# Actor-Critic Melody Learning in Songbirds

Ratnadeep Pal*
Texas A&M University
College Station, Texas
rpal@tamu.edu

Laren Spear*
Texas A&M University
College Station, Texas
larenspear@tamu.edu

## ABSTRACT

We used the actor-critic framework to model how songbirds learn to sing in a simplified environment. We discuss a proposed new model, advantage actor-critic with successive step discounting, which models brain activity more closely but has negative performance implications. We find that proximal policy optimization (PPO) performs the best when both PPO and advantage actor-critic (A2C) converge, but when they do not, A2C generally lands on a higher average normalized reward.

Github: https://github.com/larenspear/SongbirdActorCritic
YouTube: https://youtu.be/zxv7uvnorBo

## CCS CONCEPTS

• **Applied computing** → **Biological networks**; • **Computing methodologies** → **Reinforcement learning**; *Neural networks*; *Markov decision processes.*

## KEYWORDS

reinforcement learning, actor-critic, neural networks, songbirds, proximal policy optimization

**ACM Reference Format:**
Ratnadeep Pal and Laren Spear. . Actor-Critic Melody Learning in Songbirds. In *Proceedings of* . ACM, New York, NY, USA, 5 pages.

## 1 INTRODUCTION

There has long been a connection between the mathematical formulation of reinforcement learning and the natural reinforcement learning methods in the brain. The concept of reward or punishment in response to correct or incorrect actions is the basis of how humans and animals interact with their environment. As an animal grows up, it learns a variety of behaviors by interacting with its environment.

Many papers have made a connection between various parts of the brain acting in opposition to each other to make decisions in an actor-critic fashion. Wanjerkhede et al [18] compare the striatum, the main component of motor control and reward systems in the brain, to an actor-critic circuit. The two parts of the striatum, the dorsal and ventral subdivisions, are excitatory and inhibitory pathways, respectively, which mirrors actor-critic. The genesis of idea was in Montague et al (1996) [12], where the authors noted the similarities of the dopamine response in the brain to temporal difference learning methods. According to them, in the brain, the dopamine reward signals function such that a better-than-expected outcome results in a dopamine release, following the Rescorla-Wagner model. Actor-critic reinforcement learning schemes are composed of a policy (actor) and a value function (critic), such that the actor is updated with information from the critic.

A common animal to study for the purposes of learning in non-humans is the songbird. Songbirds learn by first hearing a song from a tutor, which is usually an older male in the colony. Afterwards, the bird encodes the state transitions in its brain, and then evaluates its performance syllable by syllable [5], which fits the paradigm of a Markov decision process.

In this paper, we evaluate the effectiveness of actor-critic reinforcement learning algorithms on an environment modeled after the way that songbirds learn.

## 2 RELATED WORK

According to a survey paper by Grondman et al[7] , the starting point for actor-critic algorithms is the Barto et al paper "Neuronlike adaptive elements that can solve difficult learning control problems" from 1983[1]. In that paper, Barto et al solve the cart pole problem using a combination of an associative search element (the actor) and an adaptive critic element (the critic). The authors suggest that this design is similar to the Rescorla-Wagner model of classical conditioning. In this model, reward is associated with how unexpected a reward is, such that, for example, dopamine is released in the brain upon a better than expected outcome, and successive iterations the same stimulus will result in a decrease or depletion of the dopamine reward for that stimulus.

Doya et al.[5] was one of the first papers to apply reinforcement learning to songbird song acquisition. The authors reported that their reinforcement learning model was able to match up with a real spectrogram of birdsong up to 90%. They treated the state space as continuous, with the syrinx (like the larynx in humans) producing a frequency and the hyperstriatum ventrale (HVC) determining its correctness when compared to the previously-learned state transitions. Mackevicius and Fee[10] investigated the state space of song learning and noted that, unlike a maze, getting past actions wrong does not affect future actions, a type of problem often called a "contextual bandit." Nakahara[13] calls the dopamine release in cases like this "a classic model-free signal," but notes that there is a divide between high-level model-based animal behavior and instinctive model-free dopamine responses.

While the first actor-critic methods were based on estimating value functions, actor-critic algorithms benefit greatly from the policy gradient theorem, as discussed in Konda  Tsiklis[8]. According to the policy gradient theorem, optimizing a parametrized policy such as actor-critic with respect to the expected reward guarantees convergence to at least a locally optimal policy. Given that many value function approximation methods have no convergence guarantees, this is a very positive attribute.

Given that deep neural networks are universal function approximators [9], it was only a matter of time before deep neural networks

,
,
.

were applied to actor-critic models. Some versions of actor-critic models use two neural networks, one for the actor and one for the critic, while other implementations share a single neural network. Cobbe et al.[4] argue that a shared network allows features to be shared, but note that a shared network can lead to competition between objectives, which is the basis of their Phasic Policy Gradient method (PPG), which separates the training into two stages, one that trains and one that selects features. They also show empirically that using a shared network in Proximal Policy Optimization (PPO) leads to better performance.

Proximal Policy Optimization (PPO)[17], is an extension of the actor-critic architecture that optimizes a surrogate objective function. It exhibits many of the benefits of Trust Region Policy Optimization (TRPO)[16], but with better time complexity. It uses a surrogate objective of a clipped probability ratio, which forms a lower bound on the policy's performance.

# 3 EXPERIMENTS

## 3.1 Base Conditions

We used a custom OpenAI Gym[2] environment to model the vocalizations. In this experiment, we considered the state space to be discrete, with a finite number of correct and incorrect notes, although it could have been modeled as a continuous state space as well like in Doya et al.[5]. The 12 tone equal temperament musical scale gives discrete values to notes, although frequency is a continuous measurement and there are an infinite number of notes in between each discrete note. In our implementation, the default length of a song was 4 distinct notes, with 5 potential incorrect notes. Given that songbirds start with an initial standard vocalization before beginning their song[10], our implementation reflected that as well.

We modeled the reward function after the reward prediction error (RPE) hypothesis. This hypothesis is based around the idea that our brain predicts the possibility of reward. If it gets an expected reward, a standard amount of dopamine is released. If it gets an unexpected reward, then the amount of dopamine released is greater than the standard. If it gets no reward, little or no dopamine is released. A more thorough treatment of this hypothesis is found in Glimcher[6]. In our implementation, we used +10 if a note was correct and expected to be incorrect, +5 if a note was correct and expected to be correct, and +0 if a note was incorrect and expected to be incorrect. This means that while a boost is given to new notes, the steady-state reward should be +5 per note in the song, or +20 for a 4-length motif. For reporting the rewards in training, we used a normalized reward in which all numbers were divided by this steady-state score.

## 3.2 Advantage Actor-Critic (A2C)

Following Mnih et al. [11], [3], we tested advantage actor-critic (A2C) on our environment. A2C is a synchronous version of the algorithm presented in the paper, and is a common algorithm to be present in libraries or as a student exercise in a graduate level reinforcement learning class. We used a custom shared-network implementation of A2C. The shared network is a feed forward neural network with a single hidden stage of dimension 512. The implementation was based off of the examples given by PyTorch [14].

## 3.3 Advantage Actor-Critic With Successive-Step Discounting

As has been mentioned already, the amount of dopamine released is smaller if a reward is expected. Therefore, getting a song correct multiple times in a row should result in a decrease in the dopamine secretion or decrease in reward. To model this, we changed our reward function to discount rewards even for correct action across successive episodes. We call this the advantage actor critic algorithm with successive-step discounting. We predicted this change would encourage more exploration, and the model would escape local minima more easily.

## 3.4 Proximal Policy Optimization

Proximal Policy Optimization[17], as described by the authors, is an improvement on A2C that uses a technique pulled from Trust Region methods, in which a surrogate objective function is maximized in addition to the main objective. To test Proximal Policy Optimization, we used the implementation found in Stable Baselines[15]. Given that the Stable Baselines implementation runs on any OpenAI Gym environment, this was not a difficult change.

# 4 RESULTS

PPO generally performed better than A2C, when they both converged. PPO would converge much quicker, but when neither method converged, PPO would be stuck at a lower local maximum. For a small song and a small number of erroneous notes, all of the methods performed well. We believe that this difference is because PPO needs less hyperparameter tuning, one of the reasons it is so popular.

Actor-critic with successive step discounting performed very similarly to vanilla A2C, however, when trained past convergence, the performance began to drop off. This is similar to a brain might work, except that songbirds are known to remember their songs for the rest of their lives once they've learned it to a sufficient degree.

# 5 CONCLUSION

We applied a variety of different actor-critic methods to songbird vocalization learning. Advantage actor-critic with successive step discounting models brain activity more closely but has negative performance implications. We found that PPO performed the best when both PPO and A2C converged, but when they did not, A2C generally lands on a higher average normalized reward. Given that a relatively modest state space caused both of these models not to converge, they would best be modified to use more exploration.

# ACKNOWLEDGMENTS

Thanks to Dr. Sharon for such a great class!

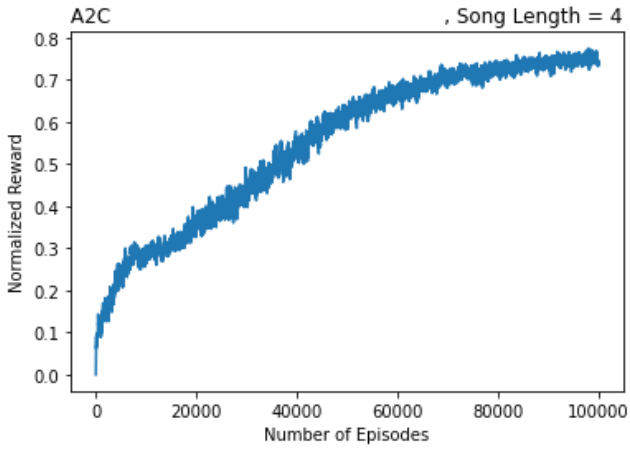


**Figure 1: Advantage Actor Critic with a length 4 song and 20 possible erroneous notes. This algorithm was able to learn the song.**

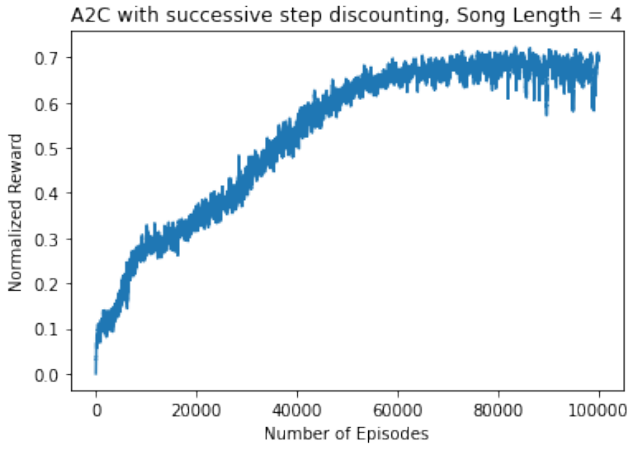


**Figure 2: Advantage Actor Critic with Successive Step Discounting with a length 4 song and 20 possible erroneous notes. This algorithm was able to learn the song.**

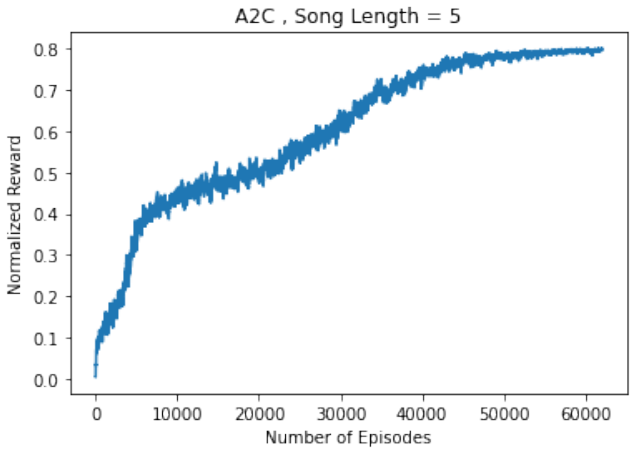


**Figure 3: Advantage Actor Critic with a length 5 song and 8 possible erroneous notes. This algorithm was able to learn the song.**

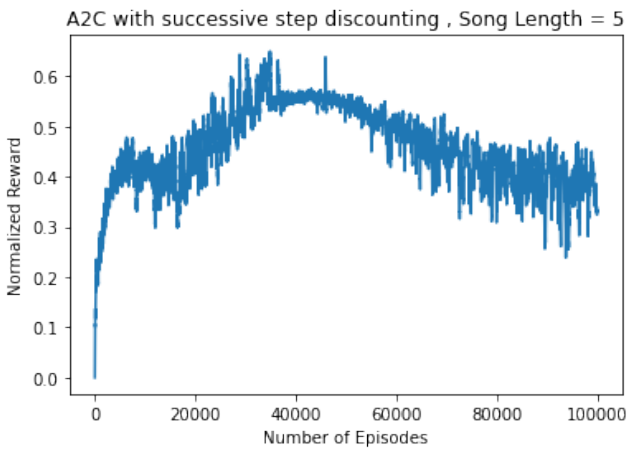


**Figure 4: Advantage Actor Critic with Successive Step Discounting with a length 5 song and 8 possible erroneous notes. This algorithm was able to learn the song.**

,
,



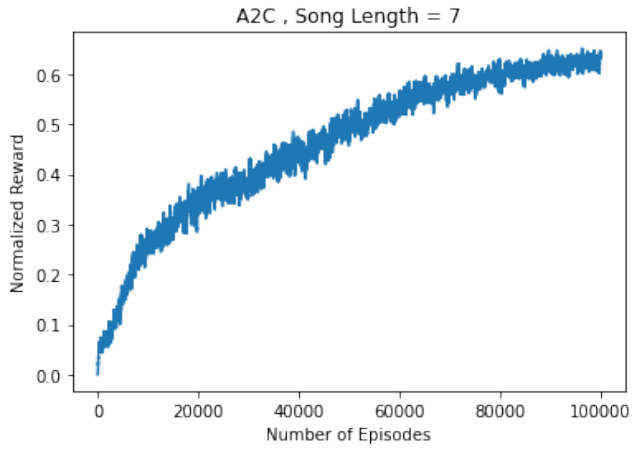


**Figure 5: Advantage Actor Critic with a length 7 song and 20 possible erroneous notes. This algorithm was able to learn the song.**

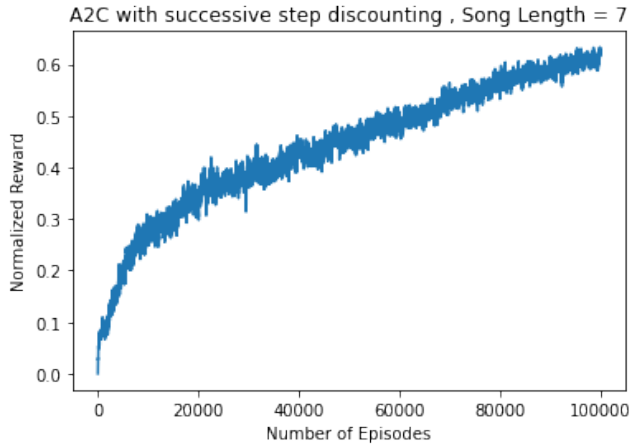


**Figure 6: Advantage Actor Critic with Successive Step Discounting with a length 7 song and 20 possible erroneous notes. This algorithm was able to learn the song.**

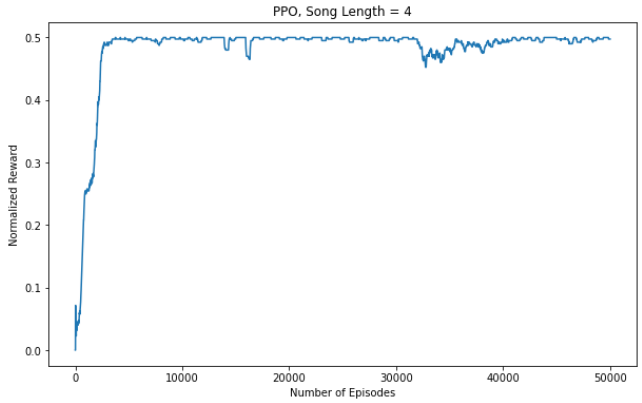


**Figure 7: PPO with a length 4 song and 20 possible erroneous notes. This algorithm was not able to learn the song.**

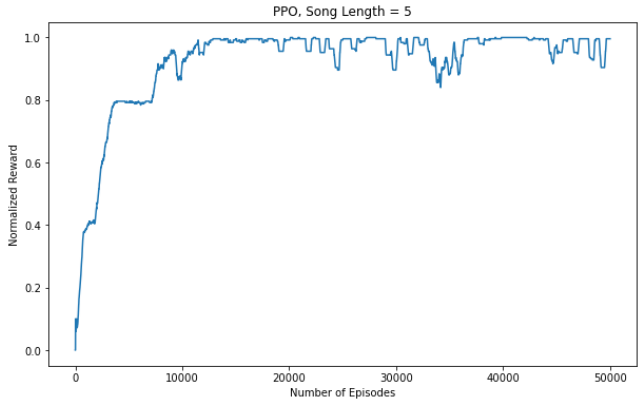


**Figure 8: PPO with a length 5 song and 8 possible erroneous notes. This algorithm was able to learn the song relatively quickly.**

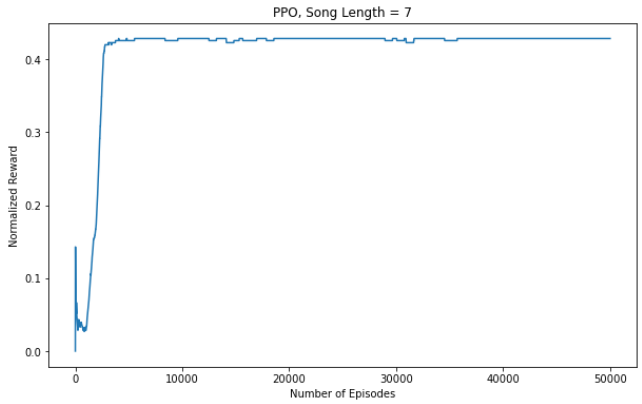


**Figure 9: PPO with a length 7 song and 20 possible erroneous notes. This algorithm was not able to learn the song.**

# REFERENCES


[1] Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. 1983. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13, 5 (Sept. 1983), 834–846. https://doi.org/10.1109/TSMC.1983.6313077

[2] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. https://doi.org/10.48550/ARXIV.1606.01540

[3] Ruidong Chen and Jesse H Goldberg. 2020. Actor-critic reinforcement learning in the songbird. *Current Opinion in Neurobiology* 65 (Dec. 2020), 1–9. https://doi.org/10.1016/j.conb.2020.08.005

[4] Karl Cobbe, Jacob Hilton, Oleg Klimov, and John Schulman. 2020. Phasic Policy Gradient. http://arxiv.org/abs/2009.04416 arXiv:2009.04416 [cs, stat].

[5] Kenji Doya and Terrence J Sejnowski. 1994. A Novel Reinforcement Model of Birdsong Vocalization Learning. In *Advances in Neural Information Processing Systems*, Vol. 7. MIT Press. https://proceedings.neurips.cc/paper/1994/hash/0a113ef6b61820daa5611c870ed8d5ee-Abstract.html

[6] Paul W. Glimcher. 2011. Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences* 108, supplement_3 (Sept. 2011), 15647–15654. https://doi.org/10.1073/pnas.1014269108

[7] Ivo Grondman, Lucian Busoniu, Gabriel A. D. Lopes, and Robert Babuska. 2012. A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (Nov. 2012), 1291–1307. https://doi.org/10.1109/TSMCC.2012.2218595

[8] Vijay Konda and John Tsitsiklis. 1999. Actor-Critic Algorithms. In *Advances in Neural Information Processing Systems*, Vol. 12. MIT Press. https://papers.nips.cc/paper/1999/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[9] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. 1993. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks* 6, 6 (Jan. 1993), 861–867. https://doi.org/10.1016/S0893-6080(05)80131-5

[10] Emily Lambert Mackevicius and Michale Sean Fee. 2018. Building a state space for song learning. *Current Opinion in Neurobiology* 49 (April 2018), 59–68. https://doi.org/10.1016/j.conb.2017.12.001

[11] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous Methods for Deep Reinforcement Learning. In *Proceedings of The 33rd International Conference on Machine Learning*. PMLR, 1928–1937. https://proceedings.mlr.press/v48/mniha16.html

[12] Pr Montague, P Dayan, and Tj Sejnowski. 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience* 16, 5 (March 1996), 1936–1947. https://doi.org/10.1523/JNEUROSCI.16-05-01936.1996

[13] Hiroyuki Nakahara. 2014. Multiplexing signals in reinforcement learning with internal models and dopamine. *Current Opinion in Neurobiology* 25 (April 2014), 123–129. https://doi.org/10.1016/j.conb.2014.01.001

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[15] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research* 22, 268 (2021), 1–8. http://jmlr.org/papers/v22/20-1364.html

[16] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. 2017. Trust Region Policy Optimization. http://arxiv.org/abs/1502.05477 arXiv:1502.05477 [cs].

[17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. http://arxiv.org/abs/1707.06347 arXiv:1707.06347 [cs].

[18] Shesharao M. Wanjerkhede, Raju S. Bapi, and Vithal D. Mytri. 2014. Reinforcement learning and dopamine in the striatum: A modeling perspective. *Neurocomputing* 138 (Aug. 2014), 27–40. https://doi.org/10.1016/j.neucom.2013.02.061