Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Neurobiology

Actor-critic reinforcement learning in the songbird

Ruidong Chen and Jesse H Goldberg

It feels rewarding to ace your opponent on match point. Here, we propose common mechanisms underlie reward and performance learning. First, when a singing bird unexpectedly hits the right note, its dopamine (DA) neurons are activated as when a thirsty monkey receives an unexpected juice reward. Second, these DA signals reinforce vocal variations much as they reinforce stimulus-response associations. Third, limbic inputs to DA neurons signal the predicted quality of song syllables much like they signal the predicted reward value of a place or a stimulus during foraging. Finally, songbirds may solve difficult problems in reinforcement learning – such as credit assignment and catastrophic forgetting – with node perturbation and consolidation of reinforced vocal patterns in motor cortical circuits. Consolidation occurs downstream of a canonical 'actor-critic' circuit motif that learns to maximize performance quality in essentially the same way it learns to maximize reward: by computing and learning from prediction errors.

Address
Department of Neurobiology and Behavior, Cornell University, Ithaca, NY 14853, United States

Corresponding author: Goldberg, Jesse H (jesse.goldberg@cornell.edu)

Current Opinion in Neurobiology 2020, **65**:1–9

This review comes from a themed issue on **Whole-brain interactions between neural circuits**

Edited by **Karel Svoboda** and **Laurence Abbott**

https://doi.org/10.1016/j.conb.2020.08.005

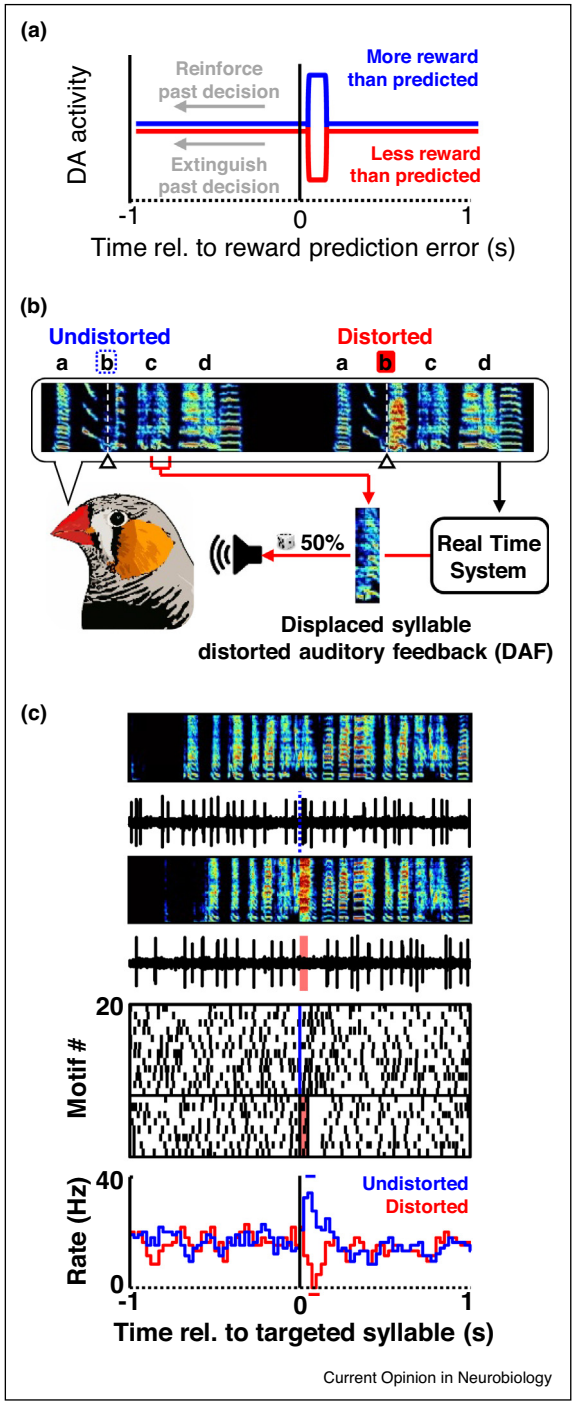
Edward Thorndike captured the essence of reinforcement learning in his Law of Effect: 'Responses that produce a satisfying effect in a particular situation become more likely to occur again in that situation, and responses that produce a discomforting effect become less likely to occur again in that situation [1].' Learning requires three pieces of information: (1) the response ('action') an animal makes; (2) the situation (or 'state') in which the action is taken; and (3) the evaluation of the outcome (effect).

Studies of hungry or thirsty animals learning for rewards have clarified how these three pieces of information are processed in dopamine-basal ganglia (BG) circuits during reinforcement learning (RL). Ventral tegmental area (VTA) dopamine (DA) neurons signal the outcome in the form of 'reward prediction error' (RPE): the difference between actual and predicted reward [2]. DA neurons exhibit bursts in response to unexpected rewards and pauses when a predicted reward is omitted (Figure 1a). In 'actor-critic' (AC) models, these DA signals control synaptic plasticity throughout the basal ganglia, including a ventral 'critic' with outputs back to VTA and a dorsal 'actor' with outputs to the motor system [3] (Figure 2a). Both subdivisions implement DA-modulated plasticity to weigh cortical (or thalamic) inputs (which encode the situation, or 'state') according to their reward value [4•]. DA-modulated plasticity mediated by critic computes predicted value of a state, that is, how much reward to expect in a given situation. Predicted state-value signals, such as ventral striato-pallidal activations to reward-associated cues or places [5], provide VTA with *prediction* information necessary to compute RPE [6••]. VTA projects back to the critic (to update reward associations, or predicted state-value) and also to the 'actor'. DA-modulated plasticity in the 'actor' weighs each state-action pair according to its predicted quality (or Q value). Q value signals may exist in dorsal striatum where the magnitude of premotor activations is strongly reward-modulated. For example, neuronal activations preceding a rightward saccade that will be rewarded are larger than the same saccade that will not [7]. Somehow, motor circuits downstream dorsal striatum convert Q into reward-maximizing action, that is, the policy (Figure 2a)[7].

Intrinsically motivated song learning

Like human speech, birdsong is a complex sequence learned by matching ongoing performance to an internal goal. Juvenile zebra finches memorize a tutor song, begin to babble, and gradually learn over weeks to sing the tutor song. Songbirds have a specialized 'song system,' and it's output RA (robust nucleus of the arcopallium), is a layer 5, primary motor cortex-like nucleus with topographic outputs to brainstem motor neurons (Figure 3). For simplicity, RA can be imagined as a piano keyboard, in which the spatial position of a neuron relates to the vocal muscle it will innervate. RA gets inputs from LMAN (lateral magnocellular nucleus of the anterior nidopallium) and HVC (proper name). LMAN is a frontal cortical nucleus that exhibits stochastic neural activity, drives vocal babbling, contributes to trial-to-trial variability in adults, and projects topographically to RA (i.e. a 'key' in LMAN has a corresponding key in RA) [8,9]. HVC is a premotor cortical nucleus that exhibits stereotyped synfire chain-like sequences of neural activity that drive the

Figure 1



Dopaminergic error signals in singing birds. **(a)** DA neurons signal better-than and worse-than predicted reward outcomes with phasic activations (blue) and suppressions (red). **(b)** Zebra finch songs are motifs consisting of a fixed sequence of syllables, for example, 'a-b-c-d.' A signal processor analyzed and distorted song in real time. A 50 ms snippet of syllable 'c' was played back during production of the target syllable 'b' (target time, black triangles and white dashed lines). Randomly interleaved target renditions were left undistorted (undistorted trials, blue dashed line). This distorted auditory feedback (DAF) induces perceived errors on target syllables. **(c)** Spectrograms with time-aligned voltage traces show responses of a DA neuron during undistorted (top) and distorted trials (bottom). Rasters and histograms show suppressions following distortions (red) and activations following the precise song time-step when error was predicted to occur but did not occur (blue). Reproduced from Ref. [21].

correspondingly stereotyped adult song [10,11•]. HVC axons ignore RA topography (i.e. span the entire keyboard), so that a single HVC axon can, in principle, learn to strike any key [12•].

An actor-critic -inspired framework for song learning

As birds mature from vocal babbling to stereotyped adult song, control of RA firing (and therefore vocal output) gradually transfers from LMAN to HVC [13,14]. Though the practicing bird does not receive external rewards for 'hitting the right note,' we propose that song learning proceeds, at least in part [15••] (Box 1), via an RL-like algorithm implemented in an 'actor-critic' circuit motif in the BG (Figure 2b).

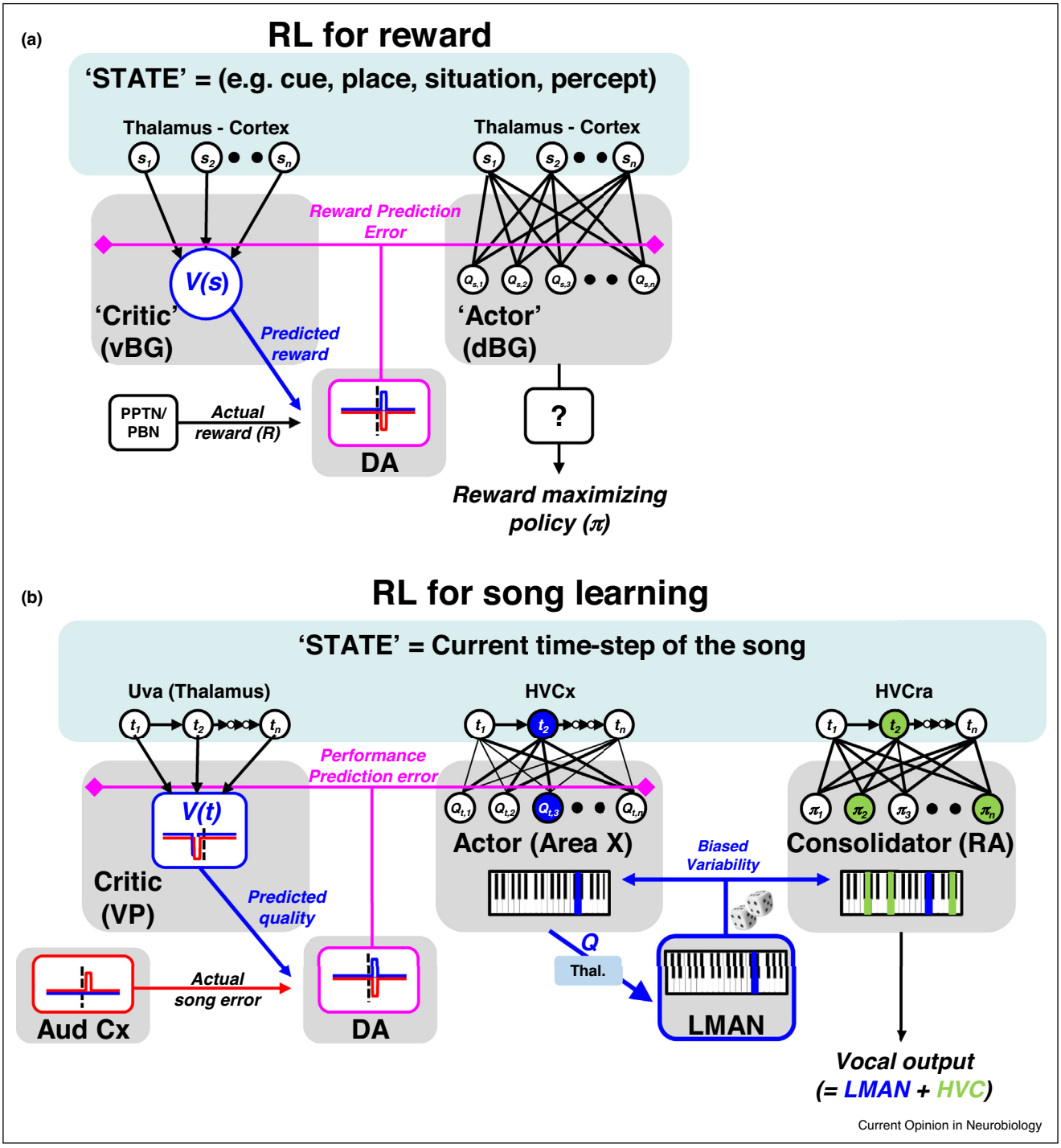
Songbird DA neurons encode RPE-like song evaluation signals

Actor-critic models construct and learn from DA RPE signals. To test for RPE-like signals during singing, we recorded antidromically identified VTA neurons that project to Area X (VTAx) while controlling perceived error with distorted auditory feedback (DAF). DAF is a 50-ms snippet of sound with the same amplitude and spectral content as normal zebra finch song that is known to drive DA and Area X-dependent reinforcement of undistorted syllable variants [16•,17,18•,19,20]. VTAx neurons, known to be dopaminergic, exhibited pauses after distortions (sounded bad) and bursts after undistorted renditions of target syllables (sounded good). Importantly, burst magnitude depended on past error probability: if one syllable was distorted with high probability, and different one with low probability, DA bursts were larger following the (more surprising) undistorted renditions of the high probability target [21]. Thus VTAx DA neurons signal performance prediction error: the difference between how good a syllable sounded and how good it was predicted to sound based on recent practice. To compute error, DA neurons need information, at each time-step, about 'just heard' auditory error as well as predicted error (predicted syllable quality).

Auditory cortex sends 'actual' (just heard) error signals to VTA

A hierarchy of auditory areas converges in a high-order VTA-projecting cortical area called AIV (ventral intermediate arcopallium) [22–24]. VTA-projecting auditory cortical neurons exhibit bursts in response to DAF during singing [25••]. Electrical microstimulation of auditory cortex drives pauses in VTAx neurons [26••] and optogenetic activation of the

Figure 2



Actor-critic RL for both reward and song learning. **(a)** The environment provides current state information, S, and current reward, r. The 'Actor' learns the quality of state/action pairs (Q(s,a)) that get converted into the reward maximizing action given the state (i.e. the *policy* $\pi(a|S)$). DA-weighted state representations in the critic compute the predicted state-value, *V(s)*. DA neurons signal RPE by taking the difference between actual, *R*, and predicted, *V(s)*, reward. **(b)** Lower left: VTA-projecting auditory cortical (Aud. Cx) neurons encode auditory error, for example, bursts following DAF [25••]. Inset schematizes firing rates during distorted (red) and undistorted (blue) renditions; vertical dashed line denotes the time-step 'targeted' with DAF (as in Figure 1b). Bursts in auditory cortex drive pauses in VTAx neurons through local VTA inhibition (not shown) [26••,28•]. The DA error signal (pink line) goes to both VP ('Critic,' left) and Area X ('Actor,' right). DA modulated plasticity in VP could weigh time-step (i.e. 'state') information according to past error. With an eiligibility trace [4•], this would explain why most VPvta neurons exhibited pauses right before the DAF target time [26••]. This predicted quality signal, similar to predicted state value in classic 'critic' circuits, could help VTAx neurons compute prediction error. DA-modulated plasticity in Area X, schematized as a keyboard due to its topographic organization, could learn

Box 1 Are supervised and unsupervised algorithms also implemented in birdsong?

Unsupervised learning can be implemented with correlation (e.g. Hebbian)-based learning rules and without explicit error or reinforcement signals. For example, during babbling the activity of a 'chhh'-producing motor neuron will be reliably correlated with a 'chhh' receptive auditory neuron. And a 'bb' motor neuron will similarly correlate with a 'bb' auditory neuron. Hebbian learning rules could create paired forward and inverse models. In the forward model, the motor system 'tells' the sensory system what is about to happen, so that the sensory consequences of movements can be predicted. In the inverse model, the sensory system can 'call upon' the motor system to produce the desired output [15••,23]. Reciprocal connections between HVC and the auditory system could instantiate these internal models [47]. In fact, a forward model could be important for extracting prediction error signals in auditory cortical areas upstream VTA. *Supervised learning*: Supervised error signals encode precisely how an outcome differed from the target, which also specifies the necessary correction (e.g. the pitch was too low, so next time move it up). Learning from supervised error signals requires an inverse model that can implement the correction. Though birds may have such internal models [23], birds surprisingly appear to solve even pitch-shifting experiments with DA reinforcement mechanisms [49].

Box 2 How might a relatively slow DA reinforcement signal improve a fast behavior?

Birds can produce reliable acoustic fluctuations with ~5–10 ms precision, the same duration as an HVC burst, the schematized 'time-step' in our model [11•]. Yet the DA reinforcement signal is ~50 ms delayed from auditory error and lasts ~100 ms [21]. How might this relatively slow signal appropriately reinforce past vocalizations? Several lines of evidence suggest that an eligibility trace in the spines of Area X MSNs last around ~0.1 s. In carefully implemented distorted auditory feedback experiments, the Brainard group discovered that DAF only reinforces vocal variations in the immediately preceding 0.1 s [16•]. They also discovered that ~0.1 s duration 'chunks' of song are reinforced even when DAF is targeted with millisecond precision to specific syllable trajectories [50]. Although it may seem optimal to independently evaluate every ~5 ms time-step, we propose that a coarser evaluation system may work for birdsong. Acoustic structure of a syllable is largely a function of air pressure and muscle activation in the syrinx, and therefore song production is better understood as a continuous trajectory through syringeal state space rather than transition between discrete states. Because neither air pressure nor muscle configuration can be instantly transformed, the action at each time-step constrains what new configurations are possible in the next. For example, an input to the syrinx that drives a 5 Hz increase in pitch would only produce the desired 500 Hz output when the preceding pitch was 495 Hz. We hypothesize that reinforcing a larger chunk of consecutive actions could reduce the dimensionality of search space and improve learning.

auditory cortical-VTA pathway extinguishes syllable variations (just like phasic suppression of the VTA-X pathway does) [27,28•,29••]. Auditory cortical areas can signal error and drive pauses in DA firing (Figure 2b, lower left). The finch auditory cortical area that projects to VTA may be functionally analogous to anterior cingulate cortex, which also may send performance error signals to VTA [30].

Ventral pallidum (VP) sends predicted syllable quality signals to VTA

Songbird VP is a mixed striatopallidal nucleus [31] that may function analogously to the critic [26••,28•]. VP is necessary for learning and receives inputs from Uva, a thalamic nucleus that sends song time-step information to HVC, and also from VTAx neurons. DA-modulated plasticity of Uva inputs could weigh time-steps according to their past error. For example, consider a song with three time-steps t_1, t_2, t_3. If t_2 is reliably correlated with error, then DA pauses (driven by auditory cortex, as described above) would be coincident with those Uva inputs active at t_2. Then DA-modulated plasticity would re-weigh these synapses, resulting in a representation in VP of error-weighted timing or, equivalently, predicted syllable quality (Figure 2b, upper left). Consistent with this idea, most antidromically identified VPvta neurons exhibited pauses in firing immediately *before* the song time-step associated with past error, exactly consistent with a predicted syllable quality signal (Figure 2b) [26••].

Put simply, imagine an animal foraging a familiar environment in search of food. It will have a memory of where it got rewards, resulting in a place-dependent reward prediction. Now imagine a bird practicing a song with many syllables. It will similarly have a memory of when in the song it made mistakes, resulting in a syllable-dependent error prediction. Thus we view VP's role in computing the predicted quality of syllables as conceptually similar to its long-established role as a 'critic' that computes the predicted reward value of cues or places.

HVC provides 'state' information in the form of what 'time it is' in the song

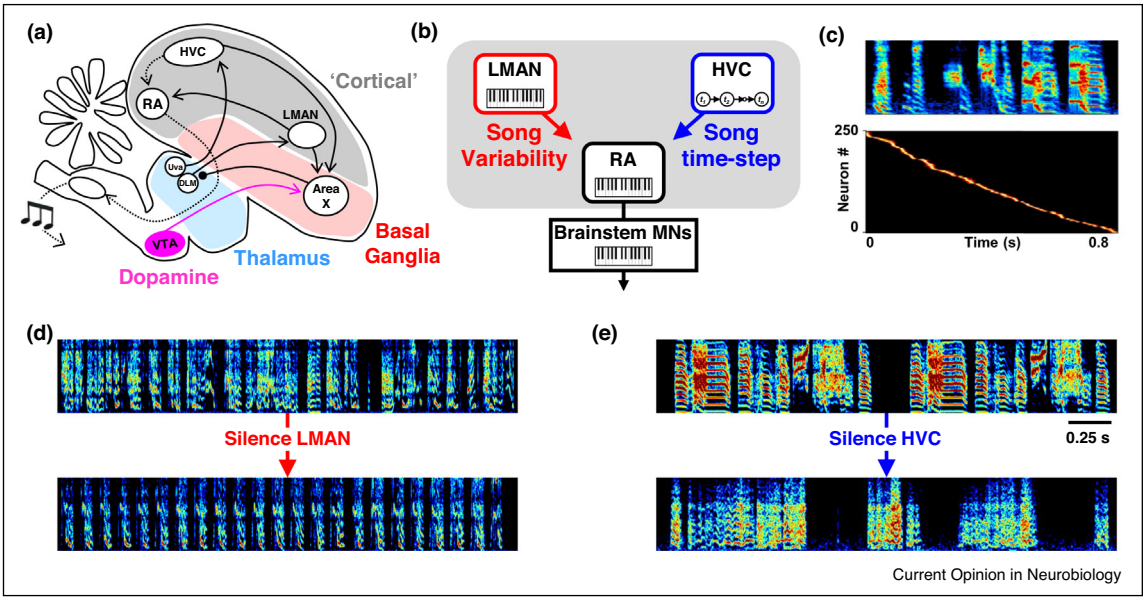
HVC time-steps (Figure 3c) are ideal 'state' representations because song policy is, in essence, learning *what* piano key to press *when* [12•]. Importantly, because the axons of single HVC or VTA neurons span the entirety of Area X, any key can in principle be learned to be struck at any given time-step [12•].

Dorsal BG nucleus Area X as part of the 'actor' that computes Q values

Area X gets three main inputs: LMAN (which provides a copy of the vocal 'guess' it made through RA), HVC (time-step) and VTAx (DA RPE). We proposed a specific learning rule in Area X based on dopamine modulated corticostriatal

the quality of each state/action pair ($Q(s,a)$). For example Area X could learn that striking blue key at t2 is high quality, and relay this signal through DLM to LMAN, resulting in a premotor 'bias' signal that, together with the HVC-driven motor program, produces the vocal output. LMAN bias, if consistently associated with better-than-predicted song outcomes, is consolidated over days into the motor program (e.g. the Area X and LMAN-driven blue key in RA eventually becomes an HVC-driven green key, lower right) [17,18•]. Note that several inputs to VTA in both mammals and songbirds are omitted for clarity.

Figure 3



Distinct variability- and timing- generating pathways in the song system. **(a-b)** The HVC→RA 'motor' pathway exhibits stereotyped neural activity that drives adult song; the LMAN→RA pathway drives vocal variability and babbling. The basal ganglia nucleus Area X receives DA inputs from VTA projects to LMAN via a motor thalamic nucleus called DLM. **(c)** HVC synfire chains track time-step in song. Spectrogram of a song motif plotted above a heatmap of HVC bursting activity. Each row shows the smoothed firing rate of a single burst. **(d)** Spectrogram from a babbling bird before (top) and after (bottom) LMAN inactivation. Note loss of vocal variability, revealing stereotyped, HVC-driven song elements. **(e)** Spectrogram from adult bird before (top) and after (bottom) HVC inactivation. Note elimination of song structure, revealing LMAN-driven vocal variability. Data from [11•]. Inputs to VTA are omitted for clarity.

plasticity: If an HVC and LMAN input to a striatal medium spiny neuron (MSN) are co-active, then an eligibility trace (Etrace) is transiently activated that 'tags' the HVC-MSN synapse [4•] (Box 2). If the LMAN activation was a 'lucky guess,' then a phasic DA burst will occur that coincides with the eligibility trace (Box 2), which would strengthen the tagged HVC-MSN synapse (we proposed the LMAN input implements node perturbation and is not itself plastic, see below). Across Area X, this process would compute Q, the quality of each action in a given state (i.e. the quality of each key strike at each time-step of the song). At each time-step of the song, a new vector of Q values (of length n = number of keys) could then *bias* LMAN to strike a new combination of high quality keys. Consistent with this model, lesions to Area X or its DA inputs block bias [19,20], and optogenetic activation of DA terminals in Area X at a specific time-step reinforces immediately preceding vocal variations [27,29••]. Thus DA-modulated plasticity in Area X can compute the quality of pressing each key at each time-step (i.e. the quality of each state/action pair, or Q(s,a)).

The idea of Area X as a Q network makes additional predictions. If performance at a given time-step is poor no matter what action is taken, then the predicted quality of all state/action pairs will be low. This occurs naturally during vocal babbling, before the onset of learning, when the predicted quality of all syllables is likely to be low. At this learning stage, all Area X output neurons exhibit phasic activations at syllable onsets, exactly where state representations (HVC activity) are also concentrated [32,33]. As Area X outputs are inhibitory, we hypothesize that this is Area X's way of saying it has not yet learned any policy to promote. These pre-syllable activations rapidly go away over days of singing experience, as birds have an opportunity to learn which state/action pairs produce good outcomes [34]. This hypothesis also predicts that all Area X output neurons recorded in adult birds should exhibit phasic rate increases prior to a song time-step targeted with DAF with 100% probability, an easy experiment.

Consolidation in a premotor – motor cortical pathway

In actor-critic models used in machine learning, deletion of the actor would have a devastating effect on task performance. Yet song remains intact following lesions to Area X or its downstream thalamo-cortical pathway (Area X-DLM-LMAN). Seminal experiments showed how LMAN-driven 'bias' is consolidated into the HVC-RA motor cortical pathway [17,18•]. When DAF was delivered only at low-pitch renditions of a target syllable, birds learned over hours to move that syllable's

pitch up. When LMAN was inactivated at the end of this day, the pitch of the target syllable immediately returned to the morning's value, showing that Area X rapidly learned to *bias* LMAN to push the song away from error (i.e. to 'educate' LMAN guesses). Yet after days of sustained pitch-up bias, LMAN inactivation no longer caused a pitch shift, meaning that the pitch-up bias had been transferred from LMAN to the HVC-RA synapse (i.e. the LMAN bias was consolidated into the HVC-RA pathway (Figure 2b lower right and legend, i.e. the blue key in RA will turn into a green key).

Songbird variations on the classic actor-critic may solve challenging problems in RL

A first unique feature of the songbird architecture is that the actor (Area X) sits upstream of a 'variability-generator' (LMAN) which in turn projects to a 'consolidator' (RA). These added thalamocortical layers between the actor and motor output may help solve two important problems in RL: credit assignment and catastrophic forgetting.

Solving credit assignment with LMAN-dependent node perturbation in Area X

One problem in RL is credit assignment: after an error, how does the brain know which of its millions synapses need to be changed? The error-backpropagation algorithm used in machine learning updates each synaptic weight based on its known unique contribution to behavioral output, but this might not be biologically plausible [35]. An alternative approach is node-perturbation, which associates the change in error caused by local stochastic fluctuations in neural activity [36]. In node perturbation, connections mediating exploratory behavioral variations are not themselves plastic – but instead they gate plasticity of other inputs that can take over to drive a successful variation. Node perturbation provided the inspiration for our proposed learning rule in Area X: Area X MSNs detect which 'guesses' (from LMAN) at which time-steps (from HVC) result in better-than-predicted outcome (from VTA) [12•]. In this model, only the HVC-MSN synapse is plastic, and the LMAN input is there to provide a 'copy' of the perturbation to vocal output caused by LMAN's collateral in RA. A recent EM study identified a micro-architecture in Area X potentially suited to implement node perturbation: HVC-MSN synapses were primarily on dendritic spines, where DA-modulated plasticity is known to occur [4•]. When a single axon contacted multiple spines of a single MSN, spines were correlated in size – a structural hallmark of Hebbian plasticity [37•]. Meanwhile, LMAN-MSN synapses were primarily on dendritic shafts, possibly situating them to gate HVC-spine plasticity.

Solving 'catastrophic forgetting' with consolidation in the motor cortical nucleus RA

Another classic problem of motor sequence reinforcement learning is knowing *when* to allow for plastic changes to a sequence. For example, zebra finches learn to sing by sequentially adding new syllables to their songs [33]. It would be maladaptive to enable plasticity in synapses important for producing syllables 'A' and 'B' that have already been mastered as the bird is attempting to learn syllable 'C.' In artificial neural networks, this is known as 'catastrophic forgetting'. Synaptic weight changes that maximize performance of newly learned behaviors can impair previously learned ones. This problem can be solved with 'elastic weight consolidation' - a process that protects synaptic weights that are useful for already-learned behaviors [38••]. Consolidation in RA may reduce catastrophic forgetting in several ways. First, after the HVC-RA pathway 'takes over' control of a specific part of the song, Area X synapses are free to learn (or unlearn) new policies without degrading ongoing vocal performance. Area X policies can 'bias' LMAN variability and, only if a bias is stable for days, will it get consolidated into the HVC-RA pathway [17,18•]. Second, neurogenesis of RA-projecting HVC neurons occurs throughout song learning, which could enable weight changes of new HVC-RA connections to occur without altering previously learned ones [39]. Third, plasticity of existing HVC-RA synapses could be gated by uncertainty – such that reliably well executed time-steps of the motor sequence are 'protected.' For example, if the bird repeatedly makes mistakes (or is distorted) at one 'difficult' time-step in the song, the predicted error is high at that specific time-step. Importantly, cholinergic inputs to RA and HVC come from VP (where predicted error signals are known to reside), inhibition in HVC is reduced during new syllables [40], and cholinergic signaling in RA is required for synaptic plasticity and for song learning [41,42]. We predict that RA projecting VP neurons exhibit bursts of activity immediately before error-prone time-steps (i.e. DAF-targeted) of the song. We predict that acetylcholine 'tells' motor cortex when a time-step with an uncertain outcome is about to occur, enabling synaptic plasticity important for consolidation specifically at this time-step of the sequence. We also predict a specific Ach-modulated heterosynaptic learning rule in RA: If LMAN, HVC and cholinergic inputs to an RA neuron are reliably coactive, then strengthen the connection strength between HVC-RA. This rule would enable an HVC time-step to 'take control' of striking a high quality key specifically at low quality time-steps, and at the same time would 'protect' existing HVC-RA synapses at reliably high quality time-steps. We hypothesize that cholinergic uncertainty signals in mammalian motor cortex could serve a similar function [43,44].

Non standard 'actor-to-critic' projections may implement advantage actor-critic

Curiously, we found that parts of the proposed 'actor' pathway (Area X, DLM and RA) project to 'the critic' VP (see also Ref. [31]), revealing projections from actor back to critic not required in standard actor-critic models.

Notably, a growing family of 'advantage actor critic' algorithms could make use of such projections [45]. In contrast to classic actor-critic where the RL signal is the difference between reward received and the predicted state value V(s), in advantage actor-critic, the RL signal to the actor (for policy update) additionally considers the 'advantage,' that is, the difference between the predicted quality of the action taken (i.e. the Q value Q(s,a)) and the predicted value of the state (V(s)). VP could inherit Q information from Area X, could compute V(s) as described in Figure 2b, and could compute the advantage as the difference between the two. This advantage could be relayed to VTA to influence its reinforcement signal.

This idea predicts that the songbird could compare actual song quality not just to the quality predicted at each time-step (V(s)), but additionally to the predicted quality of the action taken at that time-step (Q(s,a)), analogous to what an advantage actor-critic network does. A simple experiment to test this possibility would be to record VTAx DA neurons while manipulating the advantage, $A(s_t,a_t)$. This could be done by implementing conditional distorted auditory feedback in which only low-pitch variants of a specific target syllable are distorted [16•,17]. On rare catch trials, low pitch variants would instead be left undistorted. If DA neurons signal the difference between the actual outcome to the predicted outcome given the state (the target time of the song), then the magnitude of DA bursts would be the same for all undistorted target renditions, regardless of which syllable variant was produced. But if DA activity has information about the advantage, then DA bursts following undistorted renditions of the low pitch variants may be larger than bursts following undistorted renditions of high pitch variants (because low pitch renditions have been associated with histories of more error). Future recordings of VTAx neurons could therefore constrain which variant of actor-critic-like algorithms is realized in the songbird.

Summary

Many open questions remain in songbirds. Foremost, it remains unknown how auditory pathways compare the song to the tutor. This process may occur in reciprocal connections between auditory areas and HVC and likely involves both efference-copy and tutor-memory guided cancellation and evaluation of predicted acoustic outcomes [15••,46,47]. Second, our model fails to capture the real complexity of BG circuits (e.g. distinct cell types and pathways) and oversimplifies how DA signals are constructed and used. For example songbirds and mammals share indirect pathways and striatal interneuron classes whose roles in learning remain unclear [34,48]. And because VTA-projecting neurons in mammals and birds encode an incredible diversity of motor, reward, and error-related signals, it remains unclear how relatively homogenous DA error signals are computed from mixed inputs [6••,26••]. Finally, our model focuses on learning in adult birds where clear-cut time-step representations already exist in the HVC chain. It remains unclear what neural mechanisms enable HVC chains to develop in the first place [33]. Finally, we acknowledge that this review is primarily taking inspiration from actor-critic models to formulate an algorithmic-level description of song learning. Yet an implementation-level understanding may require more detailed analysis of spiking neuron models with distinct cell classes, as well as more investigation into precisely how DA modulated plasticity is implemented with eligibility traces. Such studies could in turn refine the algorithmic-level ideas presented here.

Comparative approaches can distinguish general principles from behavior-, effector-specific, and species-specific solutions to motor learning problems, and can also generate new hypotheses. For example, we predict that placing the 'actor' upstream of a 'guesser' and a 'consolidator' (as the bird's do with LMAN and RA) could lead to improved machine implementation of sequence learning. We also believe that the utility of the actor-critic framework in song learning, reward based learning, and machine learning suggests a general principle for computing and learning from prediction errors.

Conflict of interest statement

Nothing declared.

Acknowledgements

We thank Josh Dudman, members of the Goldberg lab and reviewers for constructive feedback. This work was funded by the Pew Charitable Trusts, a Klingenstein Fellowship, and the N.I.H. (R01NS094667).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Thorndike EL: *Animal Intelligence*. Darien, CT: Hafner; 1911.

2. Schultz W: **Behavioral dopamine signals**. *Trends Neurosci* 2007, **30**:203-210.

3. Joel D, Niv Y, Ruppin E: **Actor-critic models of the basal ganglia: new anatomical and computational perspectives**. *Neural Netw* 2002, **15**:535-547.

4. • Yagishita S, Hayashi-Takagi A, Ellis-Davies GC, Urakubo H, Ishii S, Kasai H: **A critical time window for dopamine actions on the structural plasticity of dendritic spines**. *Science* 2014, **345**:1616-1620

After decades of speculation about eligibility traces, the authors finally gained experimental control over striatal MSNs and their glutamatergic and dopaminergic inputs in mouse brain slice. By systematically varying delays between activation combinations (pre, post, dopamine), the authors measured how LTP depends on DA activation in a critical time window after pre/post pairing.

5. Humphries MD, Prescott TJ: **The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward**. *Prog Neurobiol* 2010, **90**:385-417.

6. •• Tian J, Huang R, Cohen JY, Osakada F, Kobak D, Machens CK, Callaway EM, Uchida N, Watabe-Uchida M: **Distributed and mixed information in monosynaptic inputs to dopamine neurons**. *Neuron* 2016, **91**:1374-1389

The authors used Rabies viral methods to image the activity of inputs to midbrain DA neurons in a simple Pavlovian task. Though DA neurons homogenously encoded RPE, diverse inputs to DA neurons from several brain regions, including VP, encoded virtually every conceivable reward-related signal. This study highlighted that the construction of DA RPE signals may be more complex than simply subtracting one input from another.

7. Ito M, Doya K: **Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit**. *Curr Opin Neurobiol* 2011, **21**:368-373.

8. Olveczky BP, Andalman AS, Fee MS: **Vocal experimentation in the juvenile songbird requires a basal ganglia circuit**. *PLoS Biol* 2005, **3**:e153.

9. Kao MH, Brainard MS: **Lesions of an avian basal ganglia circuit prevent context-dependent changes to song variability**. *J Neurophysiol* 2006, **96**:1441-1455.

10. Picardo MA, Merel J, Katlowitz KA, Vallentin D, Okobi DE, Benezra SE, Clary RC, Pnevmatikakis EA, Paninski L, Long MA: **Population-level representation of a temporal sequence underlying song production in the Zebra finch**. *Neuron* 2016, **90**:866-876.

11. • Lynch GF, Okubo TS, Hanuschkin A, Hahnloser RH, Fee MS: **Rhythmic continuous-time coding in the songbird analog of vocal motor cortex**. *Neuron* 2016, **90**:877-892

Together with Ref. [10], the authors recorded large populations of HVC neurons in single birds and found that HVC neurons discharge in 'synfire chains' that tile the entire song motif.

12. • Fee MS, Goldberg JH: **A hypothesis for basal ganglia-dependent reinforcement learning in the songbird**. *Neuroscience* 2011, **198**:152-170

A specific model of song learning inspired by RL mechanisms in mammals. We predicted the existence of DA error signals that act in Area X to link HVC time-steps to performance maximizing vocal variations.

13. Aronov D, Veit L, Goldberg JH, Fee MS: **Two distinct modes of forebrain circuit dynamics underlie temporal patterning in the vocalizations of young songbirds**. *J Neurosci* 2011, **31**:16353-16368.

14. Garst-Orozco J, Babadi B, Ölveczky BP: **A neural circuit mechanism for regulating vocal variability during song learning in zebra finches**. *eLife* 2014, **3**:e03697.

15. •• Hahnloser R, Ganguli S: **Vocal learning with inverse models**. In *Principles of Neural Coding.* Edited by Panzeri S, Quiroga P. CRC Taylor and Francis; 2013:547-564

A detailed mathematical model shows that unsupervised learning mechanisms can build forward and inverse models that pair vocal motor actions with their sensory consequences and vice versa. The authors show that these internal models can, in principle, implement vocal learning even without reinforcement signals.

16. • Tumer EC, Brainard MS: **Performance variability enables adaptive plasticity of 'crystallized' adult birdsong**. *Nature* 2007, **450**:1240-1244

Authors used pitch-contingent distorted auditory feedback (DAF) to drive experimentally controlled song learning. The authors behaviorally measured the eligibility trace for learning when DAF was delayed by $>$0.1 s, learning did not occur.

17. Andalman AS, Fee MS: **A basal ganglia-forebrain circuit in the songbird biases motor output to avoid vocal errors**. *Proc Natl Acad Sci U S A* 2009, **106**:12518-12523.

18. • Warren TL, Tumer EC, Charlesworth JD, Brainard MS: **Mechanisms and time course of vocal learning and consolidation in the adult songbird**. *J Neurophysiol* 2011, **106**:1806-1821

This paper, along with ref 17, implemented pitch-contingent distorted auditory feedback to reinforce specific pitch changes in specific syllables. Early in learning, bias to avoid distortions was driven by LMAN. Yet after days, pitch learning was consolidated into the HVC→RA pathway.

19. Hoffmann LA, Saravanan V, Wood AN, He L, Sober SJ: **Dopaminergic contributions to vocal learning**. *J Neurosci* 2016, **36**:2176-2189.

20. Ali F, Otchy TM, Pehlevan C, Fantana AL, Burak Y, Olveczky BP: **The basal ganglia is necessary for learning spectral, but not temporal, features of birdsong**. *Neuron* 2013, **80**:494-506.

21. Gadagkar V, Puzerey PA, Chen R, Baird-Daniel E, Farhang AR, Goldberg JH: **Dopamine neurons encode performance error in singing birds**. *Science* 2016, **354**:1278-1282.

22. Keller GB, Hahnloser RH: **Neural processing of auditory feedback during vocal practice in a songbird**. *Nature* 2009, **457**:187-190.

23. Giret N, Kornfeld J, Ganguli S, Hahnloser RH: **Evidence for a causal inverse model in an avian cortico-basal ganglia circuit**. *Proc Natl Acad Sci U S A* 2014, **111**:6063-6068.

24. Moore JM, Woolley SMN: **Emergent tuning for learned vocalizations in auditory cortex**. *Nat Neurosci* 2019, **22**:1469-1476.

25. •• Mandelblat-Cerf Y, Las L, Denisenko N, Fee MS: **A role for descending auditory cortical projections in songbird vocal learning**. *eLife* 2014, **3**

An auditory cortical area, AIV, is important for song learning and sends auditory error signals to VTA.

26. •• Chen R, Puzerey PA, Roeser AC, Riccelli TE, Podury A, Maher K, Farhang AR, Goldberg JH: **Songbird ventral pallidum sends diverse performance error signals to dopaminergic midbrain**. *Neuron* 2019, **103**:266-276 e264

Lesions, viral tracing, and electrophysiology show that VP is necessary for learning, is interconnected with the song system, and sends diverse error-related signals to VTA, including information about predicted syllable quality. VP is proposed to play the role of the 'critic' in AC circuits.

27. Xiao L, Chattree G, Oscos FG, Cao M, Wanat MJ, Roberts TF: **A basal ganglia circuit sufficient to guide birdsong learning**. *Neuron* 2018, **98**:208-221 e205.

28. • Kearney MG, Warren TL, Hisey E, Qi J, Mooney R: **Discrete evaluative and premotor circuits enable vocal learning in songbirds**. *Neuron* 2019, **104** http://dx.doi.org/10.1016/j.neuron.2019.07.025 559-575.e6; Epub 2019 Aug 22.PMID: 31447169

Optogenetic activation of an auditory cortical-VTA pathway and VP-VTA pathway extinguish and reinforce vocal variations, respectively.

29. •• Hisey E, Kearney MG, Mooney R: **A common neural circuit mechanism for internally guided and externally reinforced forms of motor learning**. *Nat Neurosci* 2018:1

Together with Ref. [27], authors showed that optogenetic manipulations of phasic DA signaling in Area X can reinforce immediately preceding vocal variations.

30. Kolling N, Wittmann MK, Behrens TE, Boorman ED, Mars RB, Rushworth MF: **Value, search, persistence and model updating in anterior cingulate cortex**. *Nat Neurosci* 2016, **19**:1280.

31. Gale SD, Person AL, Perkel DJ: **A novel basal ganglia pathway forms a loop linking a vocal learning circuit with its dopaminergic input**. *J Comp Neurol* 2008, **508**:824-839.

32. Pidoux M, Bollu T, Riccelli T, Goldberg JH: **Origins of basal ganglia output signals in singing juvenile birds**. *J Neurophysiol* 2014. jn 00635 02014.

33. Okubo TS, Mackevicius EL, Payne HL, Lynch GF, Fee MS: **Growth and splitting of neural sequences in songbird vocal development**. *Nature* 2015, **528**:352-357.

34. Goldberg JH, Adler A, Bergman H, Fee MS: **Singing-related neural activity distinguishes two putative pallidal cell types in the songbird basal ganglia: comparison to the primate internal and external pallidal segments**. *J Neurosci* 2010, **30**:7088-7098.

35. Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G: **Backpropagation and the brain**. *Nat Rev Neurosci* 2020:1-12.

36. Fiete IR, Fee MS, Seung HS: **Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances**. *J Neurophysiol* 2007, **98**:2038-2057.

37. • Kornfeld J, Januszewski M, Schubert P, Jain V, Denk W, Fee MS: **An anatomical substrate of credit assignment in reinforcement learning**. *BioRxiv* 2020

Serial EM reconstruction of synapses onto Area X MSNs show that HVC inputs preferentially contact spines and exhibit structural hallmarks of synaptic plasticity; meanwhile, LMAN inputs prefer shafts. This configuration, predicted in Ref. [24], is consistent with LMAN variability gated

plasticity of HVC inputs, which would support node perturbation theories of credit assignment.

38. •• Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, Milan K, Quan J, Ramalho T, Grabska-Barwinska A *et al.*: **Overcoming catastrophic forgetting in neural networks**. *Proc Natl Acad Sci U S A* 2017, **114**:3521-3526
In an artificial neural network, authors introduce 'elastic weight consolidation' as a method to address the problem of catastrophic forgetting. Synapses whose weights are important for ongoing performance are protected from further change.

39. Scott BB, Lois C: **Developmental origin and identity of song system neurons born during vocal learning in songbirds**. *J Comp Neurol* 2007, **502**:202-214.

40. Vallentin D, Kosche G, Lipkind D, Long MA: **Inhibition protects acquired song segments during vocal learning in zebra finches**. *Science* 2016, **351**:267-271.

41. Puzerey PA, Maher K, Prasad N, Goldberg JH: **Vocal learning in songbirds requires cholinergic signaling in a motor cortex-like nucleus**. *J Neurophysiol* 2018, **120**:1796-1806 http://dx.doi.org/10.1152/jn.00078.2018 Epub 2018 Jul 11. PMID: 29995601.

42. Salgado-Commissariat D, Rosenfield DB, Helekar SA: **Nicotine-mediated plasticity in robust nucleus of the archistriatum of the adult zebra finch**. *Brain Res* 2004, **1018**:97-105.

43. Ramanathan DS, Conner JM, Anilkumar AA, Tuszynski MH: **Cholinergic systems are essential for late-stage maturation and refinement of motor cortical circuits**. *J Neurophysiol* 2015, **113**:1585-1597.

44. Yu AJ, Dayan P: **Uncertainty, neuromodulation, and attention**. *Neuron* 2005, **46**:681-692.

45. Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K: **Asynchronous methods for deep reinforcement learning**. *International Conference on Machine Learning* 2016:1928-1937.

46. Mackevicius EL, Fee MS: **Building a state space for song learning**. *Curr Opin Neurobiol* 2018, **49**:59-68.

47. Roberts TF, Hisey E, Tanaka M, Kearney MG, Chattree G, Yang CF, Shah NM, Mooney R: **Identification of a motor-to-auditory pathway important for vocal learning**. *Nat Neurosci* 2017, **20**:978-986.

48. Goldberg JH, Fee MS: **Singing-related neural activity distinguishes four classes of putative striatal neurons in the songbird basal ganglia**. *J Neurophysiol* 2010, **103**:2002-2014.

49. Saravanan V, Hoffmann LA, Jacob AL, Berman GJ, Sober SJ: **Dopamine depletion affects vocal acoustics and disrupts sensorimotor adaptation in songbirds**. *eNeuro* 2019, **6**.

50. Charlesworth JD, Tumer EC, Warren TL, Brainard MS: **Learning the microstructure of successful behavior**. *Nat Neurosci* 2011, **14**:373-380.