Learning the sound inventory of a complex vocal skill via an intrinsic reward

Author(s):
Toutounji, Hazem; Zai, Anja T.; Tchernichovski, Ofer; Hahnloser, Richard H.R.; Lipkind, Dina

Publication date:
2024-03

Permanent link:
https://doi.org/10.3929/ethz-b-000668278

Originally published in:
Science Advances 10(13), https://doi.org/10.1126/sciadv.adj3824

NEUROSCIENCE

Learning the sound inventory of a complex vocal skill via an intrinsic reward

Hazem Toutounji[1,2,3]*, Anja T. Zai[4,5], Ofer Tchernichovski[6], Richard H. R. Hahnloser[4,5]†, Dina Lipkind[7]*†

Reinforcement learning (RL) is thought to underlie the acquisition of vocal skills like birdsong and speech, where sounding like one's "tutor" is rewarding. However, what RL strategy generates the rich sound inventories for song or speech? We find that the standard actor-critic model of birdsong learning fails to explain juvenile zebra finches' efficient learning of multiple syllables. However, when we replace a single actor with multiple independent actors that jointly maximize a common intrinsic reward, then birds' empirical learning trajectories are accurately reproduced. The influence of each actor (syllable) on the magnitude of global reward is competitively determined by its acoustic similarity to target syllables. This leads to each actor matching the target it is closest to and, occasionally, to the competitive exclusion of an actor from the learning process (i.e., the learned song). We propose that a competitive-cooperative multi-actor RL (MARL) algorithm is key for the efficient learning of the action inventory of a complex skill.

INTRODUCTION

Animal behavior provides a unique opportunity for understanding evolutionary solutions to complex learning problems. One prime example is learning the inventory of components for combinatorial vocal skills such as speech sounds or birdsong syllables. In both humans and songbirds, the acquisition of vocal skills is thought to be subserved by a reinforcement learning (RL) mechanism (*1*–*3*), as evidenced by dopamine signaling (*4*–*8*). Dopaminergic neurons 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 (*4*). Such error signal coding is at the heart of many hypothesized RL mechanisms of developmental birdsong learning (*6*, *8*, *13*).

The goal of RL is to maximize rewards, but internally motivated processes such as speech or birdsong learning readily occur in the absence of external rewards. Songbirds can successfully learn to imitate song recordings even in complete social isolation and without any external feedback contingent on their performance (*14*–*18*), except for sensory feedback of their own song. Learning must therefore rely on intrinsically generated reward signals that are contingent on the similarity between current and target performance. Namely, sounding like one's "tutor" must be rewarding, and/or sounding dissimilar is aversive. However, it is not known what kind of RL mechanism and which form of intrinsic reward drive the acquisition of an inventory of sounds necessary for performing combinatorial vocal sequences.

Here, we attempt to infer the intrinsic reward underlying the learning of a syllable inventory in zebra finches. We assess inventory learning in terms of pitch, which zebra finches readily imitate (*14*, *18*). Zebra finches have vocal combinatorial ability (*17*), despite typically singing a fixed syllable sequence. Juveniles learn the syllable inventory of their target song [a memorized song of an adult (*19*)] independently of syllable order, using a highly efficient "greedy" strategy (*18*). Namely, they make the minimal necessary changes to the sounds in their own vocal repertoire to match the syllables in the target song. This suggests a dedicated reward computation for syllable inventory learning that is not contingent on sequential order. Our goal is to determine the functional form of the reward, namely, how it is contingent on the pitch similarity between the syllables a bird performs and the syllables in its target song.

We develop a multi-actor RL (MARL) model where independent RL actors control the performance of distinct syllables. Akin to multi-agent learning systems (*20*), actors maximize a common intrinsic reward by nullifying an RPE (*4*). We implement the greedy and order-independent learning observed in zebra finches by competitive-cooperative interactions among actors: (i) actors compete over target syllables leading to each target being matched by the most similar actor and (ii) actors cooperate to maximize reward that increases in proportion to the number of matched targets. Such interactions require a reward function that takes all possible pairwise actor-target comparisons into account.

We test our model against several computationally simpler alternatives on the task of simulating the empirical learning trajectories (*18*) of juvenile male zebra finches that are experimentally induced to learn new syllables. We find excellent agreement with data for a competitive-cooperative reward model that acts over short distances (between actors and targets) in pitch space. The model successfully predicts a competitive hierarchy between syllables and calls in an experimental test where we incite birds to exclude an existing syllable from a song and replace it with a call. We conclude by presenting the predictions of our model for the responses of dopaminergic projections to the songbird basal ganglia during the learning of a vocal inventory.

[1]Department of Psychology, University of Sheffield, Sheffield S1 4DP, UK. [2]Insigneo Institute for in silico Medicine, University of Sheffield, Sheffield S1 3JD, UK. [3]The Neuroscience Institute, University of Sheffield, Sheffield S10 2HQ, UK. [4]Institute of Neuroinformatics, University of Zurich/ETH Zurich, Zurich CH-8057, Switzerland. [5]Neuroscience Center Zurich (ZNZ), Zurich CH-8057, Switzerland. [6]Department of Psychology, Hunter College, The City University of New York, New York, NY 10065, USA. [7]Department of Biology, York College, The City University of New York, New York, NY 11451, USA.
*Corresponding author. Email: h.toutounji@sheffield.ac.uk (H.T.); dlipkind@york.cuny.edu (D.L.)
†These authors contributed equally to this work.

RESULTS

Vocal imitation of a single syllable best agrees with a light-tailed distribution of intrinsic reward

We started with a simple one-syllable learning problem: A juvenile bird learns to adjust the pitch of a single syllable to resemble a target syllable sung by an adult tutor (Fig. 1A). We modeled the learning system as an RL agent consisting of an actor and a critic. This agent attempts to maximize an intrinsically generated reward R that is inversely related to pitch difference Δ between the bird's syllable and the target. The actor is a motor program that samples variable syllable instances P from an underlying pitch performance distribution parameterized by its mean S (Fig. 1B). After each instance, an intrinsic reward R is generated from a reward distribution that is centered on a perfect imitation of the target syllable T ($\Delta = 0$). The actor is associated with a critic that expects maximal reward for renditions near the actor's mean performance S; the critic computes the expected reward or quality Q from a distribution with the same functional form as the reward but centered at S. The actor learns by computing the difference between the received reward R and the quality Q. This difference $\delta = R - Q$ is the RPE thought to be encoded by dopaminergic neurons (*4*, *21*). Learning is driven by minimization of the square RPE (Fig. 1C). Each syllable instance leads to an update ΔS of the actor's mean performance S according to a simple iterative rule (Eq. 7; Materials and Methods). This update brings the quality Q of the instance closer to the actual reward R and makes S more similar to the target T (see Materials and Methods). Syllable instances in the vicinity of S almost always result in a negative RPE (fig. S1). That is because the critic expects maximal reward near S, whereas the actual reward is maximal only when S matches T. However, the RPE is larger (less negative) for instances on the target side of S than instances on the opposite side, which is what shifts S toward T. By construction, the RPE is zero, and learning stops when S coincides with T.

We estimated the shape of the pitch performance distributions of zebra finch syllables from published data (*18*) and found them to be Gaussian (see Materials and Methods). To estimate the distribution of the putative intrinsic reward R as a function of pitch difference Δ, we examined a continuum of hypothetical reward distributions $R(\Delta)$, which we modeled as a generalized normal distribution (*22*) with two unknown parameters: the shape β and scale ς (Fig. 1D; see Materials and Methods). The scale parameter ς determines the SD and, consequently, the width of the distribution. The shape parameter β produces a normal distribution for $\beta = 2$ as a special case. Smaller values ($\beta < 2$) correspond to heavy-tailed distributions, and larger values ($\beta > 2$) correspond to light-tailed distributions. Heavy-tailed reward distributions exert long-range influence over syllables at a large pitch difference from the target. Light-tailed distributions exert short-range influence, rewarding only syllables that are relatively similar to the target. While previous studies have proposed either short (*23*) or infinite range (*3*) of differences from a target that can affect syllable performance, the actual range is unknown.

We inferred the model parameters β and ς by fitting our actor-critic model to actual learning trajectories from syllable-matching experiments [part of which was previously published (*18*)]. Juvenile males were trained with artificial song tutors to shift the pitch of a syllable by two semitones to match a syllable in a tutor's song (Fig. 2A) or to shift two syllables toward two different targets. In the latter case (five birds), we treated the syllables' learning trajectories as independent of each other (a total of 13 syllables in eight birds). This task was accurately accomplished (Fig. 2B) regardless of whether the pitch was shifted up or down, which is consistent with our assumption that the intrinsic reward distribution is symmetrical. Three syllables (in two birds, fig. S2A) exhibited oscillating trajectories, which RL models cannot account for, and we therefore excluded these syllables from parameter inference. Because we were interested in whether reward functions are qualitatively sub- or super-Gaussian, we evaluated fit quality as a function of β by fixing β to distinct values and inferring ς via maximum likelihood estimation (MLE; $n = 10$ syllables in seven birds; Fig. 2C and fig. S2B; see Materials and Methods).

In 8 of the 10 trajectories on which model parameters were inferred, light-tailed reward distributions (β = 4, 6, or 8) resulted in better fits [i.e., lower Bayesian information criterion (BIC) scores] than normal ($\beta = 2$) and heavy-tailed ($\beta = 1$) distributions ($^*P < 0.05$ and $^{**}P < 0.01$, Benjamini-Hochberg corrected Wilcoxon signed-rank test on non-normalized BICs; Fig. 2D and fig. S2B). We obtained similar results when we estimated β jointly with other model parameters (fig. S3A). BIC scores alone, however, are insensitive to overfitted models that fail to generalize to unseen data. Therefore, we also performed a bootstrap analysis by simulating multiple learning trajectories for each maximum likelihood estimate of β (Fig. 2E and fig. S2B). Light-tailed reward distributions resulted in improved goodness of fit [i.e., smaller root mean square error (RMSE)], between empirical and simulated trajectories in 6 of the 10 cases ($^*P < 0.05$ and $^{**}P < 0.01$, Benjamini-Hochberg corrected Wilcoxon signed-rank test; Fig. 2F and fig. S3B); in 4 of the 10 cases in which the light-tailed model was not the best, the decrease in goodness of fit was small compared to heavy-tailed and normal models (Fig. 2G; thicker lines represent larger improvement by the best model compared to the second best). These results suggest that a light-tailed reward distribution guides the shift in syllable performance toward a match with a target. Namely, a syllable instance triggers a nonzero intrinsic reward only if it is sufficiently similar to a target syllable. This solution is a special case of the multisyllable inventory learning problem that we turn our attention to next.

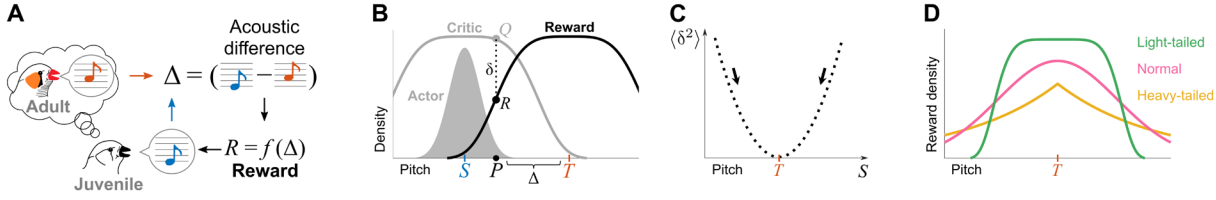

Fig. 1. Vocal imitation of a single syllable. (**A**) A juvenile male zebra finch adjusts the performance of a syllable (blue) to acoustically match its target—an adult "tutor's" syllable (orange). (**B**) An actor-critic model for learning a syllable. (**C**) The motor program mean S is updated to minimize the mean squared RPE δ, which is at its minimum when S equals T (see Eqs. 6 and 7 in Materials and Methods). (**D**) Hypothetical intrinsic reward distributions modeled as a generalized normal distribution.

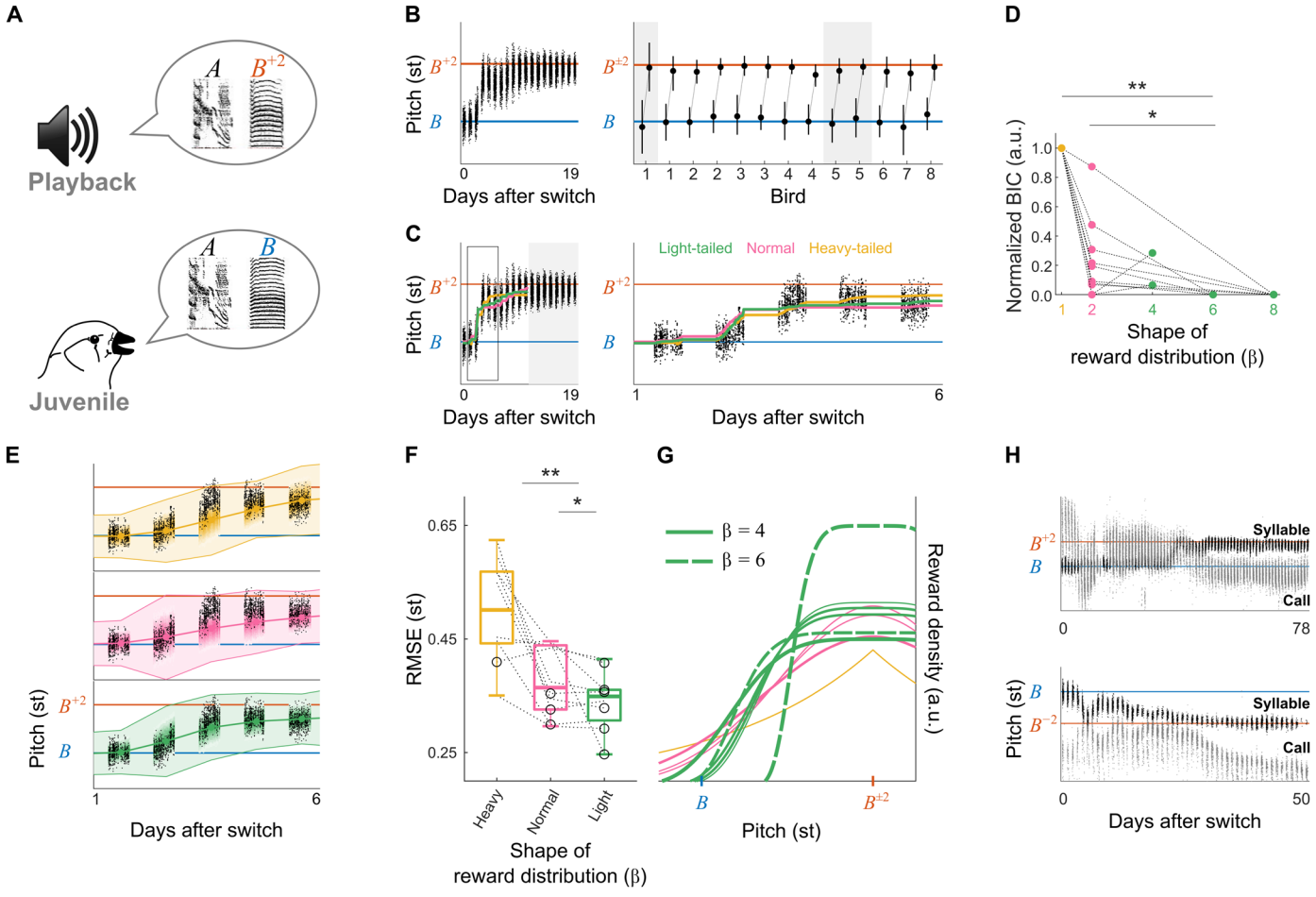

Fig. 2. Vocal imitation of a single syllable best agrees with a light-tailed distribution of intrinsic reward. (**A**) A syllable-matching experiment: A bird that has learned a two-syllable song *AB* is induced to learn a new song, AB^{+2}. The pitch of syllable B^{+2} is two semitones above or below that of *B* (an $AB \rightarrow AB^{+2}$ experiment is shown). (**B**) Left: Learning trajectory of a bird trained with the task in (A), showing the median pitch of consecutive instances of syllable *B* after the presentation of the new tutor song (day 0). Right: Learning outcomes in all birds trained with two-semitone pitch shifting tasks, showing pitch distributions [median ± 95% confidence interval (CI)] at the start point and end point of learning for each shifted syllable. Gray shading denotes three syllables excluded from parameter estimation. (**C**) Left: Inferred mean pitch shifting trajectories fitted with maximum likelihood estimation (MLE) to the learning trajectory of the bird in (B, left), assuming reward distributions with different values of β. Right: Zoom-in on the learning part of the trajectory. (**D**) Bayesian information criterion (BIC) [arbitrary units (a.u.)] for each syllable trajectory, assuming reward distributions with different values of β (dashed lines connect same-syllable data; $n = 10$). BIC scores are normalized such that 1 corresponds to the worst and 0 to the best fit. (**E**) Observed (black) and bootstrapped (color) pitch trajectories of the pitch-shifted syllable shown in (B, left). Shaded area represents median ± 50% CI of simulated trajectories. Color gradient represents bootstrap density. (**F**) 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). (**G**) Best (smallest RMSE) reward distribution for each empirical learning trajectory. (**H**) Two examples of a syllable that shifted to match a target although a call was initially closer. st, semitones.

Imitation of a multisyllable inventory best agrees with light-tailed reward distributions combined in a hierarchical sum-max operation

As a first step in modeling the learning of a multisyllable inventory, we investigated the respective roles of syllables and calls in learning. Birds use both syllables and calls in their vocal repertoire to match syllables in their target song (*17*, *18*, *24*), but previous findings (*18*) suggest that calls are only used in cases where there are more targets than syllables in a bird's song. We reinspected previously published data and found that, when there are an equal number of syllables and targets, birds used syllables to match targets even when there were acoustically closer calls in their repertoire (Fig. 2H). We therefore decided to model inventory learning as a hierarchical process that prioritizes syllables over calls: A call that is similar to a target will be ignored when a less similar syllable is also present within reward range.

Next, we set up to extend our reward model for learning a single target (Fig. 1, B and C) to the parallel learning of multiple targets (taking into account both syllables and calls in the bird's repertoire; Fig. 3A). For that purpose, we used the inferred reward distribution contingent on the pairwise pitch difference between a syllable and a target (Fig. 2) as a functional component (or a basis function); we defined a global inventory reward $\boldsymbol{R}$ as a function of all reward components R_{ij} associated with pairing target i with syllable (or call) j (Fig. 3B). This approach allowed us to generalize the model from single to multiple targets without introducing any new model parameters.

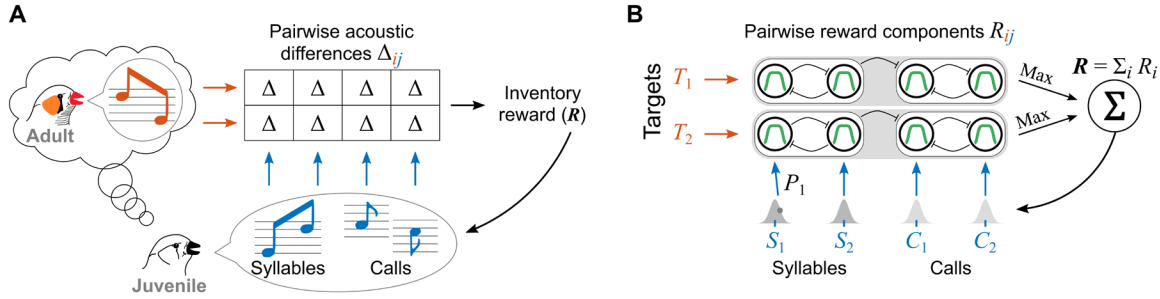

Fig. 3. Vocal imitation of multiple syllables. (**A**) Learning a multisyllable inventory, driven by a putative intrinsic scalar inventory reward ***R***, which depends on the pairwise pitch differences Δ_{ij} between the learner's vocalizations (syllables and calls) and the targets provided by a tutor. (**B**) A hypothesized computation of inventory reward ***R*** from pairwise light-tailed reward components R_{ij} (circles) that are contingent on syllable-target (or call-target) pitch differences. The reward ***R*** is the sum of partial rewards R_i computed for each target T_i via a hierarchical max operation over pairwise reward components. Inhibition (curved lines) among pairwise reward components leads to competition among actors and inhibition from syllables onto calls leads to prioritization of syllables over calls.

The global reward function ***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). Greedy matching of targets by syllables can be achieved by the maximum operation (which ignores all values smaller than the maximum). Accordingly, we constructed the inventory reward ***R*** as a sum-max operation over pairwise reward components R_{ij} (Fig. 3B): Namely, it is a sum $R = \sum_i R_i$ 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 Materials and Methods and fig. S4A). Note that such a two-stage maximum operation can naturally emerge from a light-tailed reward component 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 ***R*** is delivered to a group of actors (each as in Fig. 1B), one actor per vocalization type (syllable or call; Fig. 3B). Actors produce vocalizations independently of each other, and each learns to adjust its mean performance based on the identical reward ***R*** that they receive. We assume that reward ***R*** is delivered after each syllable or call instance (rather than at the end of a song) and that partial rewards R_i are computed with respect to the syllable or call instance just performed P_j and the means S (i.e., motor memories) of the other syllables and calls in the bird's current inventory (see Materials and Methods and fig. S4B). Upon production of an instance, an actor learns from the RPE that is defined as the difference between the inventory reward ***R*** and an expected reward Q computed by this actor's critic (see Eq. 11 in Materials and Methods). By construction, the RPE is zero, and learning stops when each target is matched by an actor's mean.

To test our model, we used data of a published set of multisyllable learning experiments (*18*), where juvenile males were induced to learn two new targets. Both new targets resembled a single syllable in the birds' song (i.e., were pitch-shifted by two semitones up and down from that syllable; Fig. 4A) but were either at a corresponding or at a non-corresponding position in the target motif. Birds predominantly shifted the pitch of the syllable to match one of the new targets, regardless of whether it was at the corresponding serial position in the motif or not and recruited a call to match the other target (7 of $n = 8$ birds; Fig. 4B and fig. S5A). Consistent with a prioritization of syllables over calls, the song syllable shifted toward the target to which it was more similar (fig. S5, A and B, left), while the call shifted to the other target, even when it was more similar to the target matched by the syllable [three of the eight cases; fig. S5, A and B, right; these findings were not published in (*18*)].

Although these data were collected before formally developing the model, the data were not used for parameter estimation (Fig. 2). We therefore used these data to test our model and compare it to alternatives (see below). We compared simulated and empirical learning trajectories (as in Fig. 2E), in terms of both learning end point and trajectory shape. We compared simulated learning trajectories generated by light-tailed, normal, and heavy-tailed variants of the model (with model parameters as inferred in Fig. 2, except for the performance distributions' initial means and daily variances that we estimated in each bird separately; Fig. 4C and fig. S5C, left and middle). The light-tailed model was significantly better at matching the empirical trajectories ($P < 0.01$, Benjamini-Hochberg corrected Wilcoxon signed-rank test; Fig. 4D, left, and fig. S5D), including the replication of a failure of a bird to match one of the targets (Fig. 4, E and F). In that bird, the performance distribution of the song syllable is relatively narrow and, consequently, lies outside the range of light-tailed reward distributions but is within the ranges of normal and heavy-tailed distributions. Therefore, only the light-tailed distribution correctly predicts the failed shifting of the syllable in this bird (Fig. 4F, top).

The light-tailed variant of our MARL sum-max model produces good fits with empirical data, but a sum-max reward is computationally intense, requiring all possible pairwise similarity evaluations at each step. We tested two alternative reward models to see whether simpler reward computations can provide comparable

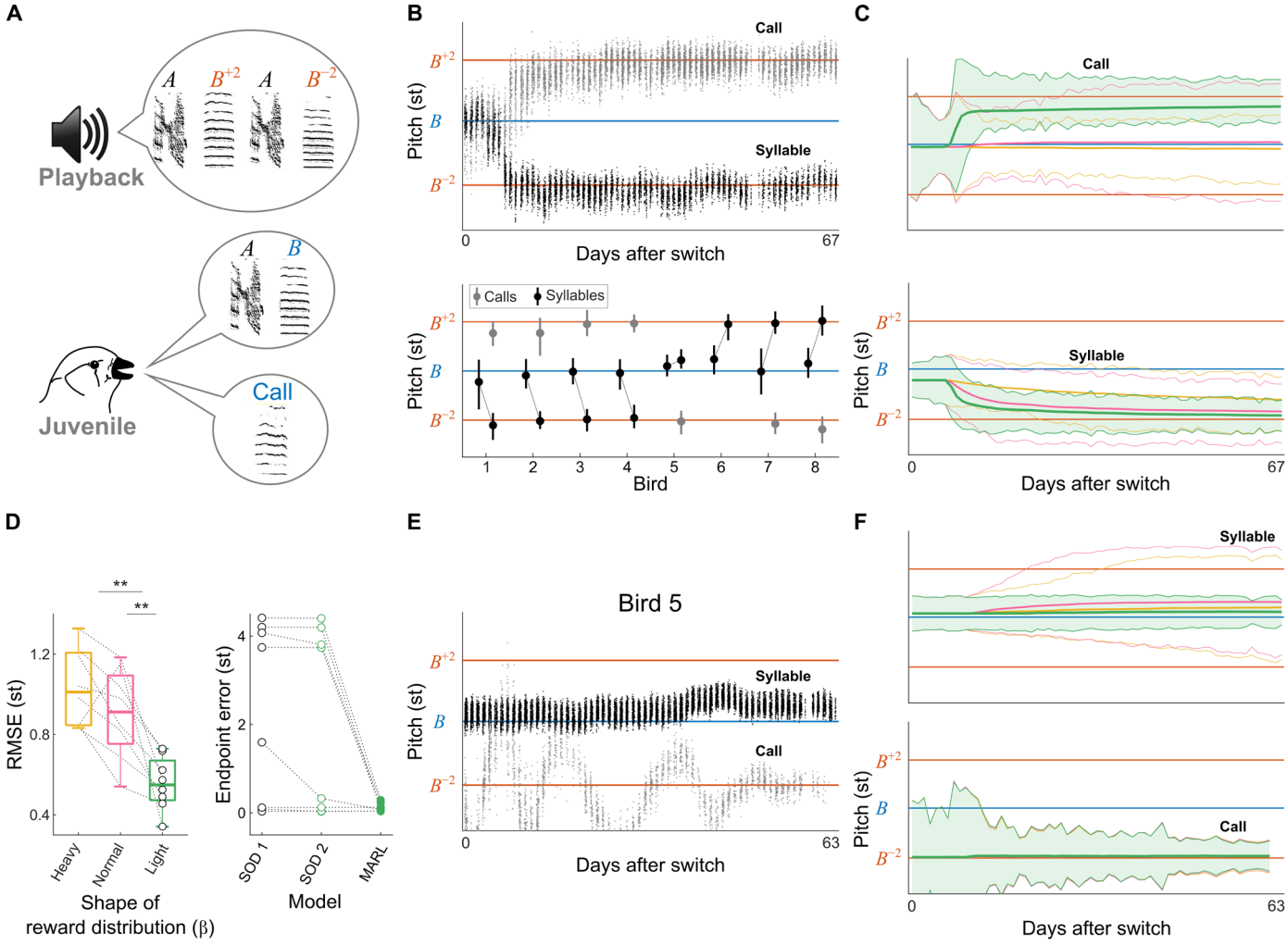

Fig. 4. Imitation of a multisyllable inventory best agrees with light-tailed reward distributions combined in a hierarchical sum-max operation. (**A**) Example of a multisyllable matching experiment, in which a bird that has learned a two-syllable song *AB* is induced to learn a new song with two new target syllables B^{+2} and B^{-2}, with pitch two semitones above and below that of *B*, respectively. The bird's repertoire includes song syllables and calls. (**B**) Top: Pitch trajectory in one bird trained with the task in (A). Bottom: Results across birds ($n = 8$), showing pitch distributions at start point and end point of learning. (**C**) Bootstrapped pitch-trajectory distributions of the call (top) and syllable (bottom) of the bird shown in (B, top). Median ± 50% CI are shown for the light-tailed (shaded), normal and heavy-tailed reward models. (**D**) Left: RMSE between experimental and bootstrapped syllable learning trajectories. Right: Comparison of the end-point simulation error between the MARL sum-max model with light-tailed reward (filled green circles) and two SOD models that follow the song learning approach proposed by Fiete and colleagues (*3*). (**E**) Pitch trajectories of song syllable *B* and call in a bird that failed to match one of the targets [bird 5 in (B), bottom]. (**F**) Same as (C) for the bird shown in (E).

results. First, we tested a sequential-order–dependent (SOD) model, previously proposed by Fiete and colleagues (*3*) (see Materials and Methods). In the SOD model, the reward delivered after each vocalization instance is a function of the similarity between that instance and the temporally aligned target (rather than of all possible pairwise similarities). In addition, target syllables have an infinite range of influence, i.e., they can attract even very dissimilar vocalizations. This is achieved via a binarized reward with an adaptive threshold that guarantees equal numbers of rewarded and unrewarded instances at any pitch distance from a target (*3*). In contrast to the MARL sum-max model, which made the correct (empirically observed) assignment of syllables to targets in eight of the eight cases, the SOD model succeeded in three cases only (Fig. 4D, right, SOD 1 model). The context-dependent evaluation algorithm in the SOD model led to an incorrect assignment of temporally aligned targets in four cases where birds shifted a syllable to a misaligned target, while the binarized reward computation did not predict the observed failure of one bird (Fig. 4E) to shift a syllable to a target. In total, the SOD model resulted in a large simulation error in five of the eight birds, compared to near zero simulation error with the MARL sum-max model. We also tested a variant of the SOD model with a light-tailed reward computation, which correctly predicted the case of failed shifting, but did not predict the matching of misaligned targets, resulting in very large errors in four of the eight birds (Fig. 4D, right, SOD 2 model; see Materials and Methods).

Next, we tested a reward model, which like the MARL sum-max model, is context-independent and competitive, but where, instead of actors competing over targets, the targets are competing over actors [a max-over-targets (MOT) model]. In the MOT model, instead of selecting the most acoustically similar actor for each target, we

select the most acoustically similar target for each actor, so that each actor ends up matching the target that is most similar to it. This model is computationally simpler than the MARL sum-max model, because each step involves only the pairwise comparisons between a given actor and each of the targets, rather than all possible comparisons. However, the MOT model would not ensure that all targets are matched in cases where a target is not closest to any syllable or call. We therefore tested model predictions in two cases where the syllable and the call start points were closer to the same target and were both within the reward range of both targets (birds 1 and 3; fig. S5C, right). We ran simulations of the MOT model using heavy-tailed, normal, and light-tailed reward components as before. In each case, the MOT model predicted that the call and syllable will match the same target, contrary to the empirical data (97 and 55% of simulations with light-tailed MOT model compared to 9 and 5% for the light-tailed MARL sum-max model). In summary, a MARL sum-max reward model with light-tailed components significantly outperformed computationally simpler models in reproducing empirical multisyllable learning trajectories.

Light-tailed sum-max intrinsic reward predicts the exclusion of a syllable from a song and its replacement with a call

The case of a bird failing to match a target (Fig. 4E) raises an interesting question: What happens when there is a target that does not have any syllables in its reward range, and a syllable that is not within the reward range of any target? In such a case, our model predicts that the syllable would not shift from its mean (and in effect be dropped from the learning process) and that a nearby call could shift to the unoccupied target (because there are no syllables in that target's reward range). We reinspected the data of the bird shown in Fig. 4E to search for a call performed in the acoustic vicinity of the unmatched target but found that no such call was present. Nevertheless, the possibility of a song syllable being left out in favor of a call triggered our interest, and we decided to further investigate this prediction of our model.

We presented juvenile male zebra finches ($n = 6$) with a learning task in which one syllable in the target song is shifted up or down by four semitones with respect to a syllable in the bird's current song ($AB \rightarrow AB^{\pm 4}$; Fig. 5A). In this experimental situation, there is a single off-target syllable in the bird's song, B, and a single unmatched target in the tutor song, $B^{\pm 4}$, but the two are separated by a relatively large pitch difference. A hierarchical sum-max reward computation based on light-tailed components predicts that song syllable B is outside the reward range of the target syllable $B^{\pm 4}$ (fig. S6A) and, therefore, that it will not shift to match the target (Fig. 5B, top). Consequently, if the bird's repertoire happens to contain a call C that is within the reward range of the target, then the call will shift to match it instead (Fig. 5B, bottom). In contrast, normal and heavy-tailed versions of the model predict that song syllable B is within the reward range of a four-semitone–shifted target and, therefore, will shift to match it, even if there is an acoustically closer call in the bird's repertoire (Fig. 5B, middle and bottom), because syllables are prioritized over calls.

None of the six experimental birds shifted song syllable B toward target $B^{\pm 4}$ (Fig. 5, C and D, and fig. S6B). In five of the six 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., Fig. 5D, birds 2 and 3). In all five cases, the call was fully or partially incorporated into the song motif (Fig. 5E), 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 (Fig. 5F; $P = 0.0625$, Wilcoxon signed-rank test; cf. fig. S3F). In addition, the performance rate of syllable B decreased significantly ($n = 6$, $P = 0.0312$, Wilcoxon signed-rank test), while call performance increased in three of the five birds (Fig. 5G), with two of the birds stopping performing syllable B altogether (bird 2, Fig. 5D; and bird 5, fig. S6B). We conclude that syllable B was functionally dropped from the song motif, and its function was transferred to a call. In the single case where no call shifted to match target $B^{\pm 4}$, the bird continued performing syllable B at the original pitch but at a lower rate (bird 4, Fig. 5D).

We tested 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 (Fig. 5H and fig. S6C). All simulations used previously estimated parameters (Fig. 2), without further adjustments to the new data except for the performance distributions' initial mean pitches and daily variances that we estimated in each bird separately. The light-tailed model was superior to normal and heavy-tailed model variants, returning smaller errors between simulated and empirical learning trajectories in five of the six birds (#: $0.05 < P < 0.1$, Benjamini-Hochberg corrected Wilcoxon signed-rank test; Fig. 5I and fig. S6D) and replicating the failure of the bird that did not have a call at the acoustic vicinity of the B^{+4} target to accomplish the learning task (bird 4, Fig. 5H). Note that under normal and heavy-tailed models, the syllable and the call can converge to the same target, which is possible when two actors are within the reward range of a single and common target (also see Discussion).

In summary, the experimental outcomes support the predictions of a light-tailed sum-max reward computation with prioritization of syllables over calls. These results demonstrate a unexpected feature of vocal learning in zebra finches, which is a direct consequence of a competitive and light-tailed reward computation: An intrinsic reward is not necessarily contingent on all syllables that a bird is performing but only on those syllables that are sufficiently similar to at least one syllable in the target song. Syllables that are too dissimilar to contribute to the reward computation end up being dropped (fully or partially) from the song motif. These findings underscore a considerable modularity and flexibility of zebra finches' developing songs and raise the interesting question of the adaptive value of such modular vocal development in birds' complex natural social environment.

The light-tailed sum-max model makes testable predictions of dopaminergic neuron tuning in the VTA during learning of a syllable inventory

The intrinsic inventory reward in our model is fed into a computation of the RPE δ, which tracks deviations from expected reward (Fig. 1B and Materials and Methods). Assuming that δ is proportional to the putative firing rate of dopaminergic neurons projecting into the avian song system, our model provides predictions for dopaminergic firing in juvenile birds learning a multisyllable inventory (Fig. 6). We present our model's predictions in a juvenile bird at the onset of learning in each of our three experimental scenarios: learning a single target syllable shifted by two semitones from a syllable in the bird's song (Fig. 2A), learning two different two-semitone–shifted targets (Fig. 4A), and learning a four-semitone–shifted target (Fig. 5A).

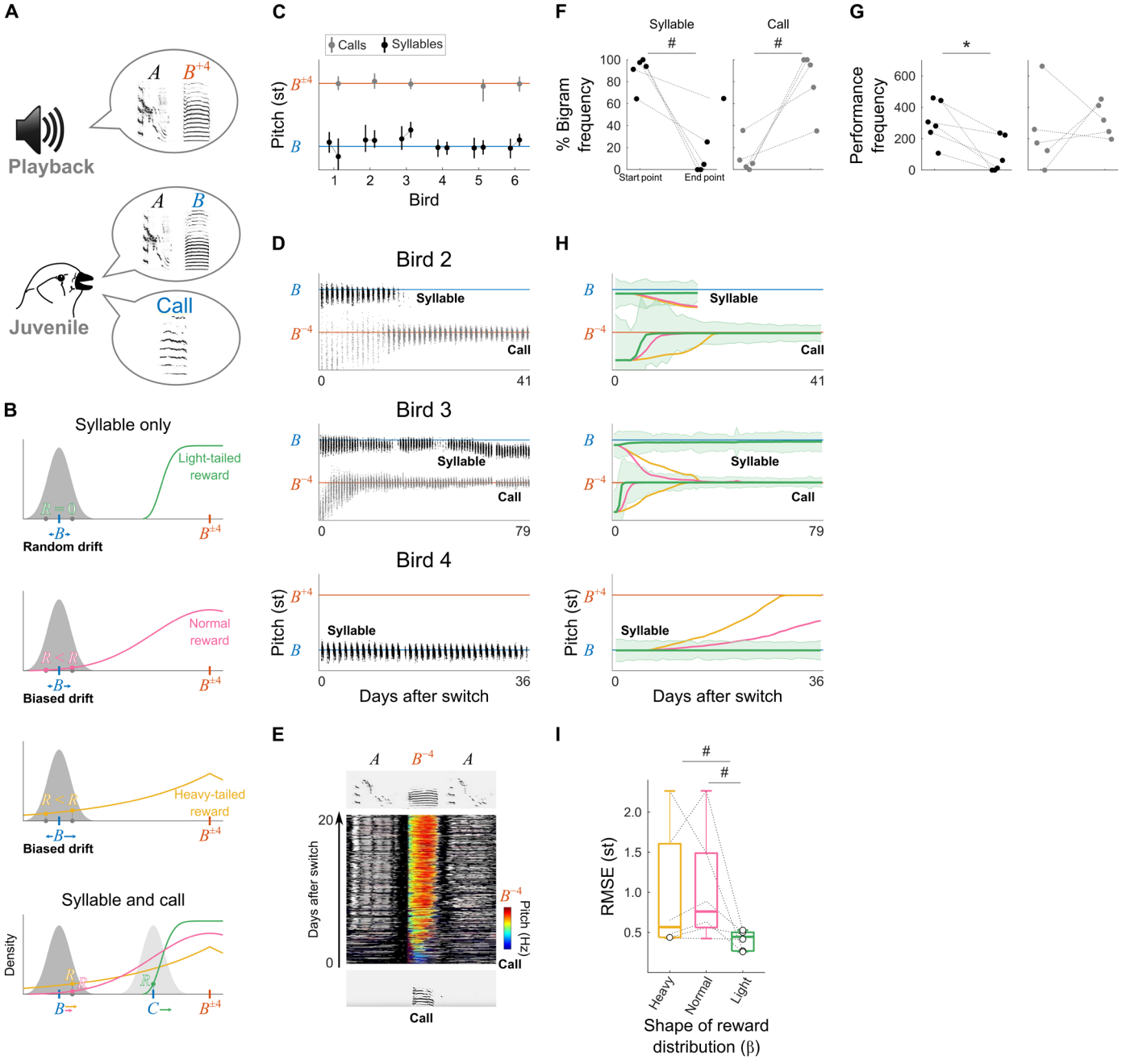

Fig. 5. Light-tailed sum-max intrinsic reward predicts the exclusion of a syllable from a song and its replacement with a call. (**A**) A syllable-matching experiment testing the prediction of a syllable being excluded from the song (an $AB \rightarrow AB^{+4}$ experiment is shown). The bird's repertoire includes song syllables and calls. (**B**) Predictions of light-tailed, normal, and heavy-tailed versions of a MARL sum-max reward model for the experiment in (A). Top and middle: The bird's repertoire only includes syllables (dark gray). Bottom: The bird's repertoire also includes calls (light gray) (**C**) Results across birds ($n = 6$) trained with the task in (A). (**D**) Pitch trajectories of three experimental birds. (**E**) Stack plot of consecutive instances of a call that shifted to match the target B^{-4} in bird 2 in (D). Colors, pitch in call instances. Grayscale, Wiener entropy in neighboring syllables. Sonograms show examples of the call performance at experiment start (bottom) and end (top). The call, initially performed outside the song motif (bottom), was incorporated into the song (top). (**F**) Left: Performance frequency of bigrams containing song syllable *B* and other motif syllables (% of total bigrams) at the start and end point across birds. Right, same as left for the call. (**G**) Performance frequencies of syllable *B* and the call (3-day average of total syllables performed) at the start and end point across birds. (**H**) Bootstrapped pitch-trajectory distributions of the syllable and call (two of the three birds) of the birds shown in (C). The shaded area represents the median ± 50% CI and is shown for the light-tailed reward simulations only. Only the median is shown for the heavy-tailed and normal reward simulations. (**I**) RMSE between experimental and bootstrapped syllable learning trajectories.

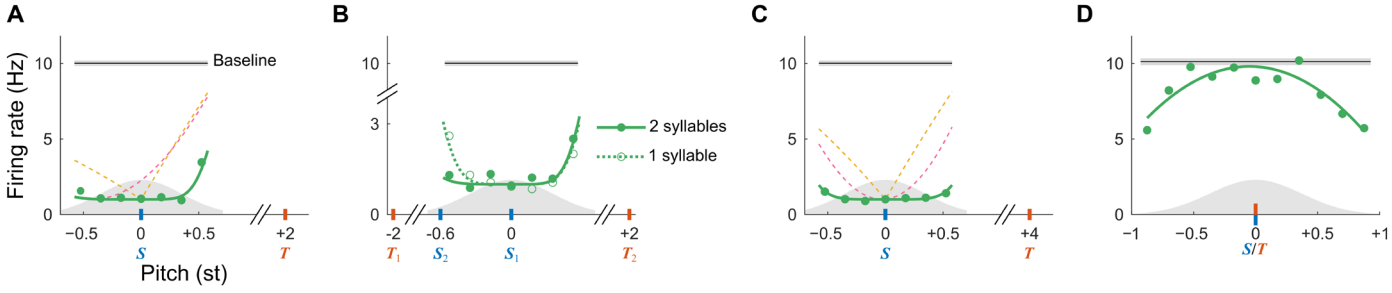

Fig. 6. The light-tailed sum-max model makes testable predictions of dopaminergic neuron tuning in VTA during learning of a syllable inventory. (**A**) Predicted VTA tuning as a function of pitch in the case of a target shifted by two semitones with respect to a song syllable. Green, theoretical (solid 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 (magenta). (**B**) Predicted VTA tuning in the case of two targets and light-tailed reward. (**C**) Same as (A) for a four-semitone–shifted target. VTA neuron simulations in (A) to (C) are based on 2000 pitch values drawn from a normal performance distribution (gray shading) with a typical SD at the onset of learning (σ_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 means ± SEM. (**D**) Predicted VTA tuning when a syllable matches a target, but the mean syllable pitch continues to drift slowly around the target due to circadian variations. The simulated VTA tuning (circles) predicted by a light-tailed reward is well fitted by a parabola (green line). Absence of pitch drift would result in a flat tuning (gray line). Simulation details as in (A) to (C).

In the case of learning a single two-semitone–shifted target (Fig. 6A), the light-tailed reward model predicts selective tuning of ventral tegmental area (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. However, 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 pitch difference Δ 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 (*4*, *25*, *26*), 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 Δ.

In the case of learning two different two-semitone–shifted targets T_1 and T_2 (Fig. 6B), VTA tuning is expected to depend not only on the performance distribution of the now sung syllable but also on the distributions of other syllables (or calls) in a bird's repertoire: When a single syllable S_1 is at a distance of two semitones from both targets, 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 S_2 proximal to one target, e.g., T_1, would break this symmetry, rendering the VTA neuron more selective to the other target T_2 by firing more following instances of S_1 in the vicinity of T_2 than instances in the vicinity of T_1.

In the case of a four-semitone target (Fig. 6C), the light-tailed reward model (in contrast to the heavy-tailed and normal reward models) predicts lack of target selectivity due to the reward being practically zero. This would lead to the syllable pitch drifting randomly around its mean (Fig. 5, B and D). Note that the zero-reward implies that the RPE $\delta \approx -Q$, which means that the firing rate increases slightly with increased distance from the mean.

Last, when a syllable matches a target (Fig. 6D), i.e., in adult birds that have learned their song, the model predicts a flat pitch tuning curve due to absence of RPE. In reality, however, small pitch fluctuations (*27*) are predicted to lead to the emergence of inverted-U tuning because the intrinsic reward decreases away from the target.

In summary, our model predicts that the syllable-related (or call-related) tuning of VTA neurons depends dynamically on the pitch difference between the vocalization and the target it is closest to. This leads to both asymmetric (Fig. 6, A and B) and symmetric (Fig. 6, B to D) tuning curves in dopaminergic neurons, and the symmetric curves can be either concave (Fig. 6C) or convex (Fig. 6D). Asymmetric and inverted-U tuning curves as in Fig. 6 (A and D) have been reported in adult birds (*21*), but our predictions on tuning in juveniles including the dependence on multiple proximal syllables and calls (Fig. 6B) remain to be explored.

DISCUSSION

Our work shows that an inventory of actions underlying a combinatorial behavior—birdsong—can be learned by a set of independent RL actors in a self-reinforced manner from an intrinsic scalar reward. By fitting alternative RL models to empirical developmental trajectories, we were able to infer and test the properties of this intrinsic reward, rejecting SOD models in favor of a greedy and context-independent model with light-tailed reward components. We tested the prediction that such learning algorithm can lead to the dropping of song syllables and the recruitment of calls instead. The model makes testable and detailed predictions for the activity patterns of dopaminergic neurons in the songbird brain during learning (Fig. 7A).

Our model assumes independence on three separate levels: among actors in the action system, among critics in the value system (see Materials and Methods), and among reward components in the reward system. We believe that the independence and modularity among actors, critics, and partial rewards promotes the expansiveness and flexibility of inventory learning, making it easily scalable. 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

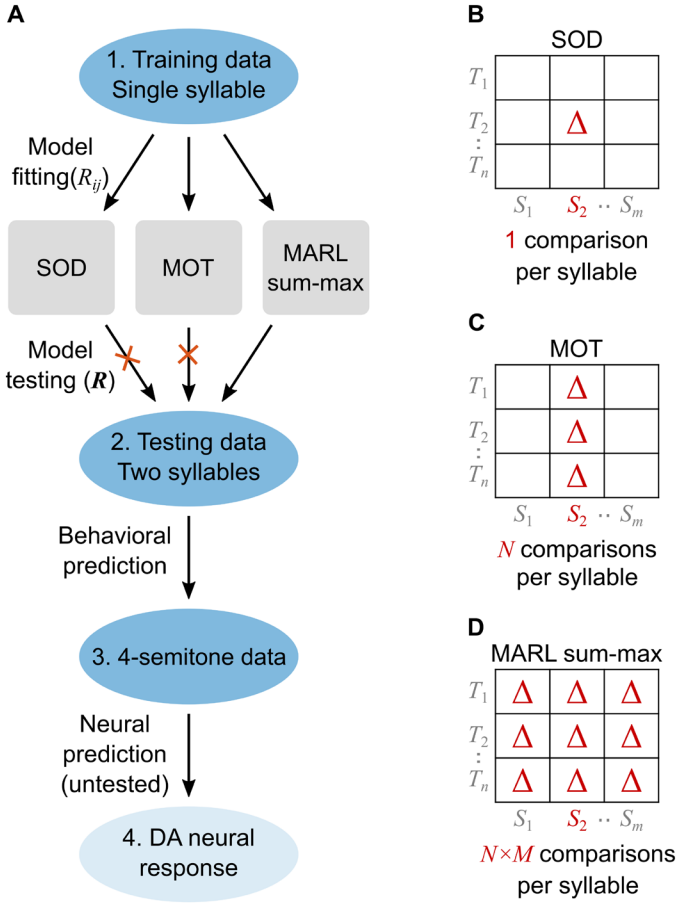

Fig. 7. A data-driven computational approach to model inference and testing. (**A**) General workflow of the study. Step 1: Fitting a single-actor RL model to data of birds that learn a single syllable [training data; Fig. 2; part of the data is from (*18*)]. We identify possible shapes and widths of R_{ij}. Step 2: Testing alternative MARL models on data of birds that learn two syllables [testing data; Fig. 4, data from (*18*)]. We use the fitted pairwise rewards R_{ij} from step 1 as components of a global inventory reward $\boldsymbol{R}$ without further parameter adjustments. Alternative models differ in the computation of $\boldsymbol{R}$ from R_{ij} components. Only the MARL sum-max model with light-tailed reward components successfully reproduces pitch trajectories in the testing data. Step 3: Testing a prediction of the light-tailed MARL sum-max model in a new behavioral experiment (four-semitone data; Fig. 5). The results confirm the prediction that birds will exclude a syllable from their song, if it is acoustically distant from all targets. Step 4: Formulating predictions for future studies of the light-tailed MARL sum-max model of the properties of DA neural responses during juvenile song learning (Fig. 6). (**B** to **D**) Alternative models of syllable inventory learning [step 2 in (A)] differ in computational complexity. Complexity is measured as the number of pairwise acoustic comparisons needed to compute the intrinsic reward after an uttered syllable (red; the same applies to uttered calls, which are not shown). The SOD model (B) is 0-order, requiring only one comparison per syllable, with the target at the same serial position. The MOT model (C) is linear, requiring comparisons between the uttered syllable and all targets. Last, the MARL sum-max model (D) is quadratic, requiring all possible comparisons between syllables and targets for each uttered syllable.

computation. The flexible goals that 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 (*28*). Moreover, our findings raise the possibility that the similarity between sung syllables and syllables a juvenile bird hears from others might be one of the factors that determine which tutors a juvenile will learn from. For example, a syllable that is dissimilar to any syllable of one tutor (which in our study birds tended to perform less or to drop altogether; Fig. 5) might in a more natural social environment shift toward a target in another tutor's song.

Computationally, however, a greedy and context-independent sum-max reward seems dauntingly complex, especially in comparison to previous actor-critic models (*2*, *3*), which assumed that syllables are learned within the sequential context of the song motif rather than independently of it [but see (*29*)]. In principle, compared to the simpler SOD and MOT models, a MARL sum-max reward circuit (Fig. 3B) needs to perform all possible pairwise comparisons at each time step (Fig. 7, B to D), a total of $N \times M$ maximum operations per song motif, each over a set of N terms (for M targets and N syllables/calls). Although this seems like a daunting number for large inventories, the computation is quite manageable in practice, because, among these $N^2 \times M$ terms, only at most $2N \times M$ are distinct—the $N \times M$ terms comparing the targets with now sung syllable instances and the $N \times M$ terms comparing the targets with the means S_j (and C_j) of the motor programs for syllables and calls not now sung. Furthermore, because the pairwise comparisons are subserved by light-tailed reward distributions, 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. Thus, light-tailed reward distributions may promote efficiency.

In terms of neural architecture, a sum-max reward circuit can be implemented by a matrix-like network of $N \times M$ pairwise comparator neurons (each neuron corresponding to one pairwise reward component; Fig. 3B). These neurons would have light tailed tuning curves, meaning that they would be tuned to small mismatches between a syllable and a target but would be unresponsive to large mismatches. Biophysically, each pairwise comparator neuron would represent information about a target syllable T_i and the mean of a motor program S_j (or C_j) either in terms of its neural activity, the synaptic inputs that it receives, or the strength of its synapses. In addition, comparators would receive auditory input P_j corresponding to feedback from a syllable instance just sung, which would replace the representation of the syllable's mean performance S_j. To implement the sum-max computation, the comparator neurons could be arranged in parallel winner-take-all modules, each of which computes a max operation with respect to a target via inhibitory synapses, followed by an output summation to obtain the inventory reward $\boldsymbol{R}$ (Fig. 3B). All these computations are feasible with neural networks.

Where would this putative sum-max reward network, and the actors and critics that learn from it, reside in the songbird brain (Fig. 8)? The comparator network computing the inventory reward $\boldsymbol{R}$ and the critics which compute the expected reward Q for each actor must reside upstream of VTA where the difference between these two signals (the RPE δ) is computed (*4*). Candidate sites for the inventory reward network are the ventral pallidum (VP) and its afferents, given that VP input to VTA is of positive valence (*30*) and therefore acts like the inventory reward $\boldsymbol{R}$ on δ [although another study found mixed signals in VP (*31*)]. The critics would be located in the ventral intermediate arcopallium (AIV) and its afferents, given that AIV input to VTA (*32*) is of negative valence (*30*) and so

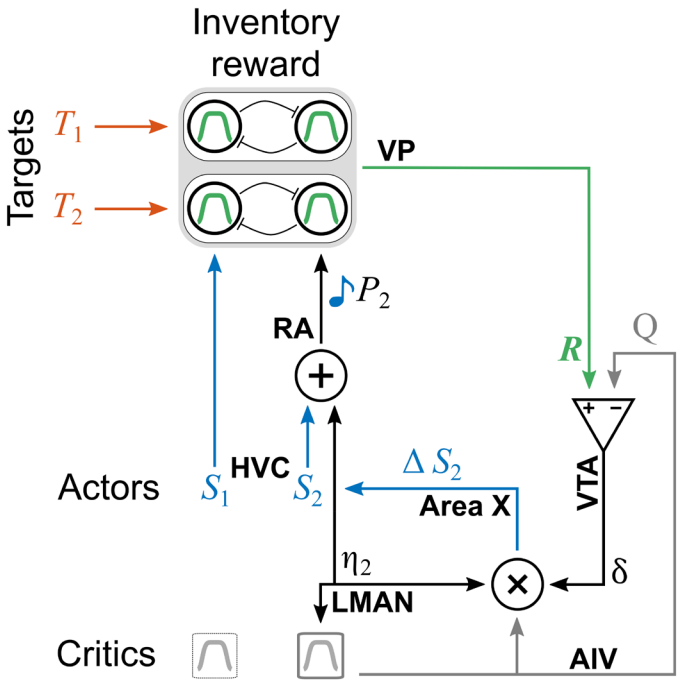

Fig. 8. Potential mapping of the MARL model for inventory learning onto the songbird brain. Signals in our model (*S*, *P*, η, ***R***, *Q*, and δ) are depicted as arrows and are labelled by the hypothesized brain areas (bold) where such signals can be found (implying the signals are computed in that area or upstream).

acts like *Q* on δ. Both the VP and the AIV should in such a case receive auditory feedback from the syllable or call instance just sung P_j, because this input is necessary for the computation of both ***R*** and *Q*. This assumption would agree with observations that neurons in VP and AIV are highly sensitive to distorted auditory feedback during singing (*30*, *31*).

Because the premotor nucleus HVC (proper name) controls the learned components of song, the lateral magnocellular nucleus of the anterior nidopallium (LMAN) is involved in vocal exploration, and both project to the robust nucleus of the arcopallium (RA), we suggest that the actors are distributed across HVC, LMAN, and RA. HVC would generate *S*, i.e., stable motor memories of each syllable or call (*33*). The individual syllable/call instances *P* performed by RA would be the sum of two components: *S* (or *C* for calls) plus a more variable component η from LMAN (*34*). This latter signal would be relayed as an efference copy to the dopaminergic recipient area X where the learning itself could take place, in agreement both with a previous hypothesis (*34*) and corroborating anatomical evidence (*35*). The update Δ*S* computed in area X would be consolidated in RA and HVC in a process that we do not explicitly model (*36*).

The experimental paradigm and the model presented here contain some simplifications that could be broadened in future studies: First, we studied pitch, mainly because of our ability to selectively manipulate this acoustic feature. It remains to be seen how this approach generalizes to other sound features. Presumably, the pairwise reward components, as well as the critics, would need to be tuned to acoustic features other than pitch and possibly compute a weighted sum across these features. It may also be that the cases of oscillating pitch trajectories we observed (fig. S2A and 4E) stem from the integration of acoustic dimensions other than pitch in birds' similarity evaluations. Second, we made the simplifying assumption that the reward components and the critics have an identically shaped tuning to pitch differences. This convenient simplification may be relaxed in a more powerful version of the model, in which critics acquire their tuning curves via function approximation methods. Third, we now treat syllables as the basic, or smallest, units of learning, but zebra finches can selectively adjust individual sub-syllabic elements toward increased target match (*24*). This calls for tutoring experiments targeting the process of learning sub-syllabic notes, and a corresponding update of an intrinsic reward model to include the evaluation of sub-syllabic structure. Fourth, despite the maximum operation we used, the greediness our model implements is not strict (as in a soft winner-take-all) and two actors can in principle converge to the same target when no other target is around (for example, Fig. 5H, middle). This possibility in our model arises from the stepwise updates Δ*S* of performance means, when two roughly equidistant means are pulled to the same target in an alternating manner. Although a light-tailed model makes this scenario rare by limiting the pitch range in which it can occur, such violation of greediness remains to be tested in future experiments. Last, birdsong learning involves both imitation and creativity (*37*–*40*), which raises the interesting question of the possible contribution of an intrinsic reward to birds' creative variations. This question needs to be experimentally studied in more complex and naturalistic social environments, where creative variations should be more apparent (*39*, *40*) than in our simplified experimental setting.

Our findings demonstrate that competition and cooperation between independent RL actors maximizing a common reward is essential for the learning of a syllable inventory in birdsong. A dedicated learning of an action inventory is crucial for combinatorial skills that rely on a small set of basic actions to flexibly generate diverse sequences. The learning algorithm that we discovered in songbirds may therefore have evolved in other learned combinatorial behaviors, including the learning of speech sound inventories of human languages.

MATERIALS AND METHODS

Experimental design

Animal care and experimental procedures were conducted in accordance with the guidelines of the US National Institutes of Health and have been reviewed and approved by the Institutional Animal Care and Use Committee of Hunter College.

Part of the experimental data presented here (all two-semitone learning experiments except for two new birds) was previously published (*18*) and used here to infer and test model parameters. Male zebra finches were bred at Hunter College and reared in the absence of adult males between days 7 and 30 after hatch. Afterward, birds were kept singly in sound attenuation chambers and continuously recorded. From days 33 to 39 until day 43, birds were passively exposed to 20 playbacks per day of a tutor song (source), occurring at random with 0.005 probability per second. On day 43, each bird was trained to press a key to hear song playbacks, with a daily quota of 20. Once birds learned the source song, we switched to playbacks of a different tutor song (target). Learning of the source was assessed by quantifying the percent similarity [Sound Analysis Pro (*41*)]

between the bird's song motifs and the source model motif in 10 randomly chosen song bouts per day. We considered the source song as being learned when the similarity to the model was at least 70%. Because the sensitive period for song learning in zebra finches ends around days 90 to 100 after hatch, we included only birds that learned the source before day 71 in the experiments. Recording and training were done using Sound Analysis Pro (*41*) and continued until birds reached adulthood (days 99 to 158 after hatch). At these ages, males are sexually mature and perform a crystalized song motif that remains unchanged for the remainder of their lives (*42*). Throughout the experimental period, birds did not receive any external stimuli contingent on their singing (aside from being able to hear themselves sing). Identical experimental conditions were used to generate previously published data used in this study [see details (*18*)].

Source and target song models were synthetically composed of natural syllables. Harmonic syllables in the source songs were pitch-shifted by two or four semitones in the target songs using GOLDWAVE (v. 5.68, www.goldwave.com). Each playback of a model included two motif renditions. To control for model-specific effects, we varied baseline pitch and pitch shift direction across experimental birds. Because source and target syllables differed by either two or four semitones (less than half an octave), birds' pitch shifting trajectories should not be affected by possible nonlinearities in pitch perception due to octave equivalence.

Data analysis

We performed song feature calculation and clustering of syllables and calls using Sound Analysis Pro (*41*) [see (*18*)]. All other analyses were performed in MATLAB (Mathworks Inc.). Unless mentioned otherwise, significance levels were adjusted using the Benjamini-Hochberg procedure (*43*) to correct for multiple comparisons.

Calculating median pitch of syllable instances

We used the following procedure to prepare pitch traces for statistical analysis and model fitting. Following syllable clustering, median pitch values of syllable instances P were converted from hertz to semitones (st) using the formula

$$P[\mathrm{st}] = 12\log_2 \frac{P\left[\mathrm{hertz}\right]}{T^0\left[\mathrm{hertz}\right]'} \tag{1}$$

where T^0 is the source pitch (i.e., the pitch of the syllable in the source song that we manipulated in the target song). The semitone scale allows us to express our results in a common reference frame in which the source pitch is zero. To minimize biasing effects of clustering errors on model fits, we excluded renditions with pitch more than three median absolute deviations from the daily median pitch (leading to elimination of less than 0.2% of data points).

Characterizing daily pitch distributions

We first restricted the distribution analysis to stable days: We assessed pitch stability across consecutive day pairs (i.e., days with no learning or no random drift) by comparing medians of pitch distributions (uncorrected Mann-Whitney U test; $\alpha = 0.001$). On the set of stable day pairs, we found pitch distribution to be nearly always Gaussian: The daily distributions (with more than 10 instances) of stable pitch were normally distributed on 730/778 days = 94% (uncorrected Kolmogorov-Smirnov test on standardized distributions; $P > 0.05$). The same was true when we tested on all days, not just the stable ones, after day-wise detrending (uncorrected Kolmogorov-Smirnov test on standardized distributions $P > 0.05$ in 1214 of 1276 days = 95%). This justifies our choice of the normal performance distribution for the actors in our RL model (see below).

RL model

Our RL model consists of three components: actors that are responsible for generating vocalization (syllable and call) instances, an intrinsic reward $\boldsymbol{R}$, and a critic that estimate the expected reward Q of an instance. We first describe the model of a single actor-critic pair suitable for learning a single target syllable.

Single-actor model

At time t, the actor generates a syllable/call instance with pitch $P(t) = S(t) + \eta(t)$, where η is drawn from a normal distribution

$$\eta(t) \sim \mathcal{N}\left[0, \sigma_B^2(t)\right] \tag{2}$$

$S(t)$ is the mean performed pitch, and $\sigma_B(t)$ is the (time-dependent) SD that sets the extent of behavioral variability. The syllable instance is followed by an intrinsically generated reward signal $R(t) := R[P(t), T]$ that depends on both the produced pitch $P(t)$ and the target pitch T. We assume that $R(t)$ is symmetric around T and is a decreasing function of the absolute pitch difference $\Delta(t) := |P(t)-T|$. Because the functional form of $R(t)$ is a priori unknown, we introduce a family of hypothetical intrinsic reward functions, i.e., we model reward as a generalized normal distribution (*22*)

$$R\left[P(t), T\right] = \frac{\beta}{2\varsigma\Gamma(1/\beta)} e^{-\left[\frac{|P(t)-T|}{\varsigma}\right]^{\beta}} \tag{3}$$

The parameter β sets the shape of the reward distribution around the target: Larger values of β correspond to distributions with lighter tails with the reward density concentrating more around the mean. Special cases of this family of distributions are normal and Laplace distributions given by $\beta = 2$ and $\beta = 1$, respectively. The positive scale parameter ς controls the width of the reward distribution. The reward SD σ_R depends on both β and ς, $\sigma_R = \sqrt{\Gamma(3/\beta)/\Gamma(1/\beta)}\varsigma$.

Learning is driven by the bird's attempt to maximize reward via minimizing the square discrepancy $\delta^2(t)$ between actual reward $R(t)$ and expected reward $E[R(t)|P(t)]$

$$\delta(t) := R(t) - E\left[R(t)|P(t)\right] \tag{4}$$

The variable $\delta(t)$ is also referred to as the RPE, which is commonly assumed to be related to the firing rate of dopaminergic neurons. To estimate the expected reward $E[R(t)|P(t)]$ associated with a vocalization $P(t)$, we introduce a parametric function $Q(t) := E[R(t)|P(t)]$, referred to as the critic. We give the critic the same functional form as the reward distribution $R(t)$, but we center its peak on the average pitch $S(t)$ instead of the target T

$$Q\left[P(t); S(t), \beta, \varsigma\right] = \frac{\beta}{2\varsigma\Gamma(1/\beta)} e^{-\left[\frac{|\eta(t)|}{\varsigma}\right]^{\beta}} \tag{5}$$

where $\eta(t) = P(t) - S(t)$ as given by Eq. 2. Our assumption that R and Q have the same functional form is plausible because presumably both R and Q are computed by neural circuits that have co-evolved to support song learning.

We derive the learning rule by minimizing the mean square error $0.5\langle\delta^2(t)\rangle_t$ using stochastic gradient descent, which is the conventional approach to continuous-action RL (*44*)

$$\begin{aligned} S(t+1) &\leftarrow S(t) - \alpha\frac{\partial}{\partial S(t)}\langle\delta^2(t)\rangle_t \\ &\leftarrow S(t) - \alpha\delta(t)\frac{\partial\delta(t)}{\partial S(t)} \\ &\leftarrow S(t) + \alpha\delta(t)\frac{\partial Q(t)}{\partial S(t)} \end{aligned} \tag{6}$$

where $\alpha > 0$ is the learning rate.

Inserting Eq. 5 into Eq. 6 leads to the following Rescorla-Wagner–type iterative update rule for the mean pitch $S(t)$

$$S(t+1) \leftarrow S(t) + \frac{\alpha\beta}{\varsigma^\beta}\delta(t)Q(t)\eta(t)|\eta(t)|^{\beta-2} \tag{7}$$

According to this rule, because $\eta(t) = P(t) - S(t)$, updates of $S(t)$ mostly point in the opposite direction of $P(t)$ (δ is rarely positive) and are proportional in magnitude to the RPE $\delta(t)$, which biases the drift of S toward T. During learning, as the RPE converges toward zero, Q approaches R, and the syllable's mean pitch $S(t)$ approaches the target pitch T.

It is important to note that, although, by construction, the goal of our model is to minimize the square RPE (δ model), an RPE is not needed in principle. Instead, we could have chosen the goal of directly maximizing the intrinsic reward [as in policy-gradient learning (*45*, *46*)]. Such a model would be simpler, as it would not require a critic for computing an expected reward, which would free up resources. Simulations of a policy gradient (or R-learning) algorithm that maximizes reward directly (rather than minimizing RPE) and that uses no critic at all showed equally good fits of the single syllable data and thus provide comparable behavioral results to our actor-critic model (fig. S3C). Under an R-learning model, assuming that dopaminergic responses in VTA encode R instead of δ, 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 (i.e., response tuning curves with the same shape as a light-tailed reward component).

However, the rather low average song-related firing rate (13 Hz) of area X–projecting VTA (VTA_X) neurons observed in adults (*47*) does not strongly support this view, suggesting instead that dopaminergic neurons encode a difference signal, as predicted by a δ model. Neither is the tendency of VTA neurons to display inverted-U tuning curves centered on a single auditory feature (*21*) in unequivocally strong support for the R-learning model; we found a similar behavior under the δ model when we let the mean pitch of an actor slowly drifts around its target, which also led to inverted-U tuning (Fig. 6D). In summary, it is now not possible to unequivocally arbitrate between the δ and R-learning models, the R-learning model may be simpler but the δ model may map better onto known firing properties.

Multiple-actor model

Next, we consider the case of many actors that learn to produce a multisyllable inventory. We introduce an independent actor for each of the juvenile's vocalization types. Actor j (j = 1, …, N) randomly produces vocalizations (a syllable or a call) drawn from a normal distribution with mean pitch S_j and SD $\sigma_{B,j}(t)$. The N actors compete to fill M targets of pitch T_i (i = 1, …, M). The actors learn from a common scalar inventory reward $\boldsymbol{R}$, which agrees with RPE being signaled by a single neuromodulator, i.e., dopamine [see the "Motivation for greedy (order-independent) learning of a syllable inventory" section].

Our model must reflect the observation that birds make the minimal changes necessary to match the target inventory, i.e., vocalization assignments are greedy and independent of sequential context. To satisfy the latter constraint of independence, we write the inventory reward $\boldsymbol{R}$ as a sum (which is insensitive to permutations) of partial rewards

$$\boldsymbol{R}(t) := \frac{1}{M}\sum_{i=1}^{M} R_i(t) \tag{8}$$

The partial reward $R_i(t)$ associated with target i must reflect the maximum across individual actor-target pairwise comparisons. Naively, to implement a competition between actors over a target, we could choose a simple maximum operation across all actors, because only one actor can score the partial reward associated with a given target, namely, the one that produced the acoustically closest instance among all vocalization types in the inventory [we could define $R_i(t) = \max_j R_{ij}(t_j)$, where $R_{ij}(t_j)$ is the partial reward associated with target T_i and $t_j \le t$ is the last time at which actor j produced a vocalization]. However, to account for the prioritization of syllables over calls illustrated in Fig. 2H and fig. S5B, we define the partial reward R_i for target i as the syllable-specific partial reward

$$R_i^S(t) := \max_{j\,\in\,\text{Sylls}} R_{ij}(t)$$

if the latter is positive (i.e., larger than some infinitesimal value ϵ), else we define it as the call-specific partial reward

$$R_i^C(t) := \max_{j\,\in\,\text{Calls}} R_{ij}(t)$$

as follows

$$R_i(t) := \begin{cases} R_i^S(t) \;;\; \text{if } R_i^S(t) > \epsilon \\ R_i^C(t) \;;\; \text{otherwise} \end{cases} \tag{9}$$

where $\epsilon = \max(R_{ij})/1000$ is a small number (see fig. S4A). According to this definition of partial reward, a distant syllable within the reward range of a target will win over a closer call, replicating the finding in Fig. 2H and fig. S5B. Note that the infinitesimal parameter ϵ is too small to affect the actual reward range, regardless of its shape.

We are left with needing to define $R_{ij}(t)$, the pairwise reward components, and relate them to the estimated reward density $R(P; T)$ of the single-actor model from the previous section. The bottleneck to consider is that we can compute the maximum operations in Eq. 9 only after all arguments are known. To avoid the need of having to wait until all calls and syllables in the inventory have been produced, e.g., at the end of a song motif, we formulate a model in which rewards and error signals are computed upon production of every syllable and call instance with minimal requirement on short-term memory (see the "Motivation for instantaneous reward" section).

We assume that the reward components $R_{ij}(t)$ associated with the ith target and the jth actor depend on the vocalization instance $P_j(t)$ just performed and the syllable motor means $S_j(t)$ [$C_j(t)$ for calls] of non-performed vocalizations

$$R_{ij}(t)=\begin{cases} R\left[P_j(t);T_i\right] & ; \quad \text{if syllable } j \text{ was performed at time } t \\ R\left[S_j(t);T_i\right] & ; \quad \text{if syllable } j \text{ was performed at time} \in [t-\tau,t) \\ 0 & ; \quad \text{otherwise} \end{cases} \tag{10}$$

Only syllables and calls that were recently performed contribute to $\boldsymbol{R}$, which is encapsulated by the time window τ = 1 min, corresponding roughly to the average duration of a singing-and-calling period (see the distributions of call-song intervals in figs. S4C and S6E).

Note that context-independent inventory learning as we propose it here is also offered in the model of Troyer and Doupe (*48*), 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 their model is far from being scalar in nature as the dopamine signal that we model.

Critics

For each actor j, we introduce a critic that tracks the expected reward $Q_j(t) \coloneqq E[R(t)|P_j(t)]$ associated with each syllable/call instance performed by that actor. The quality function $Q_j(t) = Q[P_j(t); S_j(t), \beta, \varsigma]$ of the critic takes the same form as in Eq. 5. The parameters β and ς are common to all critics, but each critic has a distinct mean pitch parameter $S_j(t)$ [$C_j(t)$ for calls] that it inherits from the actor it is paired with. When actor j produces a vocalization, only critic j produces a nonzero expected reward. All other critics produce an output of zero: $Q_{k\neq j}(t) = 0$. It follows that the total expected reward $Q(t)$ of a vocalization is simply the summed output of all critics, $Q(t) = \sum_k Q_k(t)$.

Gradient descent learning rule

Given the intrinsic reward $\boldsymbol{R}(t)$ in Eq. 8 and the expected reward $Q(t)$ as above, we define the RPE again as $\delta(t) = R(t) - Q(t)$ and minimize the mean square of that error, $0.5\langle\delta^2(t)\rangle_t$, in a stepwise manner using stochastic gradient descent. We obtain an iterative update for the mean S_j of an actor that is analogous to Eq. 7

$$S_j(t+1) \leftarrow S_j(t) + \frac{\alpha\beta}{\varsigma^\beta}\delta(t)Q_j(t)\eta_j(t)|\eta_j(t)|^{\beta-2} \tag{11}$$

Because of the term $Q_j(t)$, we see that only the mean of the actor that generated a vocalization is updated, whereas all other actors are left unchanged.

Equation 11 is in essence a Hebbian-like learning rule that multiplicatively combines three terms: an RPE δ that is common to all actors and evaluates whether the performance is better or worse than expected, a critic Q_j that restricts the update to the actor responsible for producing the vocalization, and an efference copy $\eta_j = P_j - S_j$ that provides information about whether the pitch of the current vocal instance was above or below its mean.

Such a triplet learning rule as in Eq. 11 has been predicted to exist in the basal ganglia of songbirds, and recent data provide anatomical support for its existence (*35*). According to this line of work, there is convergence of three types of signals onto medium spiny neurons in area X of the songbird basal ganglia: an error signal (like δ) that stems from VTA, a timing signal (like Q_j) from HVC that restricts learning to the right time in the song, and an efference copy signal (like η_j) from LMAN that provides information about the current exploration (see Fig. 8). These signals emerge naturally from our multi-actor model trained to minimize square RPE, providing anatomical support for our approach.

Illustration of dynamic assignment

The competitive-cooperative reward computation in Eqs. 8 and 9 entails dynamic assignments between vocalizations and targets, which we illustrate with a specific example. When multiple syllables are in the "reward range" of more than one target (fig. S4B, top), syllable-target assignments may initially vary from instance to instance, sometimes leading to an update of a motor program's mean toward one target and sometimes toward another. However, even small asymmetries in the reward magnitudes across different syllable-target pairings will lead to cooperative convergence of assignments that maximize inventory reward (fig. S4B, middle) and eventually to all targets being matched (fig. S4B, bottom). This dynamic process is reminiscent of a musical chairs game, in which pairings between players and chairs are not predetermined but are resolved in a competitive-cooperative manner, eventually resulting in each chair being occupied by a single player.

Model inference and evaluation

We infer model parameters from experimental birds induced to shift the pitch of a syllable toward a target that differs by two semitones upward or downward (Fig. 2 and Eq. 7). We included experimental birds that were trained to shift the pitch of two different syllables toward two different targets separated by at least six semitones (birds shifted each of the source syllables to match a target without competitive interactions among the syllables). Overall, the data included 13 syllables from eight birds. We excluded 3 of the 13 syllables that showed transient switching back to the source pitch after crossing the midpoint between source and target (see fig. S2A). Because were specifically interested in the part of a trajectory containing the learned shift toward the target, we extracted from each experimental dataset the learning subset of pitch values $\boldsymbol{P}_L = \{P(1), P(2), \ldots, P(t), \ldots, P(L)\}$ from D days. These days contain the shift (identified from the day at which the pitch crossed the midpoint between source and target syllable), in addition to (at most) three consecutive days of stable pitch immediately before and after the shift (we worked with pitch data from fewer than three consecutive days either when the bird started shifting on day 2 after switch or when the experiment was stopped early such that stable pitch data was not available for three full days after the switch).

Model parameters are estimated from $\boldsymbol{P}_L$ as follows. Juvenile birds in artificial tutoring experiments often show small deviations from a perfect imitation of the target song in addition to random daily fluctuations in pitch. Therefore, the mean pitch $S(t)$ of a syllable (Eq. 7) does not necessarily begin at T^0 on t = 1 and end at T on $t = L$, leading us to add the start and end pitch values $T^{0\prime}$ and T' as free model parameters. Data analysis also shows that pitch variance $\sigma_B(t)$ changes over days but is stable within a day (fig. S3G). We thus include an additional set of parameters $\boldsymbol{\sigma}_B = \{\sigma_{B,1}, \ldots, \sigma_{B,D}\}$, where D is the number of experimental days over which the dataset $\boldsymbol{P}_L$ was collected, as described above.

Birds display a variable latency between the switch to target song playback and the onset of pitch modification. In our model, this corresponds to a step change in the mean of the reward distribution from $T^{0\prime}$ (where $\delta = 0$) toward T' (where $\delta \neq 0$), initiating pitch modification. We therefore add an integer parameter c that defines

the step-change index at which the pitch target in the model changes from $T^{0\prime}$ to T'.

We infer all model parameters $\boldsymbol{\theta} = \{\alpha, \beta, c, \boldsymbol{\sigma}_B, \sigma_R, T^{0\prime}, T'\}$ including the shape β and SD σ_R (or a subset of those parameters; see below) using MLE

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{P}_L) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{t=1}^{L} \log p\left[P(t) \mid S(t;\boldsymbol{\theta})\right] \quad (12)$$

where $\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{P}_L)$ defines the log-likelihood that $\boldsymbol{\theta}$ generates data $\boldsymbol{P}_L$, and $S(t;\boldsymbol{\theta})$ is the motor mean as a function of $\boldsymbol{\theta}$ (see Eq. 7). The log-likelihood is maximized through nonlinear constrained optimization using MATLAB's active set algorithm (fmincon). Parameters $\{\alpha, \beta, \boldsymbol{\sigma}_B, \sigma_R\}$ are constrained within the interval $[0, +\infty)$, while $T^{0\prime}$ and T' are constrained to within 0.5 semitones from the true source and target pitch, T^0 and T, respectively. The step-change index $1 \le c \le L$ is rounded to the closest integer after parameter inference. To avoid convergence to a local maximum, we start the maximization routine from different initial points along a grid in parameter space (except $\boldsymbol{\sigma}_B$ that was initialized from day-wise detrended data; see above). Note that σ_R is estimated from syllables only. Determining the SD of the reward distribution of calls remains an outstanding problem.

We considered multiple model variants. We refer to the model containing the complete set of parameters $\boldsymbol{\theta}$ as the full model. We also inferred the parameters of two constrained models where either the reward shape β alone or both β and the learning rate α were fixed (i.e., we infer $\boldsymbol{\theta} = \{\alpha, c, \boldsymbol{\sigma}_B, \sigma_R, T^{0\prime}, T'\}$ or $\boldsymbol{\theta} = \{c, \boldsymbol{\sigma}_B, \sigma_R, T^{0\prime}, T'\}$, respectively). In the latter case, we set α to 0.975, the median estimated value in the full model. When fixing β, we used a range of values corresponding to heavy-tailed (β = 1), normal (β = 2), and light-tailed (β = 4, 6, and 8).

Assessing the goodness of fit using the BIC (*49*) showed comparable results across different model variants (figs. S3, A and D, and 2D).

We further evaluated the models' goodness of fit by simulating 500 pitch trajectories given the inferred model parameters and computing the RMSE between simulated and real data. This resulted in a distribution of RMSEs, D_{RMSE}, per syllable that allows us to assess a model's ability to generalize. To do so, we summarized the central tendencies of those distributions using Tukey's trimean (TM)

$$\mathrm{TM}(D_{\mathrm{RMSE}}) = \frac{Q_1 + 2Q_2 + Q_3}{4} \quad (13)$$

and compared model performance across different parameter settings over the entire set of syllables used for inferring parameters.

We found that the simplest model where both α and β were fixed significantly outperformed the more complex models (see fig. S3E), suggesting that BIC does not capture well the model's ability to generalize. We therefore used the simplest model and its inferred reward SD parameters (10 σ_R values, corresponding to the 10 syllables in the training set in Fig. 2) in all testing simulations (Figs. 4 and 5; two-semitone shift toward two potential targets and four-semitone shift toward one target; 100 simulations per σ_R value, amounting to 1000 simulations per bird in total). In these simulations, we specified other model parameters from the data directly, using the procedures described above. In case a syllable did not shift toward a target (e.g., Figs. 4E and 5D), c was set to the time at which another vocalization in the bird's repertoire (usually a call) started shifting toward the target. In the one case where no vocalization in the bird's repertoire shifted toward the target (see Fig. 5D, bird 4), c was set to its median value over the rest of the experimental dataset. After simulations, we used Tukey's trimean described above (Eq. 13) to compare model performance for different reward shapes across each experimental dataset (Figs. 2F, 4D, left, and 5I).

Motivation for instantaneous reward

In a naive MARL version where all reward components are based on produced pitch, $R_{ij} = R(P_j; T_i)$, inventory reward can only be computed after all syllables and calls have been produced, e.g., at the end of a song motif. Such a model would thus require birds to maintain recent syllable and call instances in short-term memory. However, it is unclear whether the brain would choose such a cumulative strategy of computing reward rather than making an instantaneous estimate available after each syllable and call instance. A fine temporal signaling of inventory reward agrees with highly phasic dopaminergic responses, which can occur in the middle of a motif right after an aversive external stimulus or after the omission of a stimulus (*4*). In addition, dopaminergic activity fluctuations within motifs are correlated with fluctuations in syllable performance (*21*), and there are no reports of its accumulation at motif endings, further suggesting a tight temporal correspondence between syllables and intrinsic rewards. Instantaneous reward signaling is also supported by studies involving experimental interference with birds' auditory feedback in a manner contingent on syllable pitch, showing that interference must be closely time-locked (<100 ms) to syllable performance to drive behavioral changes (*50*).

Motivation for greedy (order-independent) learning of a syllable inventory

Although zebra finches typically imitate both the syllable structure and the syllable order of their tutor's song, a recent study (*18*) has shown that the two hierarchical levels are learned independently of each other. Juvenile zebra finches use a greedy strategy of modifying each syllable that they sing to match the target that it resembles most, an efficient way for a fast acquisition of a syllable inventory. Juveniles use this greedy strategy even when it initially reduces the similarity with their tutor song at the motif level and even when a slightly less greedy strategy could have increased the match with the target at both syllable and motif levels [see figure 4 of (*18*)]. Moreover, when juveniles are experimentally presented with two equally greedy alternative ways to modify a syllable in their song, either toward a target at the "correct" position in the motif or toward a target at the "wrong" position, they choose one of the alternatives at random [see figures 3 and 5 of (*18*)]. Birds later attempt to correct sequencing errors resulting from their greedy strategy by rearranging syllable order. However, the fact that birds initially can choose to make consistent vocal changes that increase the match with the target syllable inventory but reduce the match with the target sequence indicates that syllable learning is order-independent, i.e., that learning a syllable inventory is not contingent on the similarity between the bird's and the target motifs but only on the similarity between the bird's and the target's individual syllables. Note that the assumption of order-independent learning is restricted to the level of song syllables, and it is now unknown whether it applies to other levels of the song hierarchy (e.g., sub-syllabic notes).

Null models

As null hypotheses, we first adapted the model proposed by Fiete and colleagues (*3*), where each actor j is assigned to the target T_j in a SOD manner [i.e., where vocalization $P_j(t)$ is temporally aligned to

T_j in the target song; see Fig. 4D, right]. In the SOD null model, reward at each vocalization instance is a function of the similarity between that instance and the temporally aligned target only, $\boldsymbol{R} = R_{jj} = R(P_j, T_j)$. Reward in one instantiation (SOD 1) is binarized as in the original model (*3*), where $R = 1$ when the absolute error $|P_j - T_j|$ is below a similarity threshold. The similarity threshold is adaptive (using an exponential moving average of reward history) to assure that the actor is rewarded half the time. We give the critic the same functional form as the actor (Eq. 2), to assure that $Q_j(t)$ is centered at the mean $S_j(t)$ and is a differentiable function of the mean, enabling the derivation of a learning rule with stochastic gradient descent (as in Eq. 6). A second instantiation (SOD 2) uses the same light-tailed reward distributions as in the MARL model (Fig. 4D, right).

Another MARL model (MOT model) consists of multiple actors as in the sum-max model, but rather than actors competing over a target, targets compete over an actor by defining each partial reward associated with actor j as a maximum across pairwise comparisons between actor j and all the targets (see fig. S5C, right).

Alternative MARL sum-max model variants

Our findings based on the inventory reward model in Eqs. 9 and 10 are robust to changes in model variant and model parameters. A priori, many models can signal inventory reward after every produced syllable. In another model variant that we simulated, the partial reward $R_{ij}(t)$ associated with target T_i did not depend on the motor means $S_j(t)$, but, instead, it depended on the instance memories $P_j(t)$ of syllables and calls that were last performed within τ. Results were qualitatively unchanged under this instance-memory model, for both shorter and longer windows (from $\tau = 2$ s to $\tau = 2$ hours). We also simulated a model variant that assumes a non-instantaneous intrinsic reward delivered at the end of a motif instead of after each syllable (fig. S5E) and again found that this change had no significant effect on model performance. A similar model that assumes the reward is delivered at the end of the bout also showed comparable results (fig. S5E). The robustness of our model to the variants described above means that our findings do not provide a definitive answer on whether an intrinsic reward is based on instance memories or on motor memories and on the precise temporal contingency of an intrinsic reward. Last, to verify that the parameter ϵ introduced to enforce a hierarchy between syllables and calls (Eq. 9) does not generate a bias toward light-tiled reward distributions, we performed simulations that treated calls and syllables equally by not including this parameter, which yielded comparable results (also favoring light-tailed reward distributions).

Supplementary Materials

This PDF file includes:

Figs. S1 to S6

REFERENCES AND NOTES

1. R. Sutton, A. Barto, Reinforcement learning: An introduction, (The MIT Press Cambridge, MA, 2nd edition, 2018).
2. K. Doya, T. J. Sejnowski, "A Computational Model of Avian Song Learning" in *The New Cognitive Neurosciences*, P. W. F. Poon, J. F. Brugge, Eds. (The MIT Press, ed. 2, 2002), 469–482.
3. I. R. Fiete, M. S. Fee, H. S. Seung, Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances. *J. Neurophysiol.* **98**, 2038–2057 (2007).
4. V. Gadagkar, P. A. Puzerey, R. Chen, E. Baird-Daniel, A. R. Farhang, J. H. Goldberg, Dopamine neurons encode performance error in singing birds. *Science* **354**, 1278–1282 (2016).
5. K. Simonyan, B. Horwitz, E. D. Jarvis, Dopamine regulation of human speech and bird song: A critical review. *Brain Lang.* **122**, 142–150 (2012).
6. E. Hisey, M. G. Kearney, R. Mooney, A common neural circuit mechanism for internally guided and externally reinforced forms of motor learning. *Nat. Neurosci.* **21**, 589–597 (2018).
7. L. A. Hoffmann, V. Saravanan, A. N. Wood, L. He, S. J. Sober, Dopaminergic contributions to vocal learning. *J. Neurosci.* **36**, 2176–2189 (2016).
8. L. Xiao, G. Chattree, F. G. Oscos, M. Cao, M. J. Wanat, T. F. Roberts, A basal ganglia circuit sufficient to guide birdsong learning. *Neuron* **98**, 208–221.e5 (2018).
9. J. J. Day, M. F. Roitman, R. M. Wightman, R. M. Carelli, Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nat. Neurosci.* **10**, 1020–1028 (2007).
10. W. Schultz, P. Dayan, P. R. Montague, A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
11. K. D'Ardenne, S. M. McClure, L. E. Nystrom, J. D. Cohen, BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science* **319**, 1264–1267 (2008).
12. J. Y. Cohen, S. Haesler, L. Vong, B. B. Lowell, N. Uchida, Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482**, 85–88 (2012).
13. L. Xiao, T. F. Roberts, What is the role of thalamostriatal circuits in learning vocal sequences? *Front. Neural Circuits* **15**, 724858 (2021).
14. O. Tchernichovski, P. P. Mitra, T. Lints, F. Nottebohm, Dynamics of the vocal imitation process: How a zebra finch learns its song. *Science* **291**, 2564–2569 (2001).
15. P. Marler, A comparative approach to vocal learning: Song development in white-crowned sparrows. *J. Comp. Physiol. Psychol.* **71**, 1–25 (1970).
16. S. Derégnaucourt, P. P. Mitra, O. Fehér, C. Pytte, O. Tchernichovski, How sleep affects the developmental learning of bird song. *Nature* **433**, 710–716 (2005).
17. D. Lipkind, G. F. Marcus, D. K. Bemis, K. Sasahara, N. Jacoby, M. Takahasi, K. Suzuki, O. Feher, P. Ravbar, K. Okanoya, O. Tchernichovski, Stepwise acquisition of vocal combinatorial capacity in songbirds and human infants. *Nature* **498**, 104–108 (2013).
18. D. Lipkind, A. T. Zai, A. Hanuschkin, G. F. Marcus, O. Tchernichovski, R. H. R. Hahnloser, Songbirds work around computational complexity by learning song vocabulary independently of sequence. *Nat. Commun.* **8**, 1247 (2017).
19. Y. Funabiki, M. Konishi, Long memory in song learning by zebra finches. *J. Neurosci.* **23**, 6928–6935 (2003).
20. M. Schilling, K. Konen, F. W. Ohl, T. Korthals, "Decentralized deep reinforcement learning for a distributed and adaptive locomotion controller of a hexapod robot" in *IEEE International Conference on Intelligent Robots and Systems* (Institute of Electrical and Electronics Engineers Inc., 2020), pp. 5335–5342.
21. A. Duffy, K. W. Latimer, J. H. Goldberg, A. L. Fairhall, V. Gadagkar, Dopamine neurons evaluate natural fluctuations in performance quality. *Cell Rep.* **38**, 110574 (2022).
22. S. Nadarajah, A generalized normal distribution. *J. Appl. Stat.* **32**, 685–694 (2005).
23. S. J. Sober, M. S. Brainard, Vocal learning is constrained by the statistics of sensorimotor experience. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 21099–21103 (2012).
24. P. Ravbar, D. Lipkind, L. C. Parra, O. Tchernichovski, Vocal exploration is locally regulated during song learning. *J. Neurosci.* **32**, 3422–3432 (2012).
25. H. M. Bayer, P. W. Glimcher, Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141 (2005).
26. C. D. Fiorillo, P. N. Tobler, W. Shultz, Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* **299**, 1898–1902 (2003).
27. W. E. Wood, P. J. Osseward II, T. K. Roseberry, D. J. Perkel, A daily oscillation in the fundamental frequency and amplitude of harmonic syllables of zebra finch song. *PLOS ONE* **8**, e82327 (2013).
28. H. Williams, Models for song learning in the zebra finch: Fathers or others? *Anim. Behav.* **39**, 745–757 (1990).
29. T. W. Troyer, A. J. Doupe, An associational model of birdsong sensorimotor learning I. Efference copy and the learning of song syllables. *J. Neurophysiol.* **84**, 1204–1223 (2000).
30. M. G. Kearney, T. L. Warren, E. Hisey, J. Qi, R. Mooney, Discrete evaluative and premotor circuits enable vocal learning in songbirds. *Neuron* **104**, 559–575.e6 (2019).
31. R. Chen, P. A. Puzerey, A. C. Roeser, T. E. Riccelli, A. Podury, K. Maher, A. R. Farhang, J. H. Goldberg, Songbird ventral pallidum sends diverse performance error signals to dopaminergic midbrain. *Neuron* **103**, 266–276.e4 (2019).
32. Y. Mandelblat-Cerf, L. Las, N. Denissenko, M. Fee, A role for descending auditory cortical projections in songbird vocal learning. *eLife* **3**, e02152 (2014).
33. D. Aronov, A. S. Andalman, M. S. Fee, A specialized forebrain circuit for vocal babbling in the juvenile songbird. *Science* **320**, 630–634 (2008).
34. M. S. Fee, J. H. Goldberg, A hypothesis for basal ganglia-dependent reinforcement learning in the songbird. *Neuroscience* **198**, 152–170 (2011).
35. J. Kornfeld, M. Januszewski, P. Schubert, V. Jain, W. Denk, M. S. Fee, An anatomical substrate of credit assignment in reinforcement learning. bioRxiv, doi: 10.1101/2020.02.18.954354 (2020).

36. R. O. Tachibana, D. Lee, K. Kai, S. Kojima, Performance-dependent consolidation of learned vocal changes in adult songbirds. *J. Neurosci.* **42**, 1974–1986 (2022).
37. S. Derégnaucourt, M. Gahr, Horizontal transmission of the father's song in the zebra finch (*Taeniopygia guttata*). *Biol. Lett.* **9**, 20130247 (2013).
38. R. F. Lachlan, C. A. A. van Heijningen, S. M. ter Haar, C. ten Cate, Zebra finch song phonology and syntactical structure across populations and continents—A computational comparison. *Front. Psychol.* **7**, 980 (2016).
39. O. Tchernichovski, S. Eisenberg-Edidin, E. D. Jarvis, Balanced imitation sustains song culture in zebra finches. *Nat. Commun.* **12**, 2562 (2021).
40. O. Fehér, H. Wang, S. Saar, P. P. Mitra, O. Tchernichovski, De novo establishment of wild-type song culture in the zebra finch. *Nature* **459**, 564–568 (2009).
41. O. Tchernichovski, P. P. Mitra, Sound analysis Pro user manual. [Preprint] (2004). http://ofer.sci.ccny.cuny.edu.
42. R. A. Zann, *The Zebra Finch, A Synthesis of Field and Laboratory Studies* (Oxford Univ. Press, 1996).
43. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Stat. Methodol.* **57**, 289–300 (1995).
44. H. van Hasselt, "Reinforcement learning in continuous state and action spaces" in *Adaptation, Learning, and Optimization*, M. van Wiering, M. Otterlo, Eds. (Springer, 2012).
45. A. Schwartz, "A reinforcement learning method for maximizing undiscounted rewards" in *Machine Learning Proceedings 1993* (Elsevier, 1993), pp. 298–305.
46. D. Bennett, Y. Niv, A. J. Langdon, Value-free reinforcement learning: Policy optimization as a minimal model of operant behavior. *Curr. Opin. Behav. Sci.* **41**, 114–121 (2021).
47. A. Duffy, E. Abe, D. J. Perkel, A. L. Fairhall, 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).
48. T. W. Troyer, A. J. Doupe, An associational model of birdsong sensorimotor learning II. Temporal hierarchies and the learning of song sequence. *J Neurophysiol.* **84**, 1224–1239 (2000).
49. G. Schwartz, Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
50. E. C. Tumer, M. S. Brainard, Performance variability enables adaptive plasticity of "crystallized" adult birdsong. *Nature* **450**, 1240–1244 (2007).

Acknowledgments: We thank T. Roberts for comments on draft and T. Tomka for discussion. **Funding:** This work was supported by the German Academic Exchange Service (DAAD), grant 91712284 (H.T.); the US Public Health Service, grant DC04722-137 (O.T.); Swiss National Science Foundation, grant 31003A_182638 (R.H.R.H.); NCCR Evolving Language, agreement no. 51NF40_180888 (R.H.R.H.); and the National Institutes of Health, grant 7N026-00 (D.L.). **Author contributions:** Conceptualization: H.T., A.T.Z., O.T., R.H.R.H., and D.L. Methodology: H.T., A.T.Z., R.H.R.H., and D.L. Software: H.T. Formal analysis: H.T. Investigation: H.T. and D.L. Resources: O.T., R.H.R.H., and D.L. Data curation: H.T. and D.L. Writing—original draft: H.T., R.H.R.H., and D.L. Writing—review and editing: H.T., A.T.Z., O.T., R.H.R.H., and D.L. Visualization: H.T., R.H.R.H., and D.L. Supervision: R.H.R.H. Project administration: R.H.R.H. and D.L. Funding acquisition: H.T., O.T., R.H.R.H., and D.L. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Experimental data are hosted by Dryad and are publicly available at https://doi.org/10.5061/dryad.3r2280gpp. Associated computer code (modeling, analysis, and visualization) is hosted by Zenodo and is publicly available at https://doi.org/10.5281/zenodo.10627609.

Submitted 26 June 2023
Accepted 22 February 2024
Published 27 March 2024
10.1126/sciadv.adj3824