal impact of baseline dopamine on biasing future value predictions, which may underlie some abnormalities observed in psychiatric patients and could be used to regulate risk sensitive behavior.

Additional Details. Most classic RL models rely on a single population of 'value neurons' whose estimates are positively and negatively updated by positive and negative RPEs, respectively. A direct brain implementation of this, however, requires some reformulations, as this would imply that plasticity at the cortico-striatal synapses in the MSNs would be of potentiation (LTP) and depression (LTD) for positive and negative dopamine (DA) responses, respectively. Contrary to this, experimental evidence points to signatures of LTP with DA bursts in D1-receptor-expressing MSNs (D1-MSNs) and with pauses in D2-receptor-expressing MSNs (D2-MSNs). This dichotomous effect of dopamine in plasticity has been linked to the downstream effects of striatal-projection pathways, arguing that each pathway aggregates evidence 'in favor' or 'against' a certain state, and the decision to 'approach' or 'avoid' is function of their difference. We build upon this line of reasoning.

Distributional reinforcement learning and biases in value learning in the brain. Another aspect of classic RL models, is that learning of value is driven by RPEs (δ) via the update the rule, $V \leftarrow V + \alpha \cdot \delta$ (where α is the learning rate), driving the estimate to converge to the expected value of the reward distribution. Distributional RL (DRL), on the other hand, permits an agent to learn the distribution of rewards by allowing each value neuron (V_i) to perform the updates asymmetrically for positive vs negative RPEs (α^+, α^-). Each V_i possesses a characteristic asymmetric scaling factor defined by the ratio $\tau_i = \alpha_i^+/(\alpha_i^+ + \alpha_i^-)$, which makes it converge to an estimate equal to an expectile of the distribution. In addition, there is a distribution of τ_i across the population, so that learning of the distribution is possible. At the implementation level, the asymmetric scalings (α^+, α^-) of value updates have been proposed to be reflected in the firing rates of DA evoked responses to RPEs (Dabney et al 2020).



$\delta_i = r - V_i$ ..... **Eq. 1**

$\hat{\delta}_i = \alpha_i^+ \cdot \delta_i$ if $\delta_i > 0$ .... **Eq. 2**

$\hat{\delta}_i = \alpha_i^- \cdot \delta_i$ if $\delta_i < 0$

$N_i \leftarrow N_i + \eta_G \cdot \hat{\delta}_i$ if $\delta_i > 0$... **Eq. 3**

$G_i \leftarrow G_i + \eta_G \cdot \hat{\delta}_i$ if $\delta_i < 0$

$V_i = G_i - N_i$ .....**Eq. 4**

Figure 1. Schematic of the DRL model with D1 and D2 populations. The dopamine population (δ_i) presents a distribution of asymmetric scaling factors (τ_i) determined by the gain of their firing rates responses for positive and negative RPEs (α^+, α^-) (middle row). These scaled positive or negative RPEs (**Eq. 1, 2**) are used to update the estimates of G_i or N_i, respectively (bottom row, **Eq. 3**). There is a global rescaling of the updates at this level (η_G, η_N) given by the receptor activation level of either population. The final value prediction is the difference of the estimates of G_i and N_i (**Eq. 4**)

Building on biological plasticity rules, we implemented a DRL model that includes separate D1- and D2-MSNs populations. Here, the sign of the RPE (phasic increases or pauses in DA release) will determine which population (G_i, N_i) updates its prediction. The learning rates (η_G, η_N) are given by the receptor activation levels for a given RPE magnitude, and it is assumed to be equal for all neurons within a population. This learning rate is independent of the one given by the individual DA neuron's RPE scaling that determines which expectile of the distribution each MSN will learn. The final value predictions (V_i) are the result of the pair-wise difference between the D2 and D1-MSNs estimates for a given expectile (Fig.1). This model allows DRL to learn a value distribution, while agreeing with the effects of DA in D1 and D2 MSNs plasticity. Having developed a model closer to biology, our interest was to understand how biased value predictions may arise mechanistically.

For biases to arise in DRL while still allowing to learn the form of the reward distribution, there needs to be a rescaling of the asymmetric learning rates such that they become unbalanced at a population level. In the presence of probabilistic outcomes this would make the learnt expected value (i.e., the 50th expectile) to be biased to more positive or negative values than the true one, and this would generalize to all the expectiles. In DRL, this may happen through 1) the global

modulation of asymmetric scalings (α_i^+, α_i^-) of DA evoked responses to RPEs, or 2) through the modulation of the receptor activation level (η_G, η_N) of each population for a given RPE.

The mechanistic underpinning for how DA firing rates could be independently modulated for responses to positive or negative RPEs are unclear, as it is likely that the neural factors underlying RPE computations are the same for both domains. Thus, their modulation would potentially affect both response regimes equally (i.e. change α_i^+, α_i^- in the same direction). In addition, depressive-like states are associated with suppressed spontaneous DA activity levels (REF), with no reported specific effects on DA RPEs to date. Given this, we propose a mechanism for asymmetric gains based on the D1 and D2 receptors dose-occupancy relationship. Specifically, given the sigmoidal dose-occupancy curve, a shift in baseline DA levels, would cause a global re-scaling of the receptor activation level for a given amplitude of phasic release or pause of DA, and this re-scaling would differ for each receptor class. For example, a positive shift in baseline DA would increase and decrease the sensitivity of D1- and D2-MSNs to phasic increases and decreases, respectively (Fig.2A-B). Assuming that plasticity is dependent on receptor activation level, we reach a regime of asymmetric learning rates ($\eta_G > \eta_N$) that leads to optimistic biases (Fig 2C-D). Thus, shifts in basal DA levels alone, would give rise to a global asymmetry in value updates.

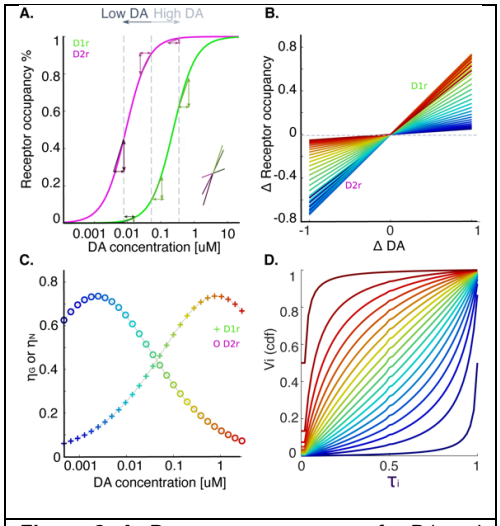


Figure 2. A. Dose-occupancy curve for D1 and D2 receptors. Receptor occupancy changes for transients in DA will depend on baseline DA concentration. **B.** Change in receptor occupancy as a function of transient DA changes for different levels of baseline DA from low (blue) to high (red). **C.** Slope of the previous curve for positive (D1r) and negative (D2r) transient DA changes as a function of baseline concentration. This slope determines the gains (η_G, η_N) for value updates in our model. **D.** Value distributions V_i (i.e. expectiles) learnt under the different levels of baseline DA for a Bernoulli distribution with p=0.5.

Optimistic biases arise from the reported changes in DA after habenula lesions. We tested our model with data from a previous study (Tian et al., 2015). Here, mice were trained in a task consisting of cues associated with reward probabilities (10%, 50%, 90%, Fig 3A), and underwent lesions of the lateral habenula (LHb). This experiment is of relevance as LHb hyperactivity is reported in depressive-like (i.e., pessimistic) states (REF). Mice that underwent lesions presented an optimistic bias in its reward-seeking behavior, as quantified from their anticipatory licking to the 50% cue with respect to the 10% and 90% cue (Fig 3B). Similar changes in cue-evoked responses of DA neurons were observed (Fig 3C). Importantly, signatures of canonical RPE-like computations and of a distributional code were present after lesions. Neither a classic RL nor a standard DRL model could entirely explain the optimistic biases. The distribution of asymmetric scaling factors expanded the range from 0 to 1 and did not present a bias with respect to controls (Fig 3D), indicating a lack of a global rescaling of α_i^+ and α_i^-. A key observation was that baseline firing rates of DA neurons after lesions were increased (Fig. 3F). Considering our model, we can argue that this would lead to increases in baseline striatal DA levels and, thus, to a global asymmetry in value updates ($\eta_G > \eta_N$). This results in a positive bias in the learnt expected value and value distribution, leading to the observed optimistic biases and changes in cue responses (Fig 3G-H). This bias did not originate from the distribution of scaling factors, as this was kept fixed in the simulations.

Value learning biases can arise from the reported changes in DA activity in mood disorders. If the broad suppression of spontaneous DA activity observed in depressive-like states leads to decreases in striatal DA, as suggested experimentally [35,REF], this would lead to a regime in which value updates are asymmetric favoring negative outcomes ($\eta_G < \eta_N$). In our model, an agent learning with these conditions would develop symptoms of depression: enhanced sensitivity to losses with respect to gains, risk aversive behavior, and more broadly, pessimistic outcome expectations.

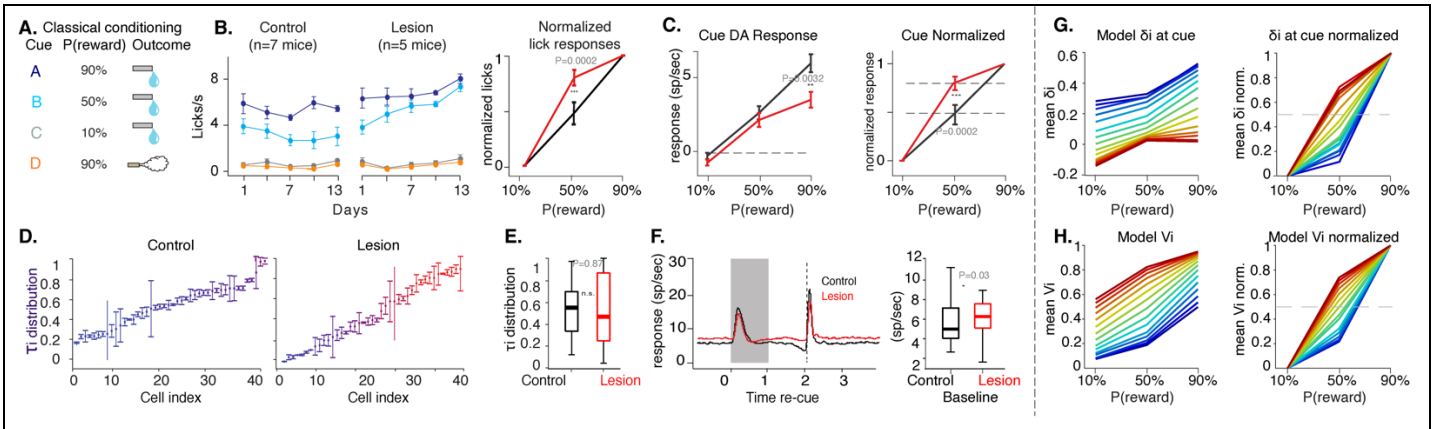


Figure 3. A. Schematic of the probabilistic task of Tian et al., 2015. **B.** Post-lesion anticipatory lick responses in mice that underwent Lhb or sham lesions. **C.** Normalized lick responses to cues predictive different probabilities of reward. **D.** Evoked DA cue responses and normalized cue responses, showing a decrease in magnitude of raw responses in lesion group (t-test; p=0.0032) that is followed by an increase in the normalized cue response to the 50% reward probability predictive cue. **D.** Distribution of asymmetric scaling factors (τ_i) derived from DA evoked outcome responses. **E.** Mean of τ_i distributions for each group (t-test; p=0.87). **F** Post-stimulus time histogram for an example neuron from control and lesion group (left), distribution of baseline firing rates (right) (t-test; p=0.03). **G.** Model simulations results for different levels of baseline DA (colors as in Fig.2). δ_i responses at cue decrease in magnitude with increases in DA baseline, as in the data after lesions (left). Normalized cue responses to the 50% cue increases (right). **H.** Increases in DA baseline leads to optimistic biases reflected in the mean and normalized mean across the V_i distribution.