Original Research Article

Temporal Credit Assignment and Reward Granularity in a Songbird-Inspired Reinforcement Learning Model

Rohan Madhok

The Lawrenceville School, 2500 Main St, Lawrence Township, NJ 08648, United States

ABSTRACT

Sequential motor learning, such as precise imitation of birdsong syllables, depends on the brain's ability to link individual actions to delayed outcomes, a challenge known as the temporal credit assignment problem. This problem arises because feedback often arrives only after a sequence unfolds, obscuring which actions drive success or error. Inspired by songbird learning, this study investigates how reward-feedback frequency (granularity) within a sequential vocal imitation task affects learning efficiency in an actor-critic reinforcement learning agent. By systematically varying reward timing while keeping cumulative feedback constant, results show that finely grained, step-by-step feedback substantially accelerates learning and improves final imitation accuracy. In contrast, sparse feedback (delivered only at midpoint or endpoint) substantially impedes learning. Even when training was extended to 20,000 episodes, the end-only condition (K = 1) never reached the success criterion, and the half-only condition (K = 2) reached it in only a small fraction of seeds. This indicates a continuous relationship between feedback frequency and learning, rather than a sharp learnability cutoff. Notably, even under the densest feedback condition (reward after every action), performance plateaued at a non-zero error. Further analysis revealed that this plateau reflects not a single limit but two sources: the fixed exploration noise in the policy, and a residual imitation error that persists once the noise is removed. These findings characterize the trade-off between reward granularity and learning efficiency in a specific reinforcement learning model and task. They also suggest directions for future investigation of reward scheduling in biologically inspired learning systems.

Keywords: reinforcement learning; temporal credit assignment; reward granularity; songbird vocal learning; proximal policy optimization; reward prediction error; sequential motor learning

INTRODUCTION

Juvenile songbirds learn to imitate a 'tutor' song they hear early in life. In dominant models of song learning, the juvenile memorizes this song, and then must learn how to imitate it by driving its vocal system to produce the desired sound sequence (1). When they start to sing, they produce highly variable vocal output, akin to human babbling, analogous to generating random guesses for how to sing (1, 2). Gradually, over a few months, practicing thousands of songs per day, birds acquire imitations of the template. This process of trial-and-error practice, where they progressively shape their vocalization based on internally generated neurobiological feedback, may resemble aspects of human speech learning. Central to this learning process are dopamine neurons that encode reward prediction errors, signals reflecting the mismatch

Corresponding author: Rohan Madhok, E-mail: madhok.rohan@gmail.com.

Accepted June 3, 2026
https://doi.org/10.70251/HYJR2348.43376384

between predicted and actual song performance (3, 4). Importantly, dopaminergic error signals are generated in real time, lagging errors by about 50-60ms (4), suggesting a process of online (and not delayed) reward feedback, as illustrated in Figure 1 (right). These neural signals enable precise temporal credit assignment, guiding the bird to reinforce successful syllables and refine or discard incorrect ones (4). In essence, the songbird's brain operates analogously to an actor-critic reinforcement learning (RL) system (5), where specific neural circuits evaluate performance at each step and provide immediate feedback signals (6).

Specifically, song learning involves a sophisticated network that includes dopamine basal ganglia pathways, conserved across vertebrates, that are also important for human speech learning. The song-learning circuit relevant to this process is shown in Figure 1 (left): the brain region HVC generates precise timing signals, LMAN introduces exploratory variability, and RA executes motor commands for song production. Critically, Area X, a component of the songbird's basal ganglia, receives RPE signals from DA neurons in the ventral tegmental area (VTA), effectively performing the role of credit assignment (4, 6).

Thus, in biological systems, credit is assigned by breaking long sequences into smaller chunks that can be rewarded (6). Figure 2 illustrates the temporal credit-assignment problem that arises when feedback is delayed across a sequence. This study investigates this biologically inspired strategy of rapid evaluative feedback in an RL model. While it is established that dopaminergic error signals operate at short latencies in songbirds (4, 6) and that frequent reward generally helps reinforcement learning, the specific question of how the frequency of evaluative feedback — at constant total reward — affects sequential motor learning has not been systematically tested in a songbird-inspired framework. This study tests the hypothesis that more frequent feedback within a sequence improves learning efficiency, and that very sparse feedback may make credit assignment infeasible within practical training budgets. The specific 8-note tutor-song target used for training is shown in Figure 3.

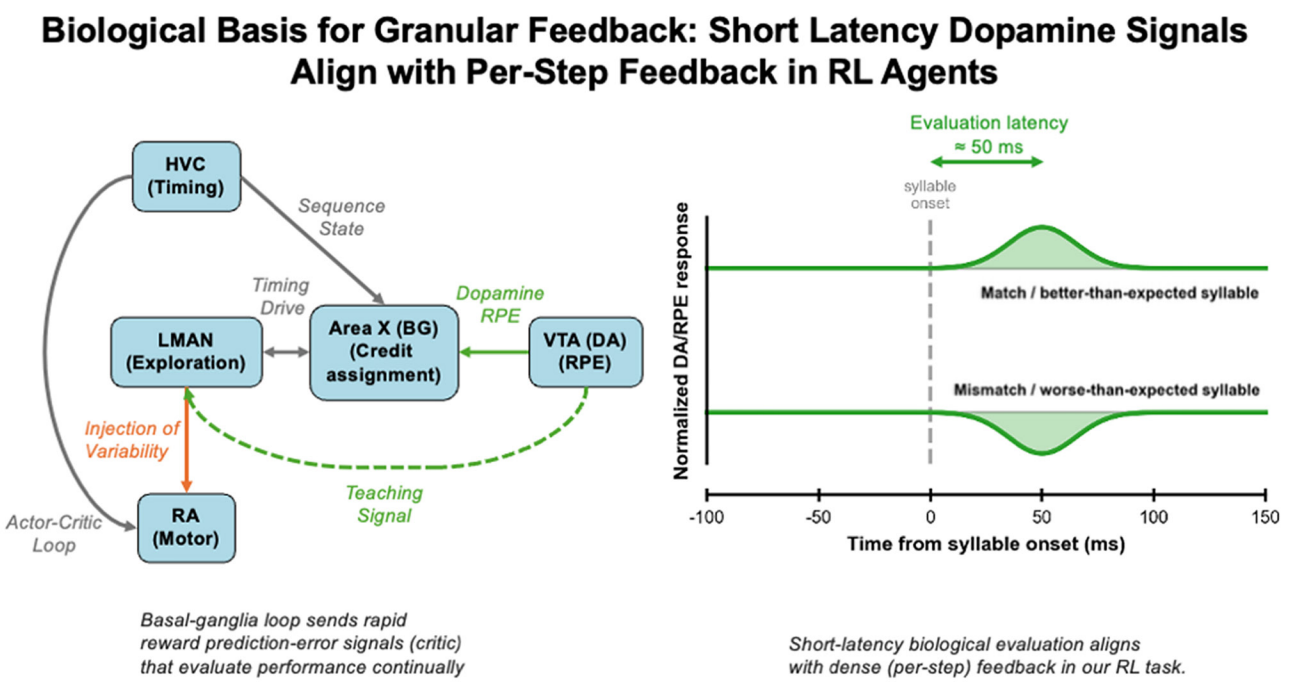

Figure 1. ***Biological basis for feedback in bird song learning.*** *Left: Simplified songbird circuit. HVC gives sequence timing; LMAN injects exploratory variability; RA projects directly to vocal motor neurons and executes motor commands; Area X (basal ganglia) undergoes reward (dopamine) modulated plasticity; VTA dopaminergic neurons send reward prediction error (RPE) signals back to Area X, which delivers a teaching signal to RA. Right: Dopamine RPEs have short latency, mirroring high feedback granularity (K).*

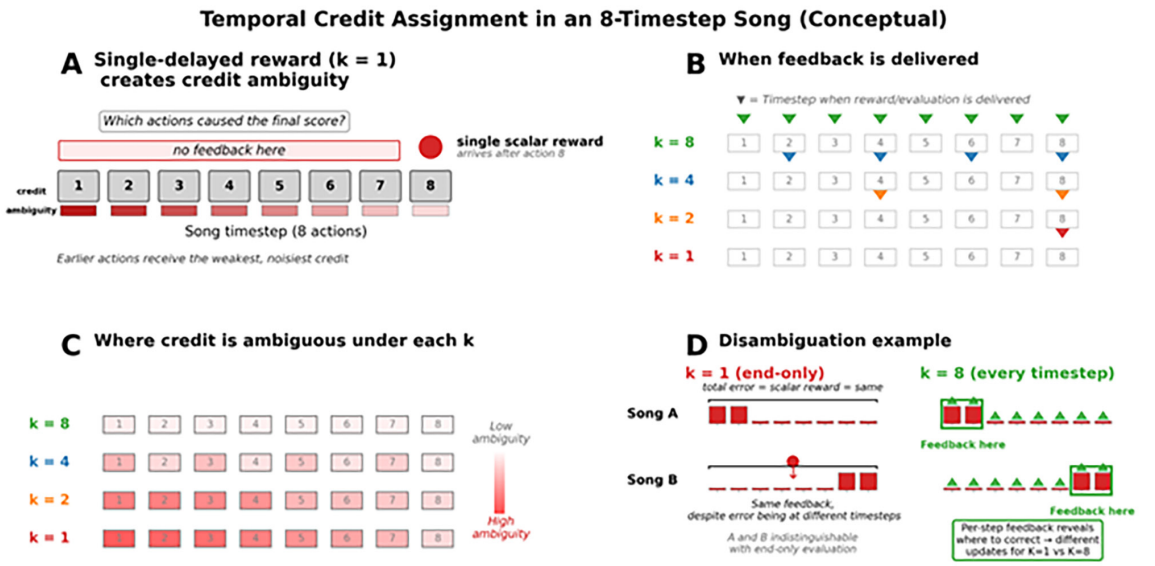

Figure 2. Temporal credit assignment diagram. *Conceptual illustration of how delayed reward feedback creates ambiguity in assigning credit to individual actions within an 8-step sequence. (A) Under sparse reward (K=1), the agent receives only one feedback signal after playing all 8 notes, resulting in high uncertainty about why actions led to success or failure. Early actions are particularly ambiguous because they receive weaker and noisier signals. (B) The timing of reward for varying feedback granularities (K=1,2,4,8). (C) Visualization of ambiguity across different granularities across different timesteps; higher K leads to reduced ambiguity. (D) Disambiguation example contrasting end-only feedback (K=1) vs. per-step feedback (K=8). Two distinct songs (A and B) produce identical total errors under sparse reward, making them indistinguishable. With dense feedback (K=8), the source of error is clearly identifiable, allowing targeted correction and efficient learning updates.*

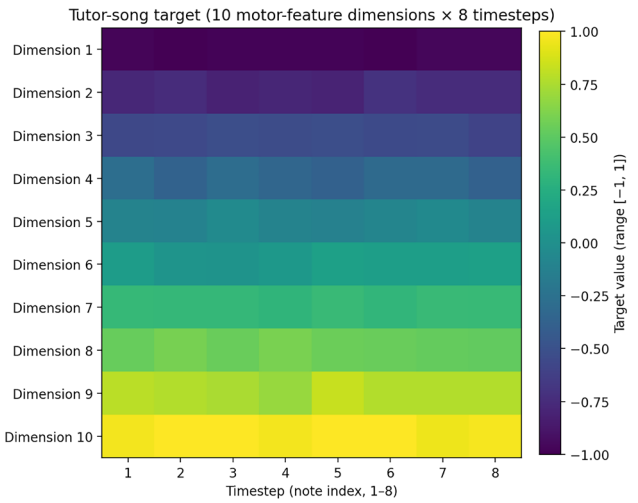

Figure 3. Tutor-song target. *The target song spans 8 timesteps, with each timestep represented by a 10-dimensional vector. The rows of the heatmap correspond to abstract motor-feature dimensions (labeled Dimension 1-10) rather than specific acoustic features. All values lie in the range [-1, 1]. The target is initialized as a linear ramp at the start of each run and undergoes slow drift during training: every 500 episodes, each entry is nudged by a small random value drawn uniformly from -0.03 to +0.03.*

METHODS AND MATERIALS

The agent's task is to imitate an 8-note target song, similar to a juvenile songbird learning a tutor song. The target song was initialized at the start of each run as a fixed 8-timestep, 10-dimensional pattern (each dimension covering the range from -1 to 1 in equal steps). To prevent the agent from memorizing one exact target, the target underwent slow drift during training: every 500 episodes, each entry of the target was nudged by a small random value drawn uniformly from (-0.03, +0.03), with values clipped to remain within [-1, 1]. The random number generator controlling drift was kept separate from the one used during training, so the target's trajectory could be reproduced exactly across runs. Each episode consists of 8 sequential notes, and the agent attempts to produce a sequence of sounds to match the target sequence. The reward granularity (K) was systematically varied to determine how often the agent receives feedback within the 8-note sequence. When K=1 (sparse reward), the agent only receives end-only feedback based on overall song imitation quality. When K=2, the agent receives two rewards, one halfway at note 4 and then one at the end. When K=4 (moderate reward), the agent receives rewards

at four evenly spaced points (after notes 2, 4, 6, and 8). Finally, when K=8 (dense reward), the agent receives immediate feedback after each note; thus, effectively providing continuous feedback. At each reward point, a negative error was computed over all notes since the previous reward (reward = - summed Euclidean distance between the agent's notes and the corresponding target notes in that segment). This ensured that the total reward across an episode was identical across all values of K, differing only in how frequently feedback was delivered. This means that the closer the agent's notes are to the corresponding target notes within a segment, the higher (less negative) the reward. This means that the manipulation isolates feedback timing rather than total evaluative signal. At each timestep, the agent observes a state consisting of two parts: a position indicator showing which timestep is currently active (a vector of 8 entries, with a 1 at the position corresponding to the current timestep and 0s elsewhere), and a 10-dimensional noise vector drawn from a small Gaussian (mean 0, standard deviation 0.1). Importantly, the tutor target itself is not part of the observation; the agent must infer the target from the reward signal alone over the course of training. This mirrors the biological setting, where a juvenile songbird internalizes the tutor song in memory early in life and then must rely on internal evaluative signals to refine its own production, rather than continuously comparing its output to an externally available target (1, 6). The action at each timestep is a 10-dimensional vector representing an abstract motor-feature vector. The 10 dimensions are not mapped to specific acoustic features, but correspond to an abstract space across which the agent must coordinate its output. Additionally, at the end of an episode, a small bonus (+1.0) is given if the entire song's cumulative error is below a threshold (near perfect imitation). This bonus is intended to simulate a dopamine surge when the juvenile bird sings the song correctly.

A feedforward actor-critic agent was trained using Proximal Policy Optimization (PPO). The architecture was kept deliberately simple: a feedforward network rather than a recurrent one isolates the credit-assignment problem from any confounds that learned temporal representations would introduce, since timing information is supplied externally through the state. The actor was a small neural network with a single hidden layer of 128 units (ReLU activation), producing a 10-dimensional mean action vector at each timestep. A hidden-layer width of 128 units was chosen as a standard size sufficient to represent the task without introducing capacity-related confounds. Actions were sampled from a Gaussian distribution centered on this mean, with a fixed standard deviation of 0.2 — a value chosen to provide enough exploration for the policy to escape local minima while remaining small enough to allow accurate imitation once the policy has learned — then clipped to the range [-1, 1] before being delivered to the environment. The critic had the same architecture as the actor and produced a single scalar value estimate for advantage computation. The position-indicator timestep input serves as an externally provided timing signal, mirroring the role of HVC in supplying sequence position to downstream motor regions in the songbird circuit (4, 6). The framework is designed to support biologically motivated extensions in future work; in the present study, the agent's learning rate and architecture are held constant across conditions, so that the effect of reward granularity can be isolated.

Each condition ($K \in \{1, 2, 4, 8\}$, song length $T = 8$ notes) was run with 10 independent random seeds. The training horizon was set to 20,000 episodes for the sparse conditions ($K = 1$ and $K = 2$), where slow convergence was anticipated. For the denser conditions ($K = 4$ and $K = 8$), training was halted at 5,000 episodes, after which the success criterion had been reached, and learning curves had stabilized. The agent's episode reward was continuously monitored to produce learning curves. Throughout the paper, performance is reported in reward units, where reward is the negative song-matching error; higher (less negative) reward indicates better imitation. Learning curves are shown as the mean with 95% bootstrap confidence intervals across 10 random-seed trials (10,000 resamples). All hyperparameters were held identical across conditions, so that differences in learning performance arose solely from the reward granularity. Complete hyperparameter values are listed in Table 1. A run was considered to have met the success criterion when the 50-episode running mean of episode reward first exceeded -12. To test whether the effect of reward granularity is specific to the 8-note baseline or generalizes to other sequence lengths, additional experiments were run with shorter songs ($T = 4$ notes) and longer songs ($T = 16$ notes). Because total episode reward scales with sequence length — longer sequences accumulate more error per episode — the success threshold of -12 was rescaled in these experiments to $-12 \times (T/8)$. This gives a threshold of -6 at $T = 4$ and -24 at $T = 16$, which keeps the per-note error budget at the criterion constant across all three sequence lengths and allows success rates to be directly compared. Results are reported under both the rescaled and the original -12

Table 1. *Training hyperparameters used across all conditions.*

Component	Setting
Algorithm	PPO with 1-step TD advantage
Actor architecture	MLP, 1 hidden layer of 128 units, ReLU
Critic architecture	MLP, 1 hidden layer of 128 units, ReLU
Action distribution	Diagonal Gaussian, fixed $\sigma = 0.2$
Action mean clamp	[-3, 3]
Action sample clip	[-1, 1]
Observation noise (state)	Gaussian, mean 0, standard deviation 0.1, 10 dimensions
Discount γ	0.99
PPO clip ε	0.2
Actor optimizer	Adam, learning rate 3×10^{-4}
Critic optimizer	Adam, learning rate 1×10^{-3}
PPO update epochs per batch	4
Batch size	1 episode
Training horizon (K = 1, K = 2)	20,000 episodes
Training horizon (K = 4, K = 8)	5,000 episodes
Tutor drift interval	500 episodes
Tutor drift magnitude	±0.03, uniform
Random seeds per condition	10

criterion, where applicable.

In the sequence-length experiment, conditions $K \in \{1, 2, 4\}$ were run at T = 4, and conditions $K \in \{1, 2, 4, 8\}$ were run at T = 16, each with 10 random seeds and 5,000 episodes per run. Combinations requiring a non-integer segment length L = T/K were excluded. Additionally, to distinguish exploration noise from policy error in the final performance plateau, each final checkpoint from the main experiment and the sequence-length experiment was loaded and re-evaluated for 100 episodes under a deterministic policy with the action sampling standard deviation set to zero.

Statistical comparisons for the main T = 8 experiment used the two-sided Mann-Whitney U test on per-seed mean reward over the final 100 training episodes. Raw p-values were Bonferroni-adjusted across the six pairwise comparisons among the four K conditions, and adjusted p-values are reported in the Results; significance was assessed against $\alpha = 0.05$. Effect sizes are reported as Cliff's δ with 95% bootstrap confidence intervals (10,000 resamples). Episodes-to-criterion are reported as means across successful seeds. The T = 4 and T = 16 sequence-length experiments were treated as robustness analyses and are summarized descriptively by seed-level convergence patterns under the rescaled success criterion. The deterministic-vs-stochastic per-note error comparison used the Wilcoxon signed-rank test on within-seed paired differences.

RESULTS

The reward granularity had a substantial impact on whether the agent succeeded in learning the task within the training budgets tested. The full learning trajectories for the four feedback-frequency conditions are shown in Figure 4.

Under sparse feedback (K = 1 and K = 2), the agent struggled to learn the task across the full 20,000-episode training horizon. The end-only condition (K = 1) never reached the success criterion across any of the 10 seeds, with a final mean reward of -22.64 (SD = 3.74). The half-only condition (K = 2) reached the criterion in only 2 of 10 seeds, with those successful seeds requiring an average of 7,818 ± 2,140 episodes; its final mean reward was -20.46 (SD = 3.90). Notably, K = 1 and K = 2 final reward distributions were not statistically distinguishable

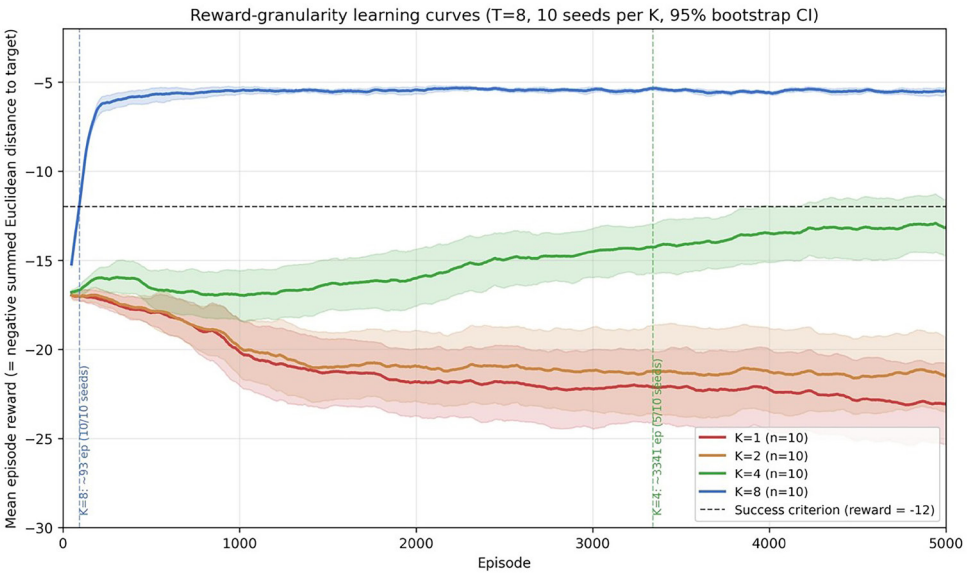

Figure 4. ***Learning curves of mean episode reward over training for each feedback frequency condition (K = 1, 2, 4, 8).*** *Reward is defined as the negative summed Euclidean distance between agent action and tutor target across the sequence; higher (less negative) reward indicates closer imitation. Curves are shown over the first 5,000 training episodes. Solid lines show the mean across 10 random-seed trials; shaded bands show 95% bootstrap confidence intervals (10,000 resamples). The horizontal dashed line at reward = -12 marks the success criterion (first crossing of the 50-episode running mean). Vertical annotations mark the mean episodes-to-criterion for the conditions that converged within this window: K = 8 at episode 93 (10 of 10 seeds) and K = 4 at episode 3,341 (5 of 10 seeds). K = 1 did not reach the criterion (0 of 10 seeds), and K = 2 reached it in only 2 of 10 seeds, converging later at approximately episode 7,818 (beyond the range shown; the sparse conditions K = 1 and K = 2 were trained for the full 20,000-episode horizon, see Methods).*

(Mann-Whitney U = 39, p = 0.43, Cliff's δ = -0.22). In contrast, more frequent feedback produced substantially more efficient learning. At K = 4, 5 of 10 seeds reached the criterion in an average of 3,341 ± 1,020 episodes; at K = 8, all 10 of 10 seeds reached the criterion in an average of just 93 ± 5 episodes. Bonferroni-adjusted p-values were significant for both adjacent-K comparisons: K = 2 vs. K = 4 (adjusted p = 0.0035, Cliff's δ = -0.92, large effect) and K = 4 vs. K = 8 (adjusted p = 0.0011, Cliff's δ = -1.00, large effect). The steepest transition between conditions is therefore localized between K = 2 and K = 4 under the conditions tested.

Learning curve dynamics: In the sparse reward conditions, the mean episode reward did not show steady improvement over training. For several K = 1 seeds, mean reward declined over the 20,000-episode horizon, consistent with policy drift in the absence of within-sequence credit signals. The two K = 2 seeds that eventually reached the criterion did so late in training, requiring an average of 7,818 episodes, indicating that learning under K = 2 is severely slowed rather than strictly impossible.

Plateau at dense reward: at K = 8, mean final reward converged at approximately -5.50 (SD = 0.33 across 10 seeds), a substantial reduction in error relative to the sparse-reward conditions, but still well above the theoretical maximum of zero. To investigate why this residual error remained, each final K = 8 model was re-evaluated with exploration noise turned off: the standard deviation of the action distribution was set to zero, so the model produced the same action every time, given the same state. Under this deterministic evaluation, the mean per-note error at K = 8 dropped from 0.685 (the value observed during training) to 0.483. For comparison, the expected error contribution from the fixed exploration noise alone is approximately 0.616 per note, so removing the noise accounts for most of the gap but not all of it. The difference between stochastic and deterministic per-note error was statistically significant (Wilcoxon signed-rank test, p = 0.002, median paired difference = -0.21). The plateau therefore has two components: a portion attributable to the constant exploration noise (which is no longer present under deterministic evaluation) and a smaller residual error that persists in the deterministic

case and reflects the limits of the policy itself.

Intermediate condition: the K = 4 condition occupied a middle ground between K = 2 and K = 8. Of 10 seeds, 5 reached the success criterion, but the successful seeds required, on average, 36 times more episodes than K = 8 to do so (3,341 versus 93). Learning efficiency under this scheme is therefore continuous rather than binary: as feedback density increases, both the success rate and the speed of convergence improve.

Sequence-length robustness: to test whether the effect of reward granularity generalizes beyond the 8-note baseline, the manipulation was replicated at shorter (T = 4) and longer (T = 16) sequence lengths (10 seeds per condition; Figure 5). At each sequence length, the density of feedback required for reliable success scaled with the horizon length. At T = 4, K = 4 (one reward per timestep) reached the rescaled success criterion in all 10 seeds, while K = 2 reached it in only 2 of 10 and K = 1 in 0 of 10. At T = 16, even K = 8 (one reward every two timesteps) was needed to reach the rescaled criterion in all 10 seeds; K = 4 reached it in only 3 of 10, and the sparser conditions in 0 of 10. Importantly, this means that the feedback frequency required for reliable learning increases as the sequence becomes longer, consistent with credit-assignment difficulty scaling with horizon length.

DISCUSSION

This study tested how the granularity of reward feedback affects learning of a sequential vocal imitation task in a biologically inspired RL framework. The results support the central hypothesis that reward granularity is a major factor in learning efficiency: dense feedback (K = 8) produced reliable, rapid convergence, while sparse feedback (K = 1 and K = 2) produced little to no convergence even with extended training. Notably, learning under K = 1 remained absent across 20,000 episodes, while learning under K = 2 was severely slowed but possible for a minority of seeds. This indicates that the effect of granularity is best characterized as a continuous difficulty curve rather than as an all-or-nothing learnability boundary. Breaking a long task into smaller rewarded segments substantially eases credit assignment within this model, supporting reward shaping as a concrete strategy in long-horizon tasks rather than reliance on more complex algorithms or architectures.

These findings have a parallel in biological systems. Songbirds likely solve the credit assignment challenge in part by leveraging short-latency dopaminergic feedback that signals partial progress, rather than waiting until the end of the song (4, 6). The present results are consistent with this strategy: in a songbird-inspired RL model, the timing of feedback alone — at constant cumulative

Figure 5 — Sequence-length ablation. T-scaled criterion = -12 × T/8.

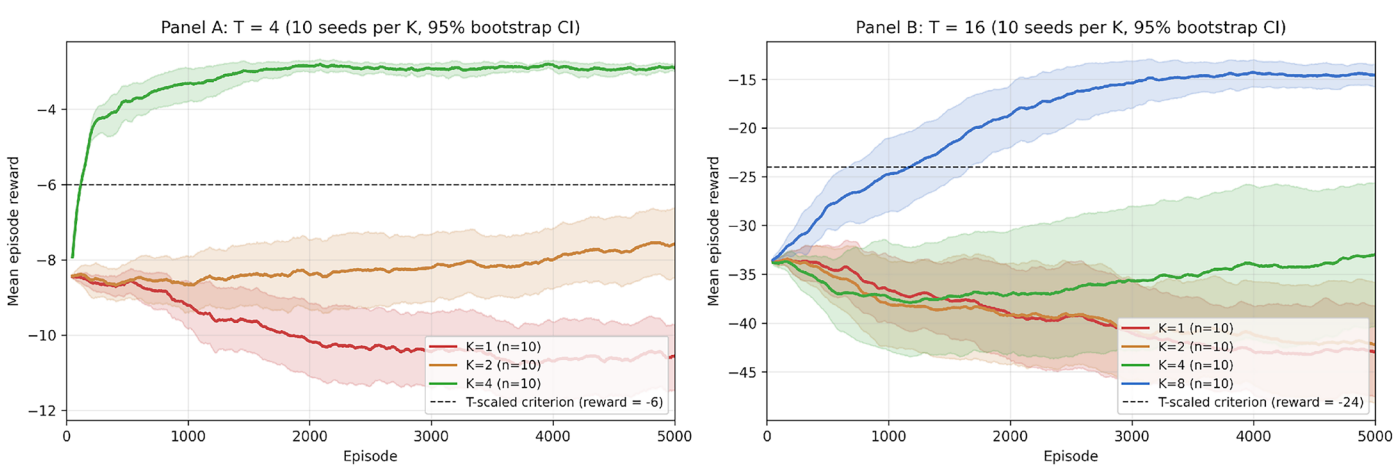

Figure 5. ***Learning curves for the sequence-length experiment.*** *Conditions at T = 4 (left) and T = 16 (right) are shown for varying K, with 10 seeds per condition. Solid lines show mean episode reward; shaded bands show 95% bootstrap confidence intervals. Horizontal dashed lines mark the success criteria rescaled for sequence length (-6 at T = 4 and -24 at T = 16), which hold the per-note error budget constant across T. The density of feedback required for reliable success scales with sequence length: K = 4 at T = 4 and K = 8 at T = 16 are the lowest densities that yielded 10 of 10 seed success at their respective rescaled criteria.*

reward — produced large differences in learning outcome. Taken together with the dopaminergic-timing data, this suggests that within-sequence credit signals may be a general feature of how biological systems learn intricate sequential behaviors.

The K = 8 plateau can be decomposed into two parts (Figure 6). The first is a contribution from the fixed exploration noise: throughout training, actions were sampled from a Gaussian with a standard deviation of 0.2, and this noise alone contributes roughly 0.62 per note in expected error. Under deterministic evaluation, where this noise contribution is removed, the average per-note error drops from 0.69 to 0.48. The second part is this residual error of about 0.48 per note that remains even after the noise is removed. This residual is consistent with a limitation of the feedforward policy in fine-tuning a 10-dimensional action based on a single scalar reward signal per segment — the kind of spatial credit assignment problem in which the agent must figure out which of its 10 output dimensions caused improvement or degradation (4, 6). This account is presented as a hypothesis: directly testing it is a promising direction for follow-up work, for example by providing per-dimension reward signals or by reducing the exploration noise during training. Alternative explanations, including limited network capacity or intrinsic stochasticity in the environment, also remain possible.

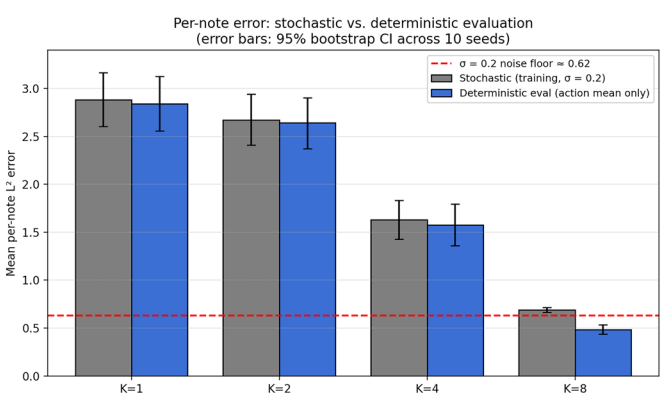

Figure 6. Decomposition of the K = 8 plateau. *Mean per-note L² error at the end of training for each K condition, under stochastic evaluation (training rollouts, fixed σ = 0.2) and deterministic evaluation (final checkpoints re-evaluated with σ = 0). Error bars show 95% bootstrap confidence intervals across 10 seeds per condition. The horizontal dashed line marks the analytic noise floor of approximately 0.62 per note, which is the expected per-note error contribution from the fixed exploration noise alone. At K = 8, the deterministic error (0.48) falls below the noise floor, indicating that the stochastic plateau is dominated by exploration noise; the residual deterministic error reflects the limits of the policy in fine-tuning the 10-dimensional action.*

Importantly, these results show that the relationship between feedback frequency and learning efficiency is continuous rather than binary, and that the amount of feedback required for reliable learning depends on the length of the sequence. The general principle that dense feedback aids reinforcement learning is well established; what these data add is a quantitative picture, showing that the sharpest change in performance occurs between K = 2 and K = 4 at the 8-note baseline, and that this transition shifts to higher densities for longer sequences. These observations may inform reward-shaping decisions in other sequential learning tasks where credit assignment is a known bottleneck.

The empirical findings of this study apply to the specific reinforcement learning model and task described above: a feedforward actor-critic agent trained on a fixed-length sequence imitation task with varying reward granularity. Within those bounds, granularity affected learning efficiency strongly and reproducibly, with the steepest transition between K = 2 and K = 4 at T = 8.

Beyond these specific findings, the results suggest several directions worth exploring further. In practical reinforcement learning, reward shaping informed by granularity considerations may speed up training in long-horizon tasks such as robotics or sequential decision-making. The dependence of learning on feedback frequency is also well established in educational psychology, and this work offers a computational perspective consistent with that body of research. In neurological rehabilitation, where motor or speech re-learning is often impaired by dopaminergic dysfunction, more frequent external cues during training may aid recovery, and direct empirical investigation in clinical and educational settings could test this.

CONCLUSION

In summary, this study demonstrates that the granularity of reward feedback is a primary factor in the learning efficiency of a sequential motor imitation task. Dense feedback produced reliable, rapid convergence; moderate feedback produced partial success at substantially greater episode cost; sparse feedback was severely impeded, with the most sparse condition failing to converge even at extended training. Furthermore, the density of feedback required for reliable convergence

scaled with the length of the sequence, with credit-assignment difficulty increasing for longer sequences. Together, these results support the broader strategy of integrating insights from biological systems into reinforcement learning design.

Future directions for this work include three main lines of investigation. First, implementing neuromodulated plasticity by allowing the agent to dynamically adjust its learning rate in response to a simulated dopamine-like signal, and testing whether such modulation reduces the residual policy error observed under K = 8. Second, isolating the spatial credit assignment component of the plateau by providing per-dimension reward signals rather than a single scalar reward per segment. Third, extending the framework to more complex tutor sequences or naturalistic acoustic features, to test whether the granularity-efficiency relationship observed here generalizes beyond the abstract feature space used in this study. By continuing to bridge neuroscience and reinforcement learning, this line of work may help clarify the principles of learning in biological systems and contribute to the development of more intelligent machines.

FUNDING SOURCES

The author received no external funding for this research.

CONFLICT OF INTEREST

The author declares that there are no conflicts of interest regarding the publication of this article.

ACKNOWLEDGEMENTS

The author is sincerely grateful to Professor Goldberg for the generous guidance, thoughtful feedback, and encouragement that shaped this project.

REFERENCES

1. Ölveczky BP, Andalman AS, Fee MS. Vocal Experimentation in the Juvenile Songbird Requires a Basal Ganglia Circuit. Schultz W, editor. *PLoS Biology*. 2005 Mar 29; 3 (5): e153. https://doi.org/10.1371/journal.pbio.0030153
2. Dhawale AK, Smith MA, Ölveczky BP. The Role of Variability in Motor Learning. *Annual Review of Neuroscience*. 2017 Jul 25; 40 (1): 479-98. https://doi.org/10.1146/annurev-neuro-072116-031548
3. Schultz W, Dayan P, Montague PR. A Neural Substrate of Prediction and Reward. *Science*. 1997 Mar 14; 275 (5306): 1593-9. https://doi.org/10.1126/science.275.5306.1593
4. Gadagkar V, Puzerey PA, Chen R, Baird-Daniel E, Farhang AR, Goldberg JH. Dopamine neurons encode performance error in singing birds. *Science [Internet]*. 2016 Dec 8; 354 (6317): 1278-82. Available from: http://science.sciencemag.org/content/354/6317/1278.full. https://doi.org/10.1126/science.aah6837
5. Chen R, Goldberg JH. Actor-critic reinforcement learning in the songbird. *Current Opinion in Neurobiology*. 2020 Dec; 65: 1-9. https://doi.org/10.1016/j.conb.2020.08.005
6. Charlesworth JD, Tumer EC, Warren TL, Brainard MS. Learning the microstructure of successful behavior. *Nature Neuroscience*. 2011 Jan 30; 14 (3): 373-80. https://doi.org/10.1038/nn.2748