**Learning the action inventory of a complex skill via an intrinsic reward**

Toutounji, H, Zai, A, Tchernichovski, O, *Hahnloser RHR and *Lipkind D

**Summary**:

Reward-prediction errors as reflected by dopaminergic neuron responses are thought to underlie the acquisition of multi-action skills such as birdsong or speech, where sounding similar to one's learning target is intrinsically rewarding. It is not known how the predicted reward itself is computed. Based on past behavioral studies showing context-independence and greediness of birdsong syllable inventory learning, we hypothesize a similarity-based intrinsic reward formed by a hierarchical competition between partial rewards attributed to individual syllables and calls. Assuming that developmental birdsong learning is driven by minimization of reward prediction error, we infer a zero partial reward for vocalizations that are at a large acoustic distance from a target. We train young zebra finches to learn a syllable inventory and find that this reinforcement learning model correctly predicts the exclusion of a syllable from a song and its replacement by a call. We present implications for dopaminergic responses in the avian song system during learning.

**Introduction**

Animal behavior provides a unique opportunity for understanding evolutionary solutions for complex learning problems. One prime example is learning the inventory of the component actions of a combinatorial behavior, a problem that humans and other animal species solve – for example, when learning the moves in a sports game, a dance, or a courtship display. For vocal combinatorial behaviors such as speech and birdsong, the learning of an action inventory — speech vocal units or song syllables — is a particularly important step.

Converging evidence from both humans and songbirds point to the involvement of dopamine signaling in the acquisition of vocal skills[1–5], providing support for an underlying reinforcement learning (RL) mechanism[6–8]. Dopaminergic neurons in animals, including humans, signal reward prediction error (RPE) by increasing or decreasing their firing rate when an appetitive outcome is respectively better or worse than expected[9–12]. Similarly, in adult songbirds, dopaminergic projections to the basal ganglia signal whether aversive singing outcomes imposed by distorted auditory feedback are better or worse than expected[5]. Such error signal coding is at the heart of many RL algorithms and is likely to underlie developmental birdsong learning[4,13,14].

The goal of reinforcement learning is to maximize rewards, but internally motivated processes such as speech or birdsong learning readily occur in the absence of external rewards and must therefore rely on intrinsically generated rewards that are contingent on the similarity between current and target performance. It is not known what kind of RL mechanism and which form of intrinsic reward drive the acquisition of an inventory of multiple actions necessary for performing combinatorial vocal sequences.

Here we attempt to infer the intrinsic similarity-based reward computation underlying the learning of a vocal inventory in birdsong. Birdsongs are typically combinatorial, i.e., made of components (syllables and calls) that are performed in variable sequences[15–17]. Combinatorial vocal ability is present even in species like zebra finches whose natural song has a fixed syllable order[18]. Recent evidence shows that young zebra finches use a highly efficient "greedy" strategy for learning the syllables of their target song (a memorized song of an adult[19]): they make the minimal necessary changes to the sounds in their own vocal repertoire to match the syllables in the target. They achieve this learning of a syllable inventory independently of learning the syllable order[20]. The independence of this learning process suggests that

there is a dedicated reward computation for guiding syllable inventory learning that is not contingent on sequential order.

To model syllable inventory learning, we adopt an actor-critic RL approach where a critic evaluates each performance of an actor in terms of its expected reward[7,21,22]. Because juveniles can independently adjust their syllables to match different targets[20,23–25], and because syllables are represented by distinct neural modules[26–29] (despite often originating from a single duplicated proto-syllable[25,30]), we assume that each syllable is associated with a distinct motor program controlled by an independent RL actor. Since zebra finches can incorporate calls into their song to match target syllables[18,20,24,31] we assume that RL actors include both syllables and calls in a bird's vocal repertoire. Given the small size of neuromodulatory brain centers and the global action of neuromodulators across the brain, it is more parsimonious to assume that all actors are affected alike by a single global intrinsic reward, akin to multi-agent learning systems[32], rather than by multiple reward signals tailored specifically for each actor. We therefore model syllable inventory learning as a multi-actor reinforcement learning (MARL) algorithm, in which a group of syllable and call actors are influenced by a scalar intrinsic reward, and act to nullify an RPE[5], defined as the difference between actual and expected reward.

Our goal is to determine the functional form of the reward – namely, how it is contingent on the acoustic similarity between the actors' performance and the syllables in the target song. An important question is how such a reward can guide the efficient allocation of resources that is evident in song learning behavior – i.e., how it can lead to greedy assignments of actors to their respective targets. A possible way to achieve greedy learning is for actors to compete over targets, leading to each target being matched by the actor whose performance is initially most similar to it. We implement such competitive matching with a reward function that performs a sum-max operation over all possible pairwise actor-target comparisons, first selecting the most acoustically similar actor for each target (max), and then adding up the contributions of the "winners" regardless of their sequential order (sum). We compare this hypothetical reward function to a sequential-order-dependent model adapted from Fiete and colleagues[8] where the reward function is specified through comparisons between actors and targets at corresponding sequential positions in the song, leading to a non-greedy learning of syllables.

If a global scalar reward indeed underlies the learning of song syllable inventories, then a model based on it should reproduce observed song learning behavior. We test diverse reward computation hypotheses by attempting to simulate the learning trajectories of juvenile male zebra finches that are experimentally induced to modify the performance of a learned song to match a new target[20]. We first use a subset of the experimental data to extract model parameters via maximum likelihood estimation, and then, without further parameter tuning, use a separate data set to test the fit between model-simulated and empirical learning trajectories.

We find excellent agreement with observed song-learning trajectories for a sum-max inventory reward where actors compete over targets across short distances in acoustic space. We show that this sum-max reward model successfully predicts a competition between syllables and calls that leads to the exclusion of a syllable from a song and its replacement by a call. We conclude by presenting the predictions of our model for the responses of dopaminergic projections to the songbird basal ganglia known to display error-related firing[5,33], during the learning of a sound inventory.

## Results

### A light-tailed distribution of intrinsic reward explains vocal imitation of a single syllable.

Before examining the full problem of learning a multi-syllable inventory, we start with a simpler one-syllable learning problem where a juvenile bird learns to adjust a single syllable to acoustically resemble

a syllable sung by an adult "tutor" (Figure 1A). We assume that learning is accomplished by an RL agent consisting of an actor and a critic (Figure 1B). This agent attempts to maximize an intrinsically generated reward signal $R$ that is inversely related to the acoustic difference $\Delta$ between the bird's syllable and the target (tutor syllable) $T$. The actor is a motor program that samples variable syllable instances from an underlying *performance distribution* with variable mean $S$. After each instance, an intrinsic reward is generated from a *reward distribution* that is centered on a perfect imitation of the target syllable. The critic computes the expected reward or quality $Q$, from a distribution with the same functional form as the reward but centered at $S$ (i.e., the critic always expects maximal reward at the performance mean; see STAR Methods). The expected reward is compared to the actual reward, and the difference $R - Q$ is the RPE $\delta$, thought to be encoded by dopaminergic neurons[5,33]. Our actor tries to nullify $\delta$ by triggering an update $\Delta S$ of the mean motor program $S$ that brings the quality $Q$ of the instance closer to the actual reward and makes $S$ more similar to the target (see STAR Methods).

For learning to occur, the performance distribution of a syllable must overlap with the intrinsic reward distribution associated with a target (Figure 1B). Whether the two distributions overlap depends not only on the acoustic difference $\Delta$, but also on their respective shape and width, which we set out to estimate from data. We measure the performance distributions of syllables performed by juvenile zebra finches and find them to be of Gaussian shape (see STAR Methods). To estimate the distribution of the putative intrinsic reward $R$ as a function of the acoustic difference $\Delta$, we examine a continuum of hypothetical reward distributions $R(\Delta)$, which we model as a generalized normal distribution[24] with two unknown parameters: the shape $\beta$ and scale $\varsigma$ (Figure 1C; see STAR Methods). The shape parameter $\beta$ produces a normal distribution for $\beta = 2$ as a special case. Larger values ($\beta > 2$) correspond to light-tailed distributions where most of the reward density is concentrated at small acoustic differences from the target (approaching a box distribution as $\beta$ increases), while smaller values ($\beta < 2$) correspond to heavy-tailed distributions, with reward density extending to large acoustic differences from the target. The scale parameter $\varsigma$ determines the standard deviation, and consequently the width of a distribution with a given shape; small values of $\varsigma$ correspond to narrow distributions and large values to wide distributions. Together, these two reward parameters determine the range of influence of a target in acoustic space, with heavy-tailed and/or wide reward distributions exerting long-range influence over dissimilar sounds; and light-tailed and/or narrow distributions exerting short-range influence, able to attract only syllables that are relatively similar to the target.

We infer the model parameters $\beta$ and $\varsigma$ by fitting our actor-critic model to actual learning trajectories from experiments in which juvenile males were trained with artificial song tutors[20] and were induced either to shift the pitch of a syllable to match a syllable in a tutor's song (Figure 1D), or to shift two syllables towards two acoustically different targets (in the latter case, we treated the syllables' learning trajectories as independent of each other). This task was accomplished with a high degree of accuracy, regardless of whether the pitch was shifted up or down by two semitones, which agrees with our assumption that the intrinsic reward distribution is symmetrical (Figures 1E and S1; 13 syllables in 8 birds; part of these data was previously published in Lipkind 2017[20]). Because we were particularly interested in whether reward functions are qualitatively sub- or super-Gaussian, we evaluated fit quality as a function of $\beta$ by fixing $\beta$ to distinct values and estimating $\varsigma$ via maximum likelihood (ML; Figures 1F and S1B; also see STAR Methods).

In 8 out of 10 pitch-shifting syllable trajectories where model parameters were inferred, we found that light-tailed reward distributions ($\beta = 4, 6$, or $8$) result in better fits (i.e., lower Bayesian Information Criterion, or BIC, scores) than normal ($\beta = 2, p < 0.05$) and heavy-tailed ($\beta = 1,\ p < 0.01$) distributions (Benjamini-Hochberg corrected Wilcoxon signed-rank test on non-normalized BICs; Figure 1G; also see Figure S1B). We obtained similar results when we estimated the shape parameter $\beta$ jointly

with other model parameters (Figure S2A). BIC scores alone, however, may yield an overfitted model that fails to generalize to unseen data. Therefore, as a second metric for model comparison, we performed a bootstrap analysis of estimated model parameters by simulating multiple learning trajectories for each ML estimate of $\beta$ (Figures 1H and S1B). Light-tailed reward distributions resulted in improved goodness-of-fit (i.e., smaller Root Mean Square Error, or RSME, between empirical and simulated trajectories) in 6 out of 10 cases (Figures 1I and S2B; in 4 of 10 cases in which the light model is not the best, the decrease in goodness-of-fit compared to heavy-tailed and normal models is small, Figure 1J). These results suggest that a light-tailed reward distribution guides the shift in performance of a syllable towards a match with a target. Namely, a syllable instance would trigger a non-zero intrinsic reward only if it is sufficiently acoustically similar to a target syllable. This solution is a special case of the multi-syllable inventory learning problem that we turn our attention to next.

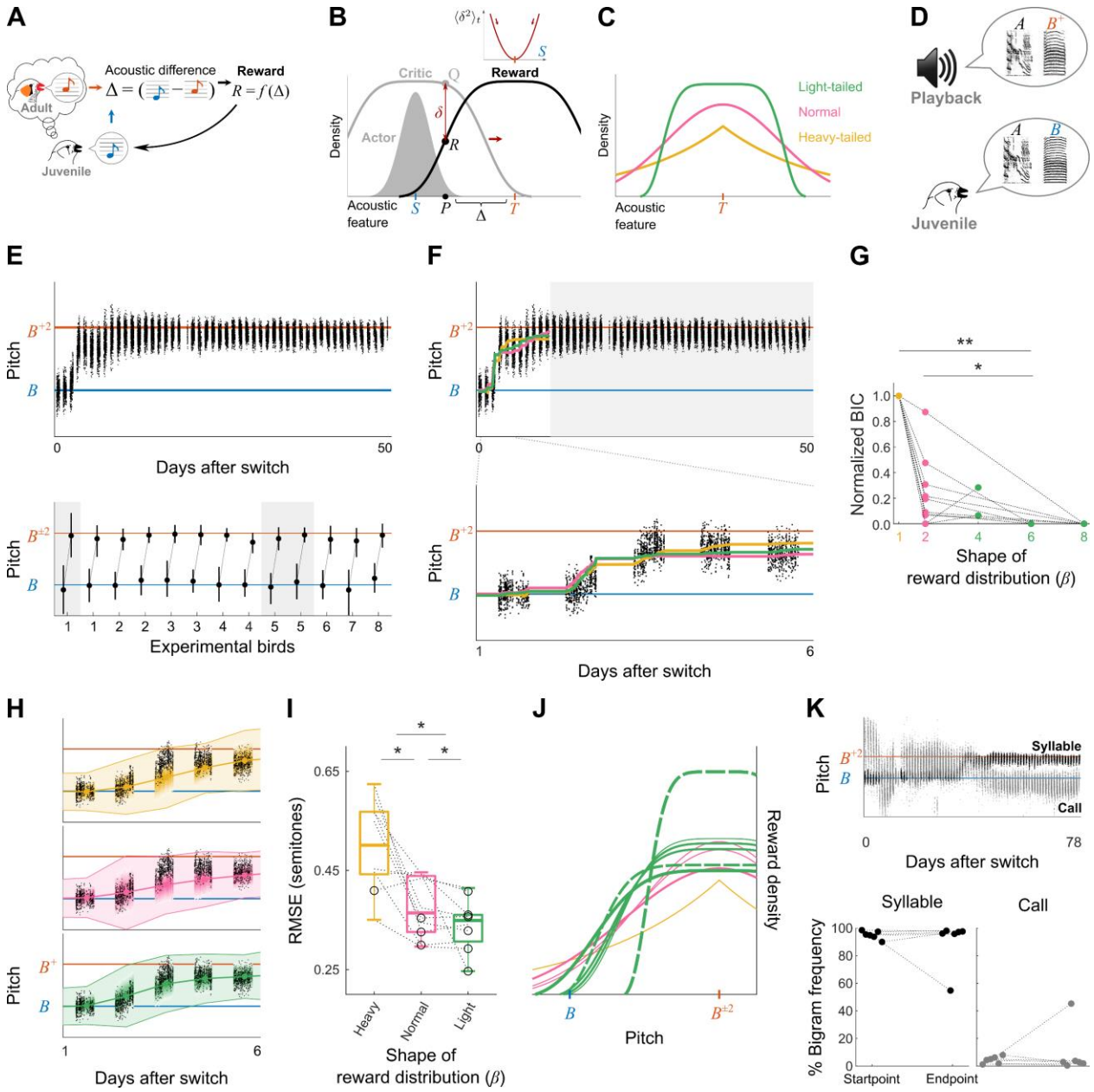

**Figure 1**: **A light-tailed distribution of intrinsic reward explains vocal imitation of a single syllable**

**A**. A juvenile male zebra finch adjusts the performance of a syllable (blue) to acoustically match an adult "tutor" syllable (orange); learning maximizes an intrinsic reward $R$ that is inversely related to the acoustic difference $\Delta$ between the juvenile and tutor syllables.

**B**. An actor-critic model for learning a song syllable: the actor is a motor program that generates variable syllable instances $P$ from a distribution with mean $S$. The critic estimates an expected reward $Q$ for each instance, which is compared to the incoming intrinsic reward $R$, resulting in a reward prediction error (RPE) $\delta = R - Q$. This error is used to update the motor program mean $S$ to reduce future errors (top inset). We assume $Q$ and $R$ have identical distributions, with $R$ being maximal at a perfect imitation of the target ($\Delta = 0$), and $Q$ being maximal at the motor program mean $S$.

**C.** Hypothetical intrinsic reward distributions modelled as a generalized normal distribution. The shape parameter beta determines whether a distribution is light-tailed ($\beta > 2$; green), heavy-tailed ($\beta = 1$; yellow) or normal ($\beta = 2$; magenta).

**D.** An example of a serial tutoring task, in which a bird that has learned a two-syllable song $AB$ is induced to learn a new song $AB^{\pm}$, where the pitch of syllable $B^{\pm}$ is two semitones above or below that of $B$.

**E**. Top, learning trajectory of an experimental bird trained with the task in **D**, showing the median pitch of consecutive instances of syllable $B$ after the presentation of the new tutor song (day 0); blue and orange lines represent the pitch of model syllables $B$ and $B^{\pm}$, respectively. Bottom, learning outcomes in all experimental birds trained with 2-semitone pitch shifting tasks, showing pitch distributions (median and 95% confidence interval, CI) at the start point and end point of learning for each shifted syllable, connected by dotted lines (13 syllables in n = 8 birds). Grey shading denotes 3 syllables excluded from parameter estimation due to lack of convergence (see Figure S1).

**F**. Top, inferred mean pitch shifting trajectories fitted with ML to the learning trajectory of the experimental bird in **E** top, assuming reward distributions with different values of $\beta$ (colors as in **C**). Bottom, zoom-in on the learning part of the pitch trajectory (cf. Figure S1).

**G**. Bayesian Information Criterion (BIC) for each pitch-shifted syllable, assuming reward distributions with different shapes (value of $\beta$; dashed lines connect same-syllable data; n = 10). BIC scores for each syllable are normalized such that 1 corresponds to the worst and 0 to the best fit. Light-tailed reward improves BIC scores compared to heavy-tailed and normal reward (*=p<0.05, **=p<0.01, Benjamini-Hochberg corrected Wilcoxon signed-rank test; cf. Figure S2).

**H.** Observed (black) and bootstrapped (color) pitch trajectories of the pitch-shifted syllable shown in F. Shaded area represents median ± 50% CI of simulated trajectories. Colors as in **C**. Color gradient represents bootstrap density (also see Figure S1).

**I.** Root-mean-square error (RMSE) between observed and bootstrapped learning trajectories (dashed lines connect same-syllable data; open circles represent the RMSE of the best reward model per syllable). Light-tailed reward improves goodness-of-fit compared to heavy-tailed and normal reward ($p < 0.05$, Benjamini-Hochberg corrected Wilcoxon signed-rank test; see Figure S2).

**J.** Best (smallest RMSE) reward distributions for each pitch-shifted syllable (solid green: $\beta = 4$, dashed green: $\beta = 6$). Thicker lines represent larger improvement by the best model compared to the second best.

**K.** Top, an example bird where a call was acoustically closer to the target than the syllable the bird ended up modifying to match it. Bottom, Left, performance frequency of bigrams containing song syllable and call at start and endpoint across birds. In 5 of the 6 experimental birds, syllable-containing bigrams and the corresponding call-containing bigrams did not change in number, indicating that the call was not incorporated into the song motif.

**Light-tailed reward distributions combined in a sum-max operation explain the learning of a multi-syllable inventory**

As a first step in modeling the learning of a multi-syllable inventory, we investigated the respective roles of syllables and calls in learning. Since birds use both syllables and calls in their vocal repertoire to match targets, we wondered whether both these vocalization types are treated equally by the reward computation mechanism, or whether calls serve as "back-up" learning precursors used only when syllables are not available. We inspected call performance in our training dataset (Figure 1) and found several instances where a call was acoustically closer to the target than the syllable the bird ended up modifying to match it (Figures 1K). This observation (which was not originally reported in Lipkind et al 2017), suggests that syllables are prioritized over calls during learning: to match a target, birds first use syllables and only when all useable syllables are exhausted, they use calls. In such a hierarchical process, a closer call to a target will get ignored when a more distant syllable is also present.

Next, we set up to extend our reward model for learning a single target syllable (Figure 1) to a model of a scalar reward that could drive the parallel learning of multiple targets using both syllables and calls in the bird's repertoire (Figure 2A). For that purpose, we used the inferred reward distribution contingent on the pairwise acoustic difference between a syllable and a target (Figures 1F-J) as a functional primitive (or a basis function); we defined a global inventory reward $\boldsymbol{R}$ as a function of all reward primitives $R_{ij}$ associated with pairing target $i$ with syllable (or call) $j$ (Figure 2B). Importantly, this approach allowed us to generalize the model from single to multiple targets without introducing any new model parameters.

The global reward function $\boldsymbol{R}$ must reproduce the context-independent and greedy syllable learning trajectories observed in zebra finches. Context-independence with respect to target positions can be achieved via the sum operation due to its associative property (a sum does not depend on the order in which the summands are added, i.e., it is permutation insensitive). Greediness with respect to syllables can be achieved by the maximum operation (which ignores all values smaller than the maximum). Accordingly, we construct the inventory reward $\boldsymbol{R}$ as a sum-max operation over pairwise reward primitives $R_{ij}$ (Figure 2B): namely, it is a sum $\boldsymbol{R} = \sum_i R_i$ of partial rewards, where the partial reward $R_i = max_j R_{ij}$ associated with target $i$ is given by the maximal pairwise reward that can be obtained for that target, i.e., it is computed with respect to the most acoustically similar vocalization. To implement the prioritization of syllables over calls, the maximum operation in our model is a hierarchical two-stage process: in a first stage, the maximum is taken over syllables, and if no syllable is within the reward range, then in a second stage, the maximum is taken over calls (see STAR Methods and Figure S3A). Note that such two-stage maximum operation can naturally emerge from a light-tailed reward primitive because the latter approximates a decision-making process driven by reward being either inside or outside some range. In contrast, normal and heavy-tailed models are less binary in nature and so an explicit decision-making step would be required to implement a prioritization of syllables over calls.

The inventory reward $\boldsymbol{R}$ is delivered to a group of actors, one actor per vocalization type. Actors produce vocalizations independently of each other, and each of them learns to adjust its mean pitch based on the identical reward $\boldsymbol{R}$ they receive. We assume that reward $\boldsymbol{R}$ is delivered after each syllable or call instance (rather than at the end of a song) and that partial rewards are computed with respect to the vocalization just performed and the means $S$ (i.e., motor memories) of the other syllables and calls in the bird's current inventory (see STAR Methods and Figure S3B).

Each actor is associated with one critic that expects maximal reward for renditions near the actor's mean pitch (as in Figure 1B). Actors learn by computing the difference between the received reward $\boldsymbol{R}$ and an expected reward computed by the critics. This difference signal is the RPE and learning in our

model is driven by minimization of the square RPE. This is achieved by iteratively drawing random syllable and call pitches from their Gaussian distributions and by updating the actors' mean pitches according to a simple iterative equation (see Eq. 11 in STAR Methods). By construction, the square RPE is zero when all targets are matched by either a syllable or a call.

We test our model using learning trajectories from a published set of multi-syllable learning experiments[10], where juvenile males were induced to learn two new targets, both of which resembled a single syllable in their current song (Figure 2C). Birds predominantly shifted the pitch of the syllable to match one of the new targets and recruited a call to match the other target (7 of n = 8 birds, Figures 2D, S3C and S4A; see STAR Methods). We compared simulated learning trajectories generated by light-tailed, normal, and heavy-tailed variants of the model to the empirical learning trajectories (with model parameters as inferred in Figure 1, with the exception of the performance distributions that we estimated in each bird separately). We found that all 3 model variants qualitatively reproduce the experimental results: a syllable shifts to match one of the targets and a call to match the other target (Figures 2E and S4B). However, the light-tailed model was significantly better at matching the empirical trajectories (Figures 2F and S4B), and in one case showed a better qualitative fit to the data, replicating a failure of a bird to match one of the targets (Figures 2G and 2H). In that bird, the performance distribution of the song syllable was relatively narrow, and consequently lay outside the range of light-tailed reward distributions but was within the ranges of normal and heavy-tailed distributions. Therefore, only the light-tailed distribution correctly predicted the failed shifting of the syllable in this bird (Figure 2H, top).

We compare the sum-max reward model to an alternative, sequential-order-dependent (SOD), model adapted from Fiete and colleagues[8]. In the SOD model, reward at each vocalization instance is a function of the similarity between that instance and the temporally aligned target. The reward is also binarized with an adaptive threshold that guarantees equal numbers of rewarded and unrewarded instances. While the sum-max context-independent MARL model makes the correct (empirically observed) assignment of syllables to targets in 8 out of 8 cases, the SOD model succeeds in 4 cases only. In one of these, the binarized reward in the SOD model does not predict the failed shifting observed in Figure 2G. In total, the SOD model resulted in a large simulation error in 5 out of 8 birds, compared to near zero simulation error with the sum-max MARL model (Figure S4C).

In summary, empirical data supports an inventory reward computation model that has the following properties: 1. it drives a competitive matching of targets by motor programs – with each target being efficiently matched by the most acoustically similar motor program. 2. The competition is hierarchical – syllable motor programs are prioritized over call motor programs. 3. The evaluation of motor programs' performance in reference to a target is near-binary rather than continuous; this means that targets can only attract motor programs that are sufficiently close to them in acoustic space.

An interesting prediction of this model is demonstrated by the single case where a bird failed to match one of the targets (Figure 2G): the possibility that a song syllable is excluded from the learning process because its performance distribution does not overlap with the reward distribution of any target. In such a case, our model predicts that the prioritization of syllables over calls could be reversed, and a nearby call in this bird could have shifted to the unoccupied target. However, our inspection of the acoustic space of all vocalizations in the vicinity of the target revealed that no such call was present. Nevertheless, the possibility of a syllable being left out of the competition in favor of a call triggered our interest and we decided to further investigate this prediction of our model.

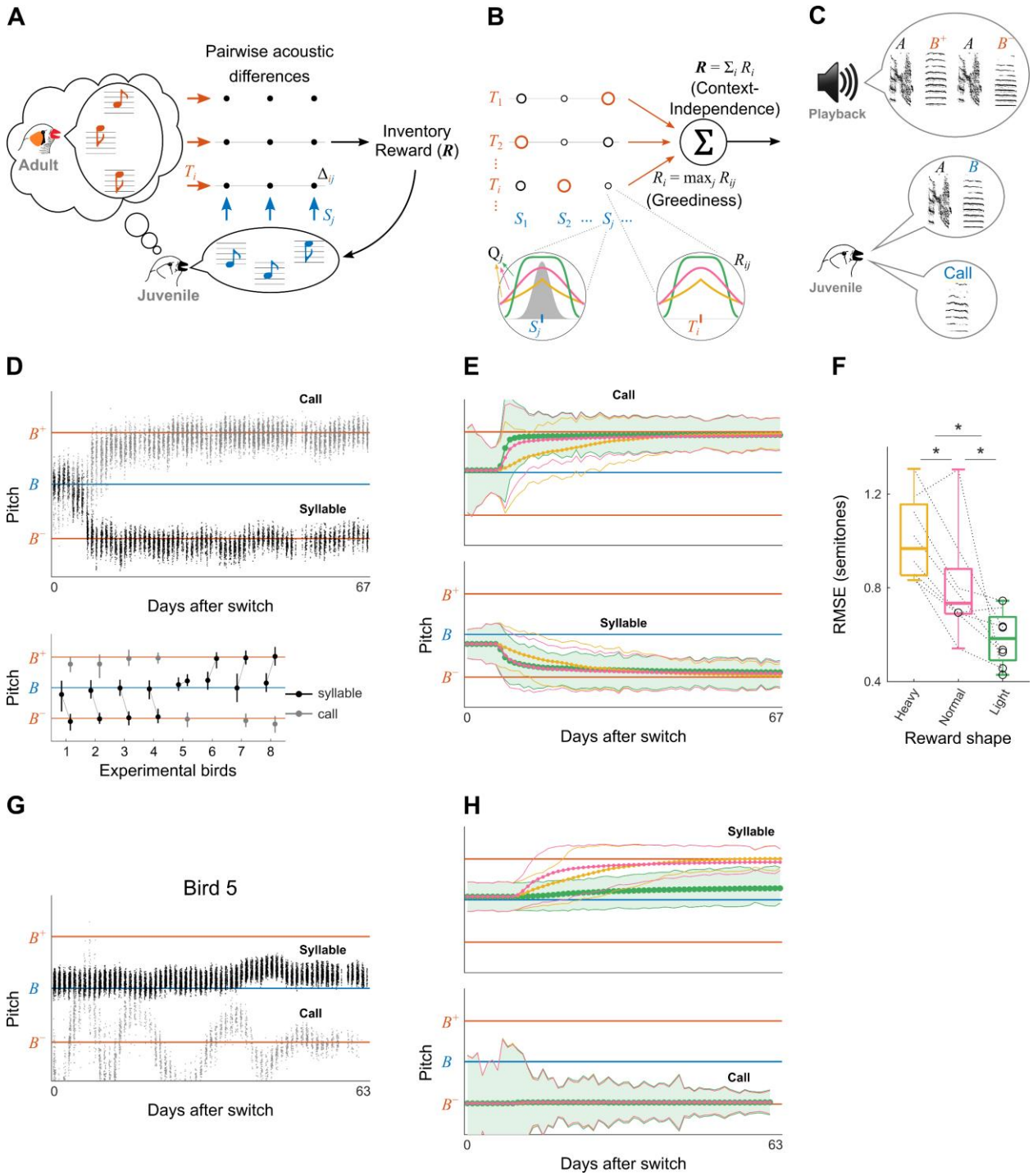


**Figure 2: Light-tailed reward distributions combined in a sum-max operation explain the learning of a multi-syllable inventory.**

**A.** Learning a multi-syllable inventory, driven by a putative intrinsic scalar inventory reward $R$. Filled black circles indicate pairwise acoustic differences $\Delta_{ij}$ between syllables in the bird's song and targets provided by a tutor.

**B.** A hypothesized computation of inventory reward $R$ from reward primitives $R_{ij}$ (circles) contingent on pairwise syllable-target acoustic differences. The reward $R$ is the sum of partial rewards $R_i$ (orange arrows) computed for

each target $T_i$; each partial reward is the maximal reward that can be obtained for that target (orange circles), namely the reward computed with respect to the acoustically closest syllable. The maximum operation in the computation of partial rewards implements greedy assignments of syllables to their targets; the summing of partial rewards implements syllable inventory learning that is independent of the order in which syllables are performed.

**C.** Example of a serial tutoring task, in which a bird that has learned a two-syllable song $AB$ is induced to learn a new song with two new target syllables $B^+$ and $B^-$, with pitch two semitones above and below that of $B$, respectively. The bird's repertoire includes song syllables and calls.

**D.** Top, pitch trajectory in one experimental bird trained with the task in **C**: the bird matched one target by shifting the pitch of song syllable $B$ (black), and the other target by shifting the pitch of a call (grey). Bottom, results across experimental birds (n = 8), showing pitch distributions at start point and endpoint of learning as in Fig 1E (see also Figure S4). Seven out of 8 birds matched one target with a song syllable (black circles), and the other target with a call (grey circles). One bird matched only one of the new targets with a call but did not shift syllable $B$ to the other target.

**E.** Bootstrapped pitch-trajectory distributions of the call (top) and syllable (bottom) of the experimental bird shown in **D** top. Green, light-tailed; Magenta, normal; and yellow, heavy-tailed reward model. Shaded area represents median ± 50% CI (see also Figure S4).

**F.** Root-mean-square error (RMSE) between experimental and bootstrapped syllable learning trajectories. Notation as in Figure 1I. Light-tailed rewards improves goodness-of-fit compared to heavy-tailed and normal rewards ($p < 0.05$, Benjamini-Hochberg corrected Wilcoxon signed-rank test).

**G.** Pitch trajectories of song syllable $B$ and call in an experimental bird that failed to match one of the targets (bird 5 in **D** bottom).

**H.** Same as **E** for the bird shown in **G** (top: syllable; bottom: call). Only the light-tailed reward model predicts the failure to shift song syllable $B$ towards a target.

### Light-tailed sum-max intrinsic reward predicts the exclusion of a syllable from a song

The short-range influence of targets in acoustic space (Figures 1I and 2F) implies that an actor of a syllable may be "left out of the game" — i.e., not contribute to the computation of inventory reward — because it is not within the reward range of any target. In such a case, a call that happens to be within range of a target, may come in to fill the target and replace the syllable in the song. This sounded like an unlikely sequence of events given that it has never been reported before to best of our knowledge.

We tested the prediction of a syllable being excluded by presenting juvenile male zebra finches (n = 6) with a learning task in which one syllable in the target song was shifted up or down by 4 semitones with respect to a syllable in the bird's current song ($AB \rightarrow AB^{\pm 4}$, Figure 3A). In this experimental situation, there is a single off-target syllable in the bird's song, and a single unmatched target in the tutor song, but the two are separated by a relatively large acoustic difference. A hierarchical sum-max reward computation based on *light-tailed* primitives predicts that song syllable $B$ is outside the reward range of the target syllable $B^{\pm 4}$ (Figure S5A) and therefore, that it will not shift to match the target. Consequently, on a hierarchically lower level, if the bird's repertoire happens to contain a call, $C$, that is within the reward range of the target, the call will shift to match it instead. In contrast, normal and heavy-tailed versions of the model predict that song syllable $B$ is within the reward range of a 4-semitone-shifted target, and therefore will shift to match it, even if there is an acoustically closer call in the bird's repertoire (Figure 3B), since syllables are prioritized over calls.

None of the six experimental birds shifted song syllable $B$ towards target $B^{\pm 4}$ (Figures 3C and S5C). In 5 out of 6 birds, $B^{\pm 4}$ was instead matched by a call (or another sound originally performed outside of the song motif), which was initially acoustically closer to the target (e.g., Figure 3C, birds 2 and 3). In all 5 cases, the call was fully or partially incorporated into the song motif (Figures 3D and 3E), replacing song syllable $B$. This was evident from an increase in the frequency of transitions between the call and motif syllables, and a concurrent decrease in the frequency of transitions between song syllable B and other motif syllables (Figure 3E, top). In addition, the performance rate of syllable B decreased with two of the birds stopping performing syllable B altogether (Figures 3E, bottom and S5B). Therefore, we conclude that syllable $B$ was functionally dropped from the song motif, and its function is transferred to a call. In the single case where no call shifted to match target $B^{\pm 4}$, the target remained unmatched, and the bird continued performing syllable $B$ at the original pitch (Figure 3C, bird 4).

The exclusion of the syllable from the song motif and its replacement by a call supports the competitive hierarchical assignment of the bird's own vocalizations to the targets predicted by a sum-max reward computation. That the song syllable did not shift towards the target, even in the absence of a call, supports the short range of influence of targets predicted by a reward computation based on light-tailed pairwise primitives. We further evaluated how well a light-tailed sum-max model fits empirical learning trajectories, compared to normal and heavy-tailed models, by simulating learning trajectories of both syllables and calls (Figures 3F, 3G and S5B). Again, all simulations used previously estimated parameters (Figure 1), without further adjustments to the new data other than the production model of each bird (initial pitch mean and daily pitch variance). We found that the light-tailed model was superior to normal and heavy-tailed model variants, returning smaller errors between simulated and empirical learning trajectories in all birds (Figure 3G), and replicating the failure of the bird that did not have a call at the acoustic vicinity of the $B^{\pm 4}$ target to accomplish the learning task (Figure 3F right). In summary, the experimental outcomes as a whole support a light-tailed sum-max reward computation with prioritization of syllables over calls.

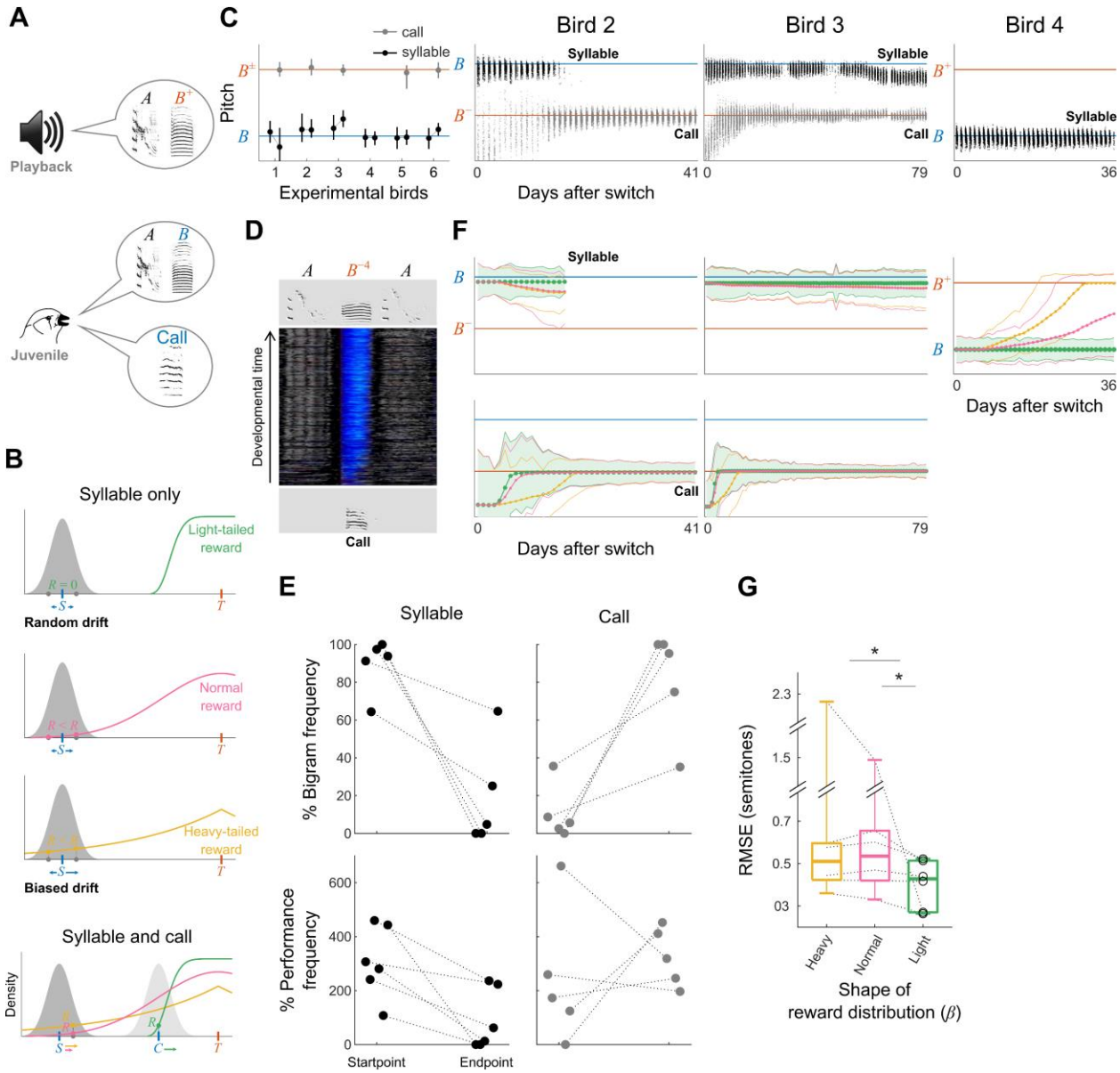


**Fig 3: Light-tailed sum-max intrinsic reward predicts the exclusion of a syllable from a song.**

**A.** A serial tutoring task testing the prediction of a syllable being dropped out of the song: a bird that has learned a two-syllable song $AB$ is induced to learn a new song $AB^{\pm 4}$ where the target $B^{\pm 4}$ is 4 semitones above or below the song syllable $B$. The bird's repertoire includes song syllables and calls.

**B.** Predictions of light-tailed, normal, and heavy-tailed versions of a MARL sum-max reward computation for the experiment in **A**. In a model based on light-tailed reward primitives, song syllable $B$ will not shift towards the target $B^{\pm 4}$ regardless of whether it is the most acoustically similar element to $B^{\pm 4}$(top), or whether a call in the bird's repertoire is acoustically closer to $B^{\pm 4}$(bottom). In contrast, normal or heavy-tailed versions of the model predict that syllable $B$ will shift to match $B^{\pm 4}$ if there is no acoustically closer call (middle and bottom).

**C.** Left, Results across experimental birds (n = 6) trained with the task in **A**. Notation as in Figure 2D. Right, learning trajectories of three experimental birds; from left to right: Bird 2 matched target $B^{\pm 4}$ with a call and stopped performing syllable $B$; Bird 3 matched target $B^{\pm 4}$ with a call and did not shift the pitch of syllable B; Bird 4 did not shift the pitch of syllable B, and did not recruit a call to match target $B^{\pm 4}$.

**D.** Raster aligned on instances of a call that shifted to match the target $B^{\pm 4}$ in a bird trained with the task in **A**; the call, initially performed outside the song motif (bottom), was incorporated into the song motif (top).

**E.** Top left, performance frequency of bigrams containing song syllable $B$ and other motif syllables (% of total bigrams) at experimental startpoint and endpoint across birds. Top right, same as left for the call that shifted to match target $B^{\pm 4}$ in 5 of the 6 experimental birds. Syllable-containing bigrams decreased from start to endpoint of learning, while call-containing bigrams increased, indicating that the call was incorporated into the song motif while syllable $B$ was dropped out. Bottom, Performance frequencies of syllable $B$ (left) and the call (right) that shifted to match target $B^{\pm 4}$ (% of total syllables performed) at the start and endpoint across birds. Syllable performance decreased while call performance increased.

**G.** Bootstrapped (color) pitch-trajectory distributions of the syllable (top) and call (bottom; 2 out of 3 birds) of the three birds shown in **C**. The shaded area represents the median ± 50% CI.

**H.** Root-mean-square error (RMSE) between experimental and bootstrapped syllable learning trajectories. Open circles represent the RMSE of the best reward model per bird. Light-tailed rewards improve goodness-of-fit compared to heavy-tailed and normal rewards ($p < 0.05$, Benjamini-Hochberg corrected Wilcoxon signed-rank test).

## Predicted tuning of error-encoding dopaminergic VTA neurons during the learning of a syllable inventory

The $\delta$ term in our model tracks deviations from expected intrinsic inventory reward (Figure 1B and STAR Methods). Assuming that $\delta$ is proportional to the putative firing rate of dopaminergic neurons, our model provides predictions for dopaminergic firing in juvenile birds learning a multi-syllable inventory.

We describe our model's predictions in a juvenile bird at the onset of learning a syllable inventory (Figure 4). We simulate the putative firing in each of our experimental scenarios: Learning a single target syllable shifted by 2 semitones from a syllable in the bird's song (Figure 1D), leaning two different 2-semitone-shifted targets (Figure 2C), and learning a 4-semitone-shifted target (Figure 3A).

In the case of learning a single 2-semitone-shifted target (Figure 4A), the light-tailed reward model predicts selective tuning of VTA neurons to pitch values at the target-adjacent tail of the performance distribution (i.e., target selectivity) so that syllable pitch is driven in the direction of higher firing rate, i.e., smaller deviation from the target. Target selectivity is restricted to a small range of pitch values at the tail, where the performance distribution overlaps with the light-tailed reward distribution. This means that VTA tuning is mostly insensitive to variations in the acoustic difference $\Delta$ between syllable and target pitch. This is consistent with reports of weak sensitivity of VTA response to the magnitude of negative RPEs upon external reinforcement[5,34,35], which comes here as a direct consequence of the intrinsic light-tailed reward. The heavy-tailed and normal reward models also predict target selectivity of VTA neuron responses but fail to account for insensitivity to variations in $\Delta$.

In the case of learning two different 2-semitone-shifted targets (Figure 4B), VTA tuning is expected to depend not only on the performance distribution of the currently sung syllable, but also on the distributions of other syllables (or calls) in a bird's repertoire: The absence of competing syllables (i.e., of syllables that are within the reward range of one of the targets) would result in a symmetric VTA tuning (although in practice, a syllable would always be slightly closer to one of the targets, resulting in a slight asymmetry). In contrast, a competing syllable proximal to one target would break this symmetry, rendering the VTA neuron more selective to the other target, thus promoting one-to-one assignments of motor programs to targets, and avoiding redundant matching of the same target. In the case of a 4-

semitone target (Figure 4C), the light-tailed reward model, unlike the heavy-tailed and normal reward models, predicts lack of target selectivity, so that syllable pitch drifts randomly around its own mean.

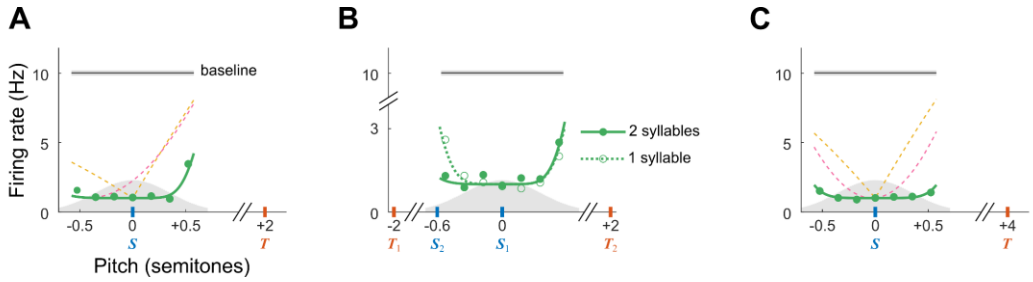


**Figure 4: Predicted tuning of error-encoding dopaminergic VTA neurons during the learning of a syllable inventory**

**A**. Predicted VTA tuning as a function of pitch in the case of a target ($T$, orange) shifted by 2 semitones with respect to a song syllable ($S$, blue). Green, theoretical (solid line) and simulated (filled circles) VTA tuning predicted by a light-tailed reward. Dashed lines, theoretical VTA tuning predicted by the heavy-tailed reward (yellow) and normal reward (pink).

**B.** VTA tuning predicted by a light-tailed reward model in the case of two targets ($T_1$ and $T_2$). When a single syllable $S_1$ is at a distance of 2 semitones from both targets, VTA tuning is symmetrical (dotted line, empty circles). The presence of a second syllable $S_2$ between $S_1$ and $T_1$ breaks this symmetry, resulting in the VTA neuron firing more in the vicinity of $T_2$.

**C.** Same as A for a 4-semitone-shifted target.

**A-C.** All VTA-neuron simulations are based on 2000 pitch values drawn from a normal performance distribution (grey) with a typical standard deviation at the onset of learning ($\sigma_B = 0.35$ semitones). Firing rates, averaged in bins of size $0.5\sigma_B$, are in the range 1 to 10 Hz, where the higher end corresponds to the on-target, baseline firing rate. The hypothetical neuron emits Poisson spikes at a rate that scales linearly with RPE within a 150 ms window, beginning 50 ms after syllable onset. Baseline shows mean +/- SEM.

## Discussion

Our work shows that an inventory of actions underlying a combinatorial behavior – birdsong – can be learned in a self-reinforced manner from a simple scalar reward by a set of independent actors. Our model assumes independence among actors, critics, and partial rewards, which makes inventory learning easily scalable. Each actor controls the acoustic mean and variance of a single syllable (or call) and it does that independently of other actors. Associated with each actor is a critic in the value system that estimates the expected reward following a syllable instance, independently of other critics. This independence among syllables in the action and value systems is mirrored by independence among targets in the reward system. Namely, the inventory reward is a sum of partial rewards, each of which is computed with respect to only one target, irrespective of the existence of other targets.

The organization of our model into independent sub-modules promotes expansiveness and flexibility of the inventory. Adding a target to the inventory turns into the simple problem of computing its associated reward and summing it up with the other partial rewards. The same goes for the removal of a target, which translates into elimination of that partial reward from the computation. The flexible goals this architecture supports may be an evolutionary adaptation for coping with changing targets in a dynamic environment and could explain the ability of juvenile zebra finches to combine syllables from

multiple tutors under natural breeding conditions[36]. Given the flexibility and efficiency it provides, such an architecture may have evolved for the intrinsically guided learning of other combinatorial behaviors besides birdsong, perhaps including the learning of speech sound inventories of human languages.

The neural organization that could carry out a sum-max reward computation seems drastically more complex than that of previous actor-critic models of birdsong learning[7,8], which assumed that syllables are learned within the sequential context of the song motif rather than independently of it (note that context-independent inventory learning is also offered in the model of Troyer and Doupe[37] where song performance is evaluated in parallel modules, one for each target syllable; however, these modules are combined via rich synaptic weights and so the overall evaluation in this model is far from being scalar in nature as the dopamine signal).

In our model, the intrinsic reward circuit needs, in principle, to perform all possible pairwise comparisons at each time step, leading to a total of $N \times M$ maximum operations per song motif, each over a set of $N$ terms (for $M$ targets and $N$ syllables). Although this sounds like a daunting number for large inventories, the computation is quite manageable in practice, since of these $N^2 \times M$ terms only at most $2N \times M$ are distinct. These are the $N \times M$ terms comparing the targets with currently sung syllable instances (which are computed on the fly and need not be held in short-term memory) and the $N \times M$ terms comparing the targets with the means $S_j$ of the motor programs (i.e., motor memories). Such computation can be carried out by a network consisting of $N\ x\ M$ pairwise comparator neurons arranged in parallel winner-take-all modules that compute a max operation with respect to each target, followed by an output summation (Figure 5A). Each pairwise comparator in this network would represent information on a target syllable and the mean motor program $S_j$ either in terms of the synaptic connections they make or the synaptic inputs they receive. Furthermore, comparators would receive auditory input $P_j$ corresponding to feedback from the syllable instance just sung, which would replace the representation of the syllable's average performance $S_j$ in case it was not just sung. Comparators belonging to the same winner-take-all module would be connected by inhibitory synapses with each other. All these computations are feasible with neural networks, but the details of a simulated model would need to be worked out.

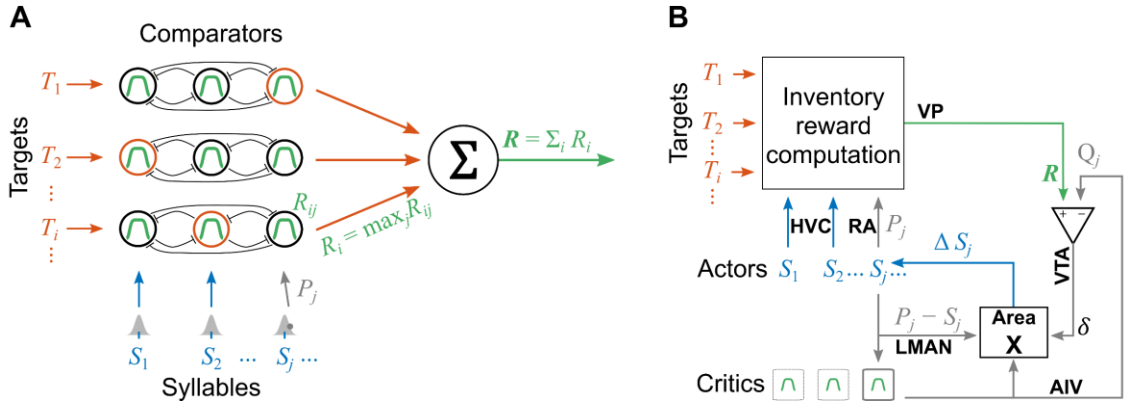


**Figure 5: Neural implementation of sum-max inventory reward and a possible mapping of the MARL model onto the songbird brain.**

**A**. To compute inventory reward, comparator neurons carry out the max operation with respect to each target through lateral inhibition of other neurons within the same winner-take-all module. At any instance, comparator firing reflects the pitch $P_j$ of the produced syllable instance and of the mean pitches ($S_1, S_2, \ldots$) of the non-produced syllables.

**B.** Potential mapping of the MARL model for inventory learning onto the songbird brain, with the following brain areas exhibiting functions analogous to the model components indicated in parentheses: VTA output ($\delta$), RA output ($P_j$), HVC output ($S_j$), VP input to VTA ($\boldsymbol{R}$) AIV input to VTA ($Q_j$), LMAN input to Area X ($P_j - S_j$), and Area X ($\Delta S_j$).

The tuning curves of the putative comparator neurons in this network implement the reward distributions associated with pairwise syllable-target comparisons. We found best fit with empirical data when these reward distributions are light-tailed. Therefore, our model predicts light-tailed tuning curves for the comparators, meaning that they would be tuned to small mismatches between a syllable and a target, but would be unresponsive to large mismatches. Thus, the seemingly many terms in the inventory reward computation may further boil down to only a few nonzero ones associated with proximal syllable-target pairs, suggesting that light-tailed reward distributions promote efficiency.

How would our model be implemented in the songbird brain (Figure 5B)? Since the RPE $\delta$ delivered by the VTA[5] is the difference between actual and expected reward, the circuits generating both inputs to the RPE computation must reside upstream of VTA. These are the putative comparator network computing the inventory reward $\boldsymbol{R}$ and the critics which compute expected reward $Q$ for each actor. Candidate sites for the inventory reward are the ventral pallidum (VP) and its afferents, given that VP input to VTA is of positive valence[38] and therefore acts like the inventory reward $\boldsymbol{R}$ on $\delta$ (although another study found mixed signals in VP[39]). The critics would be located in the ventral intermediate arcopallium (AIV) and its afferents, given that AIV input to VTA[40] is of negative valence[38] and so acts like $Q$ on $\delta$.

The actors we assume are distributed across HVC, the robust nucleus of the arcopallium (RA), and the lateral magnocellular nucleus of the anterior nidopallium (LMAN). The actors generate stable motor memories of syllables $S$, which would come from HVC[41], and the syllable instance just performed $P$ (and possibly its efference copy), which could come from RA. The latter signal is the sum of two components – a stable component $S$ delivered from HVC, and the more variable component $P - S$ from LMAN[21]. The LMAN signal $P - S$ would be relayed to the dopaminergic recipient Area X where the learning itself could take place, in agreement both with a previous hypothesis[21] (and corroborating anatomical evidence[42]. The update $\Delta S$ computed in Area X would be consolidated in RA and HVC in a process that we do not explicitly model[43].

The site in our model, where auditory feedback (AF) could enter is at the level of input to the comparators and as input to the critic (instead of the efference copy $P - S$ from LMAN). Such assumptions would agree with observations that neurons in VP and AIV are highly sensitive to distorted auditory feedback during singing[44], which means that they would compute reward and quality based on evaluations of auditory pitch.

It is important to note that, although our model foresees a use of critics in the construction of RPE, critics are not needed in principle. That is, by construction, the goal of our model ($\delta$ model) is to minimize the square RPE, which we did to align ourselves with widespread belief that dopaminergic neurons encode RPEs. However, instead, we could have chosen as goal to maximize intrinsic reward, which needs neither an expected (intrinsic) reward nor a critic. Such a model ($R$ model) would be simpler in fact, as it would free up resources used to compute expected reward. Under the $R$ model, assuming that dopaminergic responses encode $\boldsymbol{R}$ instead of $\delta$, the experimental predictions would be high firing rates in VTA neurons of adult birds with an accomplished song (i.e. triggering maximal reward) and light-tailed singing-related pitch tuning curves (essentially the tuning corresponds to the reward primitive).

However, the rather low average song-related $VTA_X$ firing rate of 13Hz in adults[45] is not in particularly strong support of this view, it suggests that dopaminergic neurons encode a difference signal, as in the $\delta$ model. Neither is the tendency of VTA neurons to display maximal firing at a single vocal target[33] in unequivocally strong support of the $R$ model. In summary, it is not possible to unequivocally arbitrate between the $\delta$ and $R$ models, the $R$ model may be simpler but the $\delta$ model may map better onto known firing properties.

Our model is a first attempt to fit a reinforcement learning model to developmental birdsong data. We studied pitch, mainly because of our ability to selectively manipulate this acoustic feature. It remains to be seen how this approach will generalize to other sound features. Presumably, the pairwise rewards will need to be tuned to these other features, and the same goes for the critics, which must be responsive to these features as well. We made the simplifying assumption that the critics and the reward basis functions have the same tuning shape to acoustic differences. This convenient simplification may be relaxed in a more powerful version of the model, in which the critic learns based on function approximation methods.

The evolutionary driving force that shapes birds' learning strategy may be efficiency, to be able to reach a target with fewer attempts and with less effort. But efficiency is a hallmark not just of evolution and its selective pressures but is similarly sought for by humans when developing machine learning algorithms. It is therefore interesting to consider possible convergences of natural and artificial learning algorithms. Our findings add to the remarkable parallels between birds' learning strategy and computational approaches in natural language processing (NLP). Previously, we have noted that birds' song learning strategy closely mirrors Word Mover's Distance (WMD), which is a powerful NLP method for document retrieval[46]. Our findings extend this analogy in terms of the efficiency gained from light-tailed reward, which mirrors the improvement of WMD gained from ignoring distant assignments between words, as detailed in the following.

Document retrieval is an NLP problem, where based on a given document one must retrieve highly similar documents from a corpus. In WMD, words in the source document (within some high-dimensional embedding space) are transported with minimal transport cost to words in the target document, and the sum of all pairwise distances is a good proxy for the semantic distance between the two documents. WMD performs a balanced transport in which all words are subject to transportation, which mirrored our previous belief that every syllable must match a target. However, it was recently found that an unbalanced version of the optimal transport problem[47], where some words can be ignored, achieves higher retrieval performance than WMD and closer to human levels, especially when documents being compared are of very different lengths. Thus, similarly to birds that ignore syllables that are distant from a target, humans seem to implicitly establish text correspondences without much consideration of dissimilar words. Both birds' song-learning strategies and human language may therefore underlie common neural mechanism that approximate unbalanced optimal transport.

One idea we would like to put forward is that songbirds' light-tailed inventory reward can serve as an inspiration for easing the computational burden of certain state-of-the art neural network language models. Namely, a similar quadratic scaling bottleneck as we find has been modeled in transformer-based neural networks and its attention mechanism[48]. On the one hand, the quadratic scaling of the computational cost on input text size has been argued to be a strength and to constitute the main reason for the success of these models. Thus, the quadratic complexity may be justified thanks to the computational benefits it provides, which is to massively inspect pairings of smaller language units to make use of the best choices. But perhaps, taking inspiration from our findings, not all pairings are equally important and many of them can be neglected. Currently, there are efforts underway to ease

the quadratic scaling in transformer networks. One possible approach suggested by our findings would be to endow the attention mechanism with light-tailed interactions, which might provide an effective means to reduce the number of word-word interactions that need to be inspected, with minimal degradation of performance.

**References:**


1. Simonyan, K., Horwitz, B. & Jarvis, E. D. Dopamine regulation of human speech and bird song: A critical review. *Brain Lang* **122**, 142–150 (2012).

2. Hisey, E., Kearney, M. G., Mooney, R. & Carolina, N. Externally Reinforced Forms of Motor Learning. **21**, 589–597 (2018).

3. Hoffmann, L. A., Saravanan, V., Wood, A. N., He, L. & Sober, S. J. Dopaminergic contributions to vocal learning. *Journal of Neuroscience* **36**, 2176–2189 (2016).

4. Xiao, L. *et al.* A Basal Ganglia Circuit Sufficient to Guide Birdsong Learning. *Neuron* **98**, 208-221.e5 (2018).

5. Gadagkar, V. *et al. Dopamine neurons encode performance error in singing birds*. http://science.sciencemag.org/.

6. Sutton, R. S. & Barto, A. G. *Reinforcement learning : an introduction*.

7. Doya, K. & Sejnowski, T. J. A Computational Model of Avian Song Learning. *The New Cognitive Neurosciences,* 469–482 (2002).

8. Fiete, I. R., Fee, M. S. & Seung, H. S. Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances. *J Neurophysiol* **98**, 2038–2057 (2007).

9. Day, J. J., Roitman, M. F., Wightman, R. M. & Carelli, R. M. Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nat Neurosci* **10**, 1020–1028 (2007).

10. Schultz, W., Dayan, P. & Read Montague, P. *A Neural Substrate of Prediction and Reward*. *Neuropsychopharmacology* vol. 28 http://www.sciencemag.org (1990).

11. Cohen, J., Haesler, S., L, V., BB, L. & Uchida, N. Neuron-type specific signals for reward and punishment in the ventral tegmental area. *Nature* **482**, 85–88 (2012).

12. D'Ardenne, K., McClure, S. M., Nystrom, L. E. & Cohen, J. D. BOLD Responses Reflecting Dopaminergic Signals in the Human Ventral Tegmental Area. *Science (1979)* **319**, 1264–1267 (2008).

13. Hisey, E., Kearney, M. G. & Mooney, R. A common neural circuit mechanism for internally guided and externally reinforced forms of motor learning. *Nat Neurosci* **21**, 589–597 (2018).

14. Xiao, L. & Roberts, T. F. What Is the Role of Thalamostriatal Circuits in Learning Vocal Sequences? *Front Neural Circuits* **15**, (2021).

15. Soha, J. A. & Marler, P. Vocal syntax development in the white-crowned sparrow (Zonotrichia leucophrys). *J Comp Psychol* **115**, 172–180 (2001).

16. Gentner, T. Q., Fenn, K. M., Margoliash, D. & Nusbaum, H. C. Recursive syntactic pattern learning by songbirds. *Nature* **440**, 1204–7 (2006).

17. Immelmann, K. Song development in the zebra finch and other Estrildid finches. in *Bird vocalizations* (ed. Hinde, R.) 61–74 (Cambridge University Press, 1969).

18. Lipkind, D. *et al.* Stepwise acquisition of vocal combinatorial capacity in songbirds and human infants. *Nature* **498**, 104–108 (2013).

19. Funabiki, Y. & Konishi, M. Long memory in song learning by zebra finches. *J Neurosci* **23**, 6928–35 (2003).

20. Lipkind, D. *et al.* Songbirds work around computational complexity by learning song vocabulary independently of sequence. *Nat Commun* **8**, (2017).

21. Fee, M. S. & Goldberg, J. H. A hypothesis for basal ganglia-dependent reinforcement learning in the songbird. *Neuroscience* **198**, 152–70 (2011).

22. Chen, R. & Goldberg, J. H. Actor-critic reinforcement learning in the songbird. *Current Opinion in Neurobiology* vol. 65 1–9 Preprint at https://doi.org/10.1016/j.conb.2020.08.005 (2020).

23. Liu, W.-C., Gardner, T. J. & Nottebohm, F. Juvenile zebra finches can use multiple strategies to learn the same song. *Proc Natl Acad Sci U S A* **101**, 18177–18182 (2004).

24. Ravbar, P., Lipkind, D., Parra, L. C. & Tchernichovski, O. Vocal exploration is locally regulated during song learning. *Journal of Neuroscience* **32**, 3422–3432 (2012).

25. Tchernichovski, O., Mitra, P. P., Lints, T. & Nottebohm, F. *Dynamics of the Vocal Imitation Process: How a Zebra Finch Learns Its Song*. http://science.sciencemag.org/.

26. Kozhevnikov, A. A. & Fee, M. S. Singing-related activity of identified HVC neurons in the zebra finch. *J Neurophysiol* **97**, 4271–4283 (2007).

27. Hahnloser, R. H. R., Kozhevnikov, A. A. & Fee, M. S. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* **419**, 65–70 (2002).

28. Tian, L. Y. & Brainard, M. S. Discrete Circuits Support Generalized versus Context-Specific Vocal Learning in the Songbird. *Neuron* **96**, 1168-1177.e5 (2017).

29. Vallentin, D., Kosche, G., Lipkind, D. & Long, M. A. Neural circuits: Inhibition protects acquired song segments during vocal learning in zebra finches. *Science (1979)* **351**, 267–271 (2016).

30. Okubo, T. S., Mackevicius, E. L., Payne, H. L., Lynch, G. F. & Fee, M. S. Growth and splitting of neural sequences in songbird vocal development. *Nature* **528**, 352–357 (2015).

31. Okubo, T. S., Mackevicius, E. L., Payne, H. L., Lynch, G. F. & Fee, M. S. Growth and splitting of neural sequences in songbird vocal development. *Nature* **528**, 352–357 (2015).

32. Schilling, M., Konen, K., Ohl, F. W. & Korthals, T. Decentralized deep reinforcement learning for a distributed and adaptive locomotion controller of a hexapod robot. in *IEEE International Conference on Intelligent Robots and Systems* 5335–5342 (Institute of Electrical and Electronics Engineers Inc., 2020). doi:10.1109/IROS45743.2020.9341754.

33. Duffy, A., Latimer, K. W., Goldberg, J. H., Fairhall, A. L. & Gadagkar, V. Dopamine neurons evaluate natural fluctuations in performance quality. *Cell Rep* **38**, (2022).

34. Bayer, H. M. & Glimcher, P. W. *Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal*.

35. Fiorillo, C. D. Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. (2015) doi:10.1126/science.1077349.

36. Williams, H. Models for song learning in the zebra finch: fathers or others? *Anim Behav* (1990) doi:10.1016/S0003-3472(05)80386-0.

37. Troyer, T. W. & Doupe, A. J. An associational model of birdsong sensorimotor learning I. Efference copy and the learning of song syllables. *J Neurophysiol* **84**, 1224–39 (2000).

38. Kearney, M. G., Warren, T. L., Hisey, E., Qi, J. & Mooney, R. Discrete Evaluative and Premotor Circuits Enable Vocal Learning in Songbirds. *Neuron* **104**, 559-575.e6 (2019).

39. Chen, R. *et al.* Songbird Ventral Pallidum Sends Diverse Performance Error Signals to Dopaminergic Midbrain. *Neuron* **103**, 266-276.e4 (2019).

40. Mandelblat-Cerf, Y., Las, L., Denissenko, N. & Fee, M. A role for descending auditory cortical projections in songbird vocal learning. *Elife* **2014**, 1–23 (2014).

41. Aronov, D., Andalman, A. S. & Fee, M. S. *A Specialized Forebrain Circuit for Vocal Babbling in the Juvenile Songbird*. http://science.sciencemag.org/.

42. Kornfeld, J. *et al.* An anatomical substrate of credit assignment in reinforcement learning. *bioRxiv* (2020) doi:10.1101/2020.02.18.954354.

43. Tachibana, R. O., Lee, D., Kai, K. & Kojima, S. Performance-Dependent Consolidation of Learned Vocal Changes in Adult Songbirds. *The Journal of Neuroscience* **42**, 1974–1986 (2022).

44. Keller, G. B. & Hahnloser, R. H. R. Neural processing of auditory feedback during vocal practice in a songbird. *Nature* **457**, 187–90 (2009).

45. Duffy, A., Abe, E., Perkel, D. J. & Fairhall, A. L. Variation in sequence dynamics improves maintenance of stereotyped behavior in an example from bird song. *Proc Natl Acad Sci U S A* **116**, 9592–9597 (2019).

46. Huang, G., Guo, C., Kusner, M., Sun, Y. & Sha, F. Supervised Word Mover's Distance. *Advances in Neural* (2016).

47. Chizat, L., Peyré, G., Schmitzer, B. & Vialard, F.-X. Unbalanced Optimal Transport: Dynamic and Kantorovich Formulation. (2015).

48. Vaswani, A. *et al.* Attention is all you need. *ArXiv* (2017).