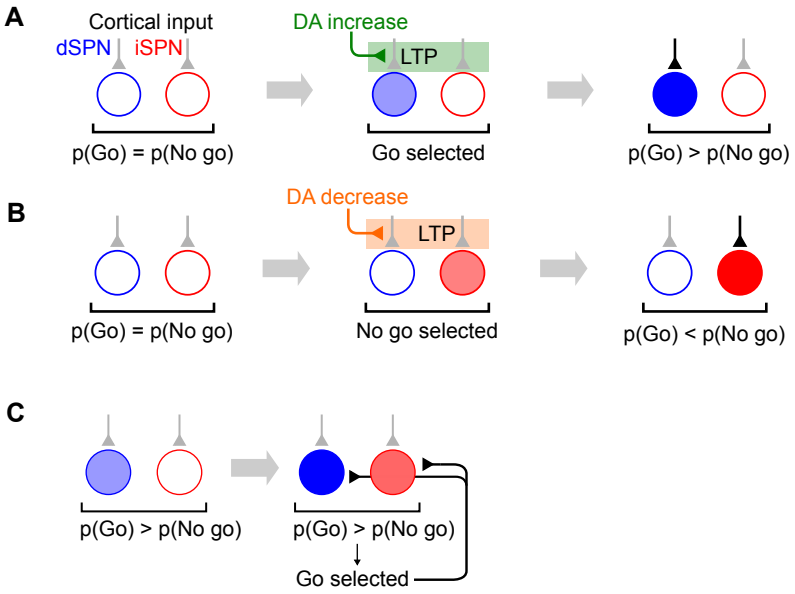
Supplemental information

Model of go/no-go task



Supplemental Fig. 1: Go/no-go task. **A.** Example in which dSPN plasticity produces correct learning behavior in a go/no-go task. Left: cortical inputs to the dSPN and iSPN are equal prior to learning. Shading of corticostriatal connections indicates synaptic weight, and shading of blue and red circles denotes dSPN/iSPN activity. Middle: the "go" response is selected, corresponding to elevated dSPN activity. In this example, the "go" response is rewarded, leading to elevated DA activity and thus potentiation of the dSPN input synapse. Right: in a subsequent trial, cortical input to the dSPN is stronger, increasing the likelihood of selecting the "go" response. **B.** Example in which iSPN plasticity produces incorrect learning behavior in a go/no-go task. Left: same as panel B. Middle: the "no go" response is selected, corresponding to elevated iSPN activity. In this example, the "no-go" response is punished, leading to decreased DA activity and thus potentiation of the iSPN input synapse. Right: in a subsequent trial, cortical input to the iSPN is stronger, decreasing the likelihood of selecting the "go" response. **C.** Illustration of the efference model in a go/no-go task. Left: feedforward SPN activity driven by cortical inputs. Right: once the "go" response is selected, the dSPN and iSPN are both excited by efferent input, which is combined with their original input. As a result, both the dSPN and iSPN are more active than prior to action selection, but the dSPN is still more active than the iSPN.

Relationship between sum mode activity and future difference mode activity

In the main text we provided an argument for why sum mode activity drives changes to future difference mode activity, assuming a linear $f^{\mathrm{d/iSPN}}(\delta)$ and linear neural activation functions. Here we generalize this argument to more general learning rules and activation functions ϕ, assuming only that $f^{\mathrm{dSPN}}(\delta)$ is monotonically increasing, $f^{\mathrm{iSPN}}(\delta)$ is monotonically increasing, and $\phi(\cdot)$ is monotonically increasing. We have that $y^{\mathrm{d/iSPN}} = \phi(\mathbf{w}^{\mathrm{d/iSPN}} \cdot \mathbf{x})$, and $\delta\mathbf{w}^{\mathrm{d/iSPN}} = (f^{\mathrm{d/iSPN}}(\delta) \cdot y^{\mathrm{d/iSPN}})\mathbf{x}$. Thus, in the limit of small small weight updates, we can write:

$$\begin{aligned}
\Delta(y^{\text{dSPN}} - y^{\text{iSPN}}) &= \Delta\phi(\mathbf{w}^{\text{dSPN}} \cdot \mathbf{x}) - \Delta\phi(\mathbf{w}^{\text{iSPN}} \cdot \mathbf{x}) \\
&\approx \phi'(\mathbf{w}^{\text{dSPN}} \cdot \mathbf{x})(\Delta\mathbf{w}^{\text{dSPN}} \cdot \mathbf{x}) - \phi'(\mathbf{w}^{\text{iSPN}} \cdot \mathbf{x})(\Delta\mathbf{w}^{\text{iSPN}} \cdot \mathbf{x}) \\
&\propto \phi'(\mathbf{w}^{\text{dSPN}} \cdot \mathbf{x})(f^{\text{dSPN}}(\delta) \cdot y^{\text{dSPN}}\mathbf{x} \cdot \mathbf{x}) - \phi'(\mathbf{w}^{\text{iSPN}} \cdot \mathbf{x})(f^{\text{iSPN}}(\delta) \cdot y^{\text{iSPN}}\mathbf{x} \cdot \mathbf{x}) \\
&= \|x\|^2 \left(\phi'(\mathbf{w}^{\text{dSPN}} \cdot \mathbf{x})(f^{\text{dSPN}}(\delta) \cdot y^{\text{dSPN}}) - \phi'(\mathbf{w}^{\text{iSPN}} \cdot \mathbf{x})(f^{\text{iSPN}}(\delta) \cdot y^{\text{iSPN}})\right) \\
&\propto c^{\text{dSPN}} f^{\text{dSPN}}(\delta) y^{\text{dSPN}} + (-c^{\text{iSPN}} f^{\text{iSPN}}(\delta) y^{\text{iSPN}}). \qquad (24)
\end{aligned}$$

where c^{dSPN} and c^{iSPN} are nonnegative because ϕ' is always nonnegative by assumption. Snce by assumption $f^{\text{d/iSPN}}$ are increasing/decreasing, respectively, the first term of the above sum has nonnegative correlation with δy^{dSPN} and the second term has nonnegative correlation with δy^{iSPN}. Thus, changes $\Delta(y^{\text{dSPN}} - y^{\text{iSPN}})$ to difference mode activity are always nonnegatively correlated with sum mode activity. If we assume that efferent excitation is always sufficiently strong that $c^{\text{dSPN}} = \phi'(\mathbf{w}^{\text{dSPN}} \cdot \mathbf{x})$ and $c^{\text{iSPN}} = \phi'(\mathbf{w}^{\text{iSPN}} \cdot \mathbf{x})$ are positive, and that there are no values of δ for which $f^{\text{d/iSPN}}(\delta)$ both have zero derivative, we can further guarantee that changes to difference mode activity will always be *positively* correlated with sum mode activity.

Generalizing the model to a distributed code for actions

In our model simulations in the main text we assumed for convenience that there is a single dSPN and iSPN that promote and suppress each available action, respectively. It is more realistic to model the code for action as distributed among many SPNs. Our model generalizes easily to this case; all that is necessary is for the efferent activity following action selection to excite the vectors (for both dSPNs and iSPNs) in population activity space corresponding to that action. To demonstrate this, we conducted a simulation with $N = 1000$ dSPNs and iSPNs each, $S = 10$ input cues (one-hot input vectors), and $A = 10$ actions, with one correct action for each input state. Feedforward SPN activity is given by

$$y_i^{\text{dSPN}} = \phi\left(\sum_{j=1}^{M} w_{ij}^{\text{dSPN}} x_j\right) \qquad (25)$$

$$y_i^{\text{iSPN}} = \phi\left(\sum_{j=1}^{M} w_{ij}^{\text{iSPN}} x_j\right) \qquad (26)$$

The log-likelihood of an action a being performed is proportional to

$$\ell_a = \sum_{i=1}^{N} \zeta_{ai}^{\text{dSPN}} y_i^{\text{dSPN}} - \zeta_{ai}^{\text{iSPN}} y_i^{\text{iSPN}} \qquad (27)$$

where ζ_{ai}^{dSPN} and ζ_{ai}^{iSPN} are randomly sampled uniformly in the interval [0, 1] and then normalized so that each vector $\zeta_{\mathbf{a}}^{\text{dSPN}}$ and $\zeta_{\mathbf{a}}^{\text{iSPN}}$ has norm 1. Thus, the contribution of each dSPN/iSPN to the promotion/suppression of each action is randomly distributed.

In the efference model, following selection of an action a^*, activity of the SPNs associated with action $a*$ is updated as follows, so that efference activity excites the modes $\zeta_{\mathbf{a}^*}^{\text{dSPN}}$ and $\zeta_{\mathbf{a}^*}^{\text{iSPN}}$ associated with the selected action:

$$y_i^{\text{dSPN}} \leftarrow \phi\left(c_{\text{efference}} \cdot \zeta_{a^*i}^{\text{dSPN}} + \sum_{j=1}^{M} w_{ij}^{\text{dSPN}} x_j\right) \tag{28}$$

$$y_i^{\text{iSPN}} \leftarrow \phi\left(c_{\text{efference}} \cdot \zeta_{a^*i}^{\text{iSPN}} + \sum_{j=1}^{M} w_{ij}^{\text{iSPN}} x_j\right) \tag{29}$$

$$\tag{30}$$

We also experiment with a generalization of the canonical action selection model to this distributed action tuning architecture, in which following action selection, SPN activity is set to

$$y_i^{\text{dSPN}} \leftarrow \zeta_{a^*i}^{\text{dSPN}} \tag{31}$$

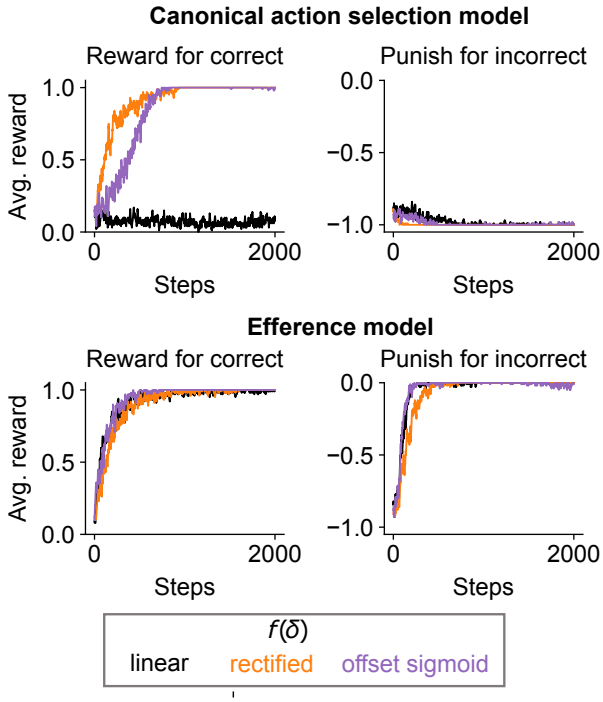
$$y_i^{\text{iSPN}} \leftarrow \left(\max_{i'} \zeta_{a^*i'}^{\text{iSPN}}\right) - \zeta_{a^*i}^{\text{iSPN}} \tag{32}$$

$$\tag{33}$$

In this model, dSPNs are excited in proportion to their contribution to the currently selected action and iSPNs are suppressed in proportion to their degree of inhibition of the currently selected action.

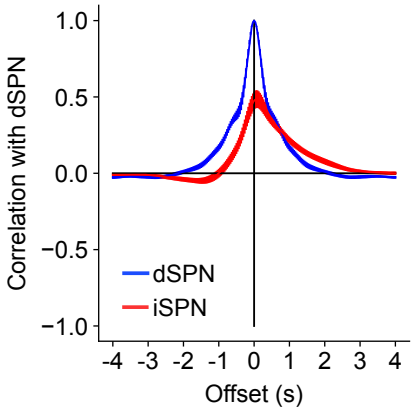
The plasticity rules used are the same as in the main text.

We find that the results of the main text – that the canonical action selection model fails to learn from negative rewards, while the efference model successully learns from both reward protocols – is replicated (Supp. Fig. 2).



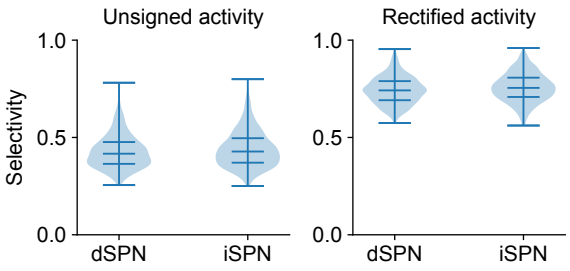
Supplemental Fig. 2: Performance of striatal RL models with a distributed code for actions on a task with 10 cortical input states, 10 available actions, and one correct action for each input state.

Photometry analysis with reversed indicators



Supplemental Fig. 3: Same as Fig. 5C, but performing the analysis on subjects with reversed assignment of indicators to SPN types.

Comparison of selectivity of dSPNs and iSPNs



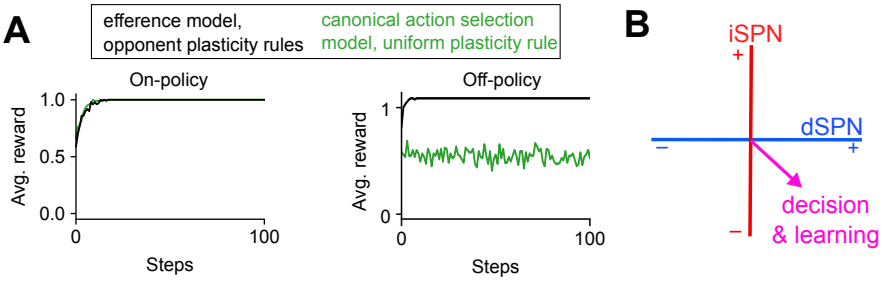
Supplemental Fig. 4: Comparison of dSPN and iSPN tuning selectivity. Violin plots indicate the distribution of selectivity values across all neurons computed using Eq. 34, using either unsigned (left) or rectified (right) z-scored activity as the raw measure of a neuron's tuning to a behavioral syllable. Horizontal lines indicate the $0, 25, 50, 75, 100$ percentile values of the distribution.

To test whether dSPNs or iSPNs exhibit greater or less specificity in their tuning to behaviors, we computed the selectivity of each neuron in the imaging data of Fig. 6. For each neuron, we computed its average z-scored activity a_i in response to each of the behavioral syllables $i \in \{1, ..., A\}$ in the dataset. Common measures of selectivity require a nonnegative measurement of a neuron's tuning to a given condition. Thus, we conducted the analysis in two ways, using either the unsigned activity $|a_i|$ or the rectified activity $\max(a_i, 0)$ as the measure of the neuron's tuning t_i to syllable i. The selectivity was then computed using the following expression introduced in prior work (Treves and Rolls, 1991; Willmore and Tolhurst, 2001):

$$\frac{\left(\frac{1}{A}\sum_i t_i\right)^2}{\frac{1}{A}\sum_i t_i^2} \tag{34}$$

This value ranges from 0 to 1, and higher value indicates that fluctuations in a neuron's activity are driven primaril by one or a few behavioral syllables. The results are shown in Supp. Fig. 4. The selectivity values are fairly modest (consistent with a distributed code for actions) and comparable between dSPNs andn iSPNs.

Alternative model with shared plasticity rule among all SPNs



Supplemental Fig. 5: Comparison to counterfactual model in which iSPNs use the same plasticity rule as dSPNs. A. Left: performance of simulated striatal RL system using efference model with the opponent dSPN/iSPN plasticity rules used elsewhere in the paper (black, same as Fig. 3E), and a system using the canonical action selection model and identical dSPN and iSPN plasticity rules (green). Right: same as left panel, but in an off-policy setting in which another pathway controls behavior during and always chooses the correct action, and the performance of the striatal RL system is evaluated over time. Here the Q-learning model of dopamine activity is used. B. In the counterfactual model in which iSPNs use the same plasticity rule as dSPNs, activity in the difference mode (dSPN - iSPN) influences (via plasticity) changes in future difference mode activity that affect decision-making.

The issues identified in Fig. 2 with the canonical action selection model are a consequence of the iSPN plasticity rule. From a normative perspective is interesting to consider why the empirically observed iSPN plasticity rule might be advantageous, compared to an alternative model in which iSPNs share the same plasticity rule as dSPNs. For instance, this alternative model can solve the two-alternative forced choice task of Fig. 2 with both positive and negative reward protocols (Supp. Fig. 5A, left). However, the limitations of this alternative model are revealed in the off-policy learning setting, where the Q-learning algorithm is required. In this case, SPN activity must encode Q-values associated with each action, but in the canonical action selection model, these values are disrupted by the updates to SPN activity following action selection. This is because the activity updates in the canonical action selection model modify difference mode activity, which (when dSPN and iSPN plasticity rules are the same) is needed for learning (Supp. Fig. 5B). As a result, the predicted Q-values are inaccurate, and the model has difficulty learning the true value of each action. We demonstrate this in the two-alternative forced task in an off-policy learning protocol where an oracle chooses the correct action on each trial, and the striatal pathway's ability to solve the task independently is evaluated. The efference activity model has no issue due to the orthogonality of the efferent activity and difference modes as described above, but the canonical action selection model fails to solve the task (Supp. Fig. 5A, right).

We note that non-orthogonality of the activity mode used for learning and behavior could cause other problems besides impairing the system's ability to implement off-policy learning algorithms; for instance, even in an on-policy setting it could interfere with sequential action selection at rapid timescales.

Models used for dopamine analysis

We experimented with models that predict transition probabilities $P(s_{t-1}, s_t)$ based on average dopamine activity $D(s_{t-1}, s_t)$ associated with each transition.

Q-learning model: In the Q-learning model, the mouse maintains an internal estimate of the value $Q(s_{t-1}, s_t)$ of each transition between syllables. In the absence of explicit rewards, the dopamine activity associated with a syllable transition is predicted to be: $D(s_{t-1}, s_t) = \max_{s'} Q(s_t, s') - Q(s_{t-1}, s_t)$. We inferred a set of Q-values by initializing a Q-table with all zero values and running gradient descent on the Q-table to minimize the mean squared error between the predicted and empirical values of $D(s_{t-1}, s_t)$. These inferred Q-values were used to predict behavioral transition probabilities according to: $\hat{P}(s_{t-1}, s_t) = \frac{e^{\beta(s_{t-1})Q(s_{t-1}, s_t)}}{\sum_{s'} e^{\beta(s_{t-1})Q(s_{t-1}, s')}}$. We did not fit the value of $\beta(s_{t-1})$ but rather chose it to be the reciprocal of the standard deviation of $Q(s_{t-1}, s')$ across all s', to ensure a reasonable dynamic range in predicted transition probabilities.

$V(s)$ TD learning model: In this model, the mouse maintains an internal estimate of the value $V(s)$ of each syllable, and the predicted dopamine activity at each transition is $D(s_{t-1}, s_t) = V(s_t) - V(s_{t-1})$. We fit the vector of values $V(s)$ to minimize the mean squared error of predicted and empirical $D(s_{t-1}, s_t)$. The predicted transition probabilities in this modl (which are independent of the previous syllable s_{t-1}) are: $\hat{P}(s_{t-1}, s_t) = \frac{e^{\beta V(s_t)}}{\sum_{s'} e^{\beta V(s')}}$ with β chosen to normalize the $V(s')$ to have standard deviation 1, as in the previous models.

Action value model: In this model, we assume that dopamine activity simply reflects the probability of each transition rather than encoding a prediction error; that is, we assume $P(s_{t-1}, s_t) = \frac{D(s_{t-1}, s_t)}{\sum_s D(s_{t-1}, s)}$.

State value model: In this model, we assume that dopamine activity simply reflects the probability of each behavioral syllable being chosen and is independent of the previous syllable. That is, we compute the average dopamine activity $D(s)$ associated with each syllable s, and predict $P(s_{t-1}, s_t) = \frac{D(s_t)}{\sum_s D(s)}$.