nature portfolio

Peer Review File

Reinforcement-trained recurrent networks reproduce human beat-synchronization dynamics

Corresponding Author: Ms Yassaman Ommi

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors develop a recurrent neural network that is trained to synchronize “taps” in synchrony to a stimulus through a reinforcement learning paradigm. They find that in order for the model to learn the task and replicate features of human behavior, the reward function needs to have two key terms: a term that asymmetrically penalize late taps more than early taps and an interval matching term. The model is shown to reproduce important features of human behavioral and macaque neural data. Using the model, the authors also hypothesize that NMA, a signature feature of human behavior, results from perceptual error in estimating the exact time of the stimulus. Unlike previous models, this model addresses how the ability of synchronizing movements to stimuli might arise over development which is an important contribution. The paper is well written and the results are well placed in the context of what is known already. My comments are generally minor and can be found below.

Comments:
- Why is the interval reward formulated as it is? Why not just use the absolute value of -|IOI-ITI|? The 0.2 coefficient suggests that this is less important relative to the asymmetric reward term. Was this coefficient fit to something? How do your results change (mount of NMA, correction, convergence rate, etc.) if this coefficient is changed?

- From my understanding, there is no training the requires the model to produce the “taps” in the absence of the stimulus? If this was part of the training perhaps the model would then learn the IOI without needing to include the interval matching reward term explicitly?

- The authors make the prediction that domaine response magnitude would be modulated depending on whether the tap is early vs. late. I am a little confused if the observation that being late is penalized more than early is feature of biology (ie dopamine systems) or culture? Relatedly, why would dopamine be asymmetric for early vs. late for humans but not for other animals?

- I would be excited to know how the model learns to synchronize at the neural circuit level. Is the solution similar or different from mechanisms that have been proposed? Perhaps this is planned for future work, so I will just leave this comment as encouragement.

Reviewer #2

(Remarks to the Author)
This manuscript addresses what reinforcement (reward) structures could lead to the development of human-like sensorimotor synchronization—specifically, metronome tapping with predictive timing. The authors implement an LSTM-based recurrent network trained with a standard PPO strategy in a tapping task. They compare four reward schemes that combine (i) symmetric vs. asymmetric penalization of tap–cue asynchrony and (ii) an interval-matching term. They report that the asymmetric + interval scheme is most effective and yields several “human-like” signatures: asymmetric error correction

matching human phase correction data, emergence of negative mean asynchrony (NMA) under jitter/noise, and low-dimensional neural trajectories resembling macaque neural activity recordings (circular PCA trajectories with tempo-dependent scaling), plus a “period-2” alternation interpreted as subjective rhythmitization.

The overall idea—using deep RL to compare candidate reinforcement strategies for the development of sensorimotor synchronization—is potentially interesting to computational psychology/neuroscience researchers. However, several central claims currently read as over-generalized relative to what the simulations establish, and some key aspects of statistical validity and robustness are underspecified. Overall, I think it needs additional controls, clearer reporting, and more rigorous quantification in its current form. I list several comments as follows without specific order, and hope some of them will be helpful for the improvement of the manuscript.

Majors

1) As stated in the Abstract, the manuscript’s major claim is that “the most successful of these schemes rewards early taps more than late ones and additionally incentivizes tempo matching”. But the evidence currently shows one engineered reward combination (scheme) training an LSTM PPO agent to display certain signatures; it does not yet show that these ingredients are necessary, sufficient, or robust to alternative formulations. The authors might want to include more control simulations:
a) vary only the asymmetry while holding the rest fixed and similar for interval term weight/shape, missed-tap penalties etc.
b) Compare against at least one alternative plausible learning signal (e.g., reward on prediction error of cue timing; reward on phase error rather than absolute time).
Without these controls, the conclusion risks being an artifact of reward scaling and exploration dynamics rather than a principled insight about biological reinforcement.

2) The paper states “To ensure fair comparison across reward schemes, we conducted a guided hyperparameter search…” (Methods: “guided hyperparameter search…”) and later lists different final hyperparameters per agent (Table 3). This makes it hard to interpret failures (e.g., symmetric+interval “did not substantively improve”) as due to reward structure rather than optimization/tuning interactions. If each reward condition is allowed its own tuning budget, “success” partially reflects tuning success. Conversely, if tuning budgets were unequal or early-stopped, failures may be premature. Thus, the author might want to pre-register (or specify) a fixed tuning protocol and budget per scheme (number of trials, search space, selection criterion) and report distributions across many seeds with identical hyperparameters (or a small shared set), plus a sensitivity analysis showing results are stable across a reasonable hyperparameter neighborhood.

3) The manuscript states “Three random seeds were used per agent” (Methods), but most claims are demonstrated by example traces/figures and not summarized with uncertainty. Given the known instability of RL, three seeds are typically not enough to claim reliable emergence of subtle signatures like asymmetric phase correction, period-2 alternation, or NMA scaling. The authors might want to provide across-seed and across-training-run summary statistics (means, SDs) for key outcomes: synchronization accuracy, NMA under each jitter level, generalization performance, and neural trajectory metrics, etc. And use robust metrics rather than a demonstration of an example trial.

4) The manuscript highlights similarity to macaque MPC findings: “trajectories traced out circles… faster tempi producing smaller radius circles… taps occurring consistently in the same region…” (Results). But PCA trajectories from RNNs often appear cyclic in low dimensions for periodic tasks. The author might want to provide quantitative measures, for instance, circularity measure, radius scaling with IOI, phase-locking value of tap times to trajectory angle, and strength of period-2 component, etc.

5) The manuscript repeatedly uses terms like “quantitatively matches human data” (Abstract) and “closely matches” (Results; Fig. 7 discussion), but provides little formal quantification: no confidence intervals, no effect sizes, no seed-to-seed variability. For example, “Testing over a range of perturbation sizes showed…” and “quantitatively compared … with human data” (page 8) but the quantitative metric is not described.

6) Related to the above, statistically, this manuscript currently relies heavily on qualitative demonstrations, example traces, and overlays (e.g., Fig. 7 overlay onto human figure). Key claims (error correction asymmetry, NMA scaling, neural trajectory similarity) should be supported with formal quantification, uncertainty estimates, and seed-level replication.

Minors
1) In Reward scheme 2: “r=ttap−tnext” is described as penalizing asynchrony relative to next cue (Methods). Since rewards are “negative penalties, with 0 maximal reward,” this implies early taps (negative) are “better” than late taps (positive), but the text should explicitly state the optimization direction and clipping/normalization (e.g., is reward summed per tap? per timestep, per episode?).

2) The missed penalty is given as in Eq. 6. IOI is in seconds or ms? Elsewhere, IOI is in ms (400–700 ms). This needs unit consistency and a clear explanation of magnitude relative to asynchrony penalties.

3) The manuscript states that it excluded tempos between 500–600 ms, and performance was “indistinguishable” (Generalization section). Please specify: how many excluded tempi; what metric statistical test supported this claim.

4) Several places say “perfect” based on example trials (e.g., Asymmetric agent; Asymmetric+interval agent). Please report mean absolute asynchrony (and SD) across many trials and tempi.

5) Phrases like “provide evidence that similar reward-driven processes may shape synchronization circuits across species” (Asymmetric agent section) are a bit strong, given the abstraction level and lack of quantitative neural comparison.

6) Some inconsistencies: for instance, “IOI 400–700 ms” in page 2; later “450–750ms” training range” in page 6. Please clarify.

7) Figure quality needs to be improved. For instance, in Figure 5e, please denote the labels for each colored trace.

8) Some phrases are not consistent. For example, for the second reward scheme, “Asymmetric” and “Next cue” are used alternatively. Try to be consistent on one.

9) In equation 3, please define r_nearest. According to the context, it is referring to scheme 1, but better to spell it out.

Reviewer #3

(Remarks to the Author)
The paper entitled « Reinforcement-trained recurrent networks reproduce human beat-synchronization dynamics » by authors Yassaman Ommi, Matin Yousefabadi, and Jonathan Cannon presents a neural network model of auditory rhythms synchronization, foundational to music and dance.
The hypothesis is whether reward functions are required to develop tempi in a series of metronomic cues. Several experiments are done, with and without reward, on the beat, before after, or just before. Some results are consistent with human-like tendency to tap ahead of the cues, what is called negative match asynchrony (NMA).
A discussion is then made on biological plausibility, link to human infant and monkey sense of taps in audiomotor synchronization.

The paper is clearly written and well-organized. The introduction explains well the problematics and references are properly added. The figures and conclusion are well also. However, the method and experiments, and conclusion based on them can be argued partly.
For instance, the method section presents shortly the LSTM network. A better presentation of the parameters of the LSTM network should be presented, like the learning rates the difference between the different gates. Giving only the number of units is not enough to understand well the model.

Also, it is not clear how the information is encoded into the Neural Network: the binary metronome signal (0 or 1?) is given to the 128 units of the LSTM ? Does it means it is a 128 length binary vector ? What is the resolution also of the timesteps for the LSTM: is it in milliseconds ?

Moreover, embodiment is not explicitely taken into account and argued in the experiment section or the discussion. The system does not model a true sensorimotor network with identified sensor network and motor network. for instance, the delays due to action is not addressed.

This is important in the tapping experiment because we can observe some psychophysical features propr to human perception. Here is a text transcribed from paper 10.1080/03640210701801941:

(...) Gilden, Thornton, and Mallon (1995) asked people to tap each time they thought that a specific time interval had elapsed (in the range 0.3 sec–10 sec). They then calculated the temporal displacement from a “regular” rhythm from each successive interval.
These displacements showed correlations across a wide range of time scales—and indeed, the power spectrum of revealed a power law, which is characteristic of scale-invariance (the exponent of this power law was roughly –1, a noise-structure often observed in physical data; Handel & Chung, 1993).

please refer to these different papers.
it would be interesting to study or explain also of a power-law arise or not in your model.

In the neural modeling, the paper does not attempt also to model specifically the dopamine circuit of the basal ganglia with the reward functions. Some references to those bio-inspired models should be added.
Moreover, some dynamical models based on reservoir computing on this topic of time perception/generation have to be added like :
Laje and Buonomano, « Robust timing and motor patterns by taming chaos in recurrent neural networks ». Nature Neuroscience 2013
Y. Kawai, J. Park, I. Tsuda, M. Asada Learning long-term motor timing/patterns on an orthogonal basis in random neural networks Neural Netw., 163 (2023), pp. 298-311
Yuji Kawai, Takashi Morita, Jihoon Park, Minoru Asada, Oscillations enhance time-series prediction in reservoir computing with feedback, Neurocomputing, Volume 648, 2025

Could you describe the difference of your model and these ones? Why not have taken dynamical systems model that are dynamic?

Considering the four reward functions. They are well described.
One of them, the asymmetric model, appear to replicate well, the hebbian Spike-Timing Dependent Plasticity (STDP) learning mechanism. Could you describe the link/difference between your model and this one?

One strong argument against your model is the duration of the experiments which is very long >100 hours. These training times are they in line with human performance? Why choosing this time scale ? Is it possible to speed up learning convergence ?

It appears a strong drawback of the method on synchronization. Some analysis and models with more rapid phase synchronization require less duration.
The authors should give argument on that.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
My concerns have been sufficiently addressed.

Reviewer #2

(Remarks to the Author)
I thank the authors for their careful revision and detailed rebuttal. The revised manuscript is substantially improved relative to the original version. In particular, I appreciate the increase from three to ten random seeds per agent, the addition of seed-level summary statistics and uncertainty estimates, the clearer reporting of hyperparameter search procedures and shared-hyperparameter controls, and the expanded quantitative comparison with human data. The clarification of reward terms, units, missed-tap penalties, tempo range, and other methodological details also makes the manuscript much easier to evaluate.

Overall, the authors have addressed most of my previous concerns. The factorial comparison of asymmetric vs. symmetric reward and interval vs. no-interval reward is now described more explicitly, and the added across-seed statistics provide stronger evidence that the main behavioral patterns are not due to single-run examples. The revised NMA analysis, using tapping variability rather than the less robust perceptual-noise comparison, is also more convincing. I also appreciate that the authors moderated some of the stronger claims and added appropriate caveats regarding the qualitative nature of the neural-trajectory comparison.

I have only a few remaining minor comments listed below.

1，Although the language has been improved, some phrases such as “qualitatively and quantitatively” or “closely resembles human data” may still sound somewhat stronger than the analyses warrant, especially where the comparison is primarily based on overlap with human variability rather than formal model comparison. I suggest slightly softening these statements, e.g., “falls within the range of human data” or “captures key qualitative and quantitative features.”

2，The authors now appropriately acknowledge that cyclic PCA trajectories are expected in periodic tasks. Still, the text could make it even clearer that the neural comparison is mainly descriptive and hypothesis-generating, rather than a strong mechanistic match to macaque data.

3，Please check for a few remaining typographical or formatting issues in the revised manuscript and figures, including some figure-label readability and minor wording issues such as “would have the option of continue to tap.” These do not affect the scientific conclusions but should be cleaned up before publication.

Reviewer #3

(Remarks to the Author)
The authors corrected and improved the manuscript accordingly to the reviewers' questions. I have no other questions.

Author remarks: We thank the reviewers for their careful reads, thoughtful feedback, and interesting ideas! In addition to addressing their concerns below, we opted to change the following:

- **We changed one of the agent/human comparisons, replacing human results based on a not-so-robust correlation between "perceptual noise" and NMA with results based on a more robust correlation between tapping noise and NMA, as well as a new discussion of the justification and limitations of this approach. We also added various details about our reanalysis of data to Methods as requested by the revision table.**
- **We augmented the other human/agent comparison with the best fit line from the comparison paper and with an additional agent trained with jitter.**
- **We removed one agent's "stochastic policy" test, which added yet another dimension along which we could systematically vary all the agents but which we felt did not add to the story (and would have required another 10 days to train across 4 agents and 10 seeds with our current setup).**
- **We reordered out two comparisons with human data – asymmetric error correction test and NMA test – for better flow.**

one of the agent/human comparisons, replacing human results based on a not-so-robust correlation between "perceptual noise" and NMA with results based on a more robust correlation between tapping noise and NMA, as well as a new discussion of the limitations of this approach. We augmented the other human/agent comparison with the best fit line from the comparison paper and with an additional agent trained with jitter. Guided by the journal, we have also refined our statistical methods for analyzing the human data. We also removed one agent's "stochastic policy" test, which added yet another dimension along which we could systematically vary agents but which we believe did not add to the story.

All substantively new text is written in blue.

Reviewer #1 (Remarks to the Author):
The authors develop a recurrent neural network that is trained to synchronize "taps" in synchrony to a stimulus through a reinforcement learning paradigm. They find that in order for the model to learn the task and replicate features of human behavior, the reward function needs to have two key terms: a term that asymmetrically penalize late taps more than early taps and an interval matching term. The model is shown to reproduce important features of human behavioral and macaque neural data. Using the model, the authors also hypothesize that NMA, a signature feature of human behavior, results from perceptual error in estimating the exact time of the stimulus. Unlike previous models, this model addresses how the ability of synchronizing movements to stimuli might arise over development which is an important contribution. The paper is well written and the results are well placed in the context of what is known already.

Thank you!

My comments are generally minor and can be found below.

Comments:

Why is the interval reward formulated as it is? Why not just use the absolute value of -|IOI-ITI|? The 0.2 coefficient suggests that this is less important relative to the asymmetric reward term. Was this coefficient fit to something? How do your results change (mount of NMA, correction, convergence rate, etc.) if this coefficient is changed?

Author's response: We thank the reviewer for these questions. The interval term uses a squared deviation rather than the absolute deviation because the quadratic form is gentle near the target while penalizing large interval errors more steeply, which produced smoother and more stable learning in our experiments than the absolute-value form; the cap prevents large interval errors early in training from dominating the gradient.

The 0.2 coefficient was not fit to any external dataset. It was chosen during development because it improved training stability and convergence by balancing the interval term against the dominant asynchrony term, so that interval matching guides learning without overwhelming the primary synchrony objective. We did not perform a systematic sweep of this coefficient, and we now describe it in the Methods as a fixed implementation choice rather than a tuned parameter. We agree that characterizing how the amount of NMA, error correction, and convergence rate vary with this coefficient is an interesting question, and we note it as a worthwhile direction for future systematic analysis.

From my understanding, there is no training the requires the model to produce the "taps" in the absence of the stimulus? If this was part of the training perhaps the model would then learn the IOI without needing to include the interval matching reward term explicitly?

Authors' response: This is an insightful observation. Requiring the agent to continue tapping in the absence of the metronome is indeed formally equivalent to providing an interval-matching reward: in both cases, the agent must internalize and reproduce the inter-onset interval independently of incoming sensory cues. In a continuation task the agent is evaluated on how accurately it sustains the learned IOI once the metronome is silenced—exactly what an interval-matching reward incentivizes during training. To make this connection explicit, we added a synchronization–continuation experiment to the revised manuscript (Section 3.5), in which the Asymmetric+Interval agent is tested on a continuation phase without audible cues. The agent maintains the beat with a mean ITI error of 9.49 ± 0.26 ms, confirming that interval-matching reward is the functional equivalent of training with a silent-phase continuation requirement.

We could indeed train the model first to produce rhythms at cued tempos by rewarding it for matching a desired interval. Interval-based rewards would precede asynchrony-based

rewards rather than coming at the same time. This would be a developmentally reasonable approach to curriculum training, and might reduce the need for tempo matching per se. But it would have the added complication of a cue that is subsequently shut off for synchronization. We intend to try this out in future work, and we now mention it under Future Directions. Thanks for the interesting idea!

The authors make the prediction that domaine response magnitude would be modulated depending on whether the tap is early vs. late. I am a little confused if the observation that being late is penalized more than early is feature of biology (ie dopamine systems) or culture? Relatedly, why would dopamine be asymmetric for early vs. late for humans but not for other animals?

Authors' response: We hypothesize that the asymmetric penalty is not unique to humans—on the contrary, we propose that asymmetric dopaminergic reinforcement of predictive actions is a general principle of vertebrate reward learning. Organisms that act in anticipation of a sensory event (e.g., lunging for prey as it prepares to flee) gain a survival advantage over those that react after the fact; a reward signal that more strongly penalizes late responses than early ones would therefore be selected for across species. What may distinguish humans (and potentially other musical species) is the additional interval-matching component: the capacity and motivation to reproduce a precise inter-beat interval. Under our account, the asymmetric reward term reflects a phylogenetically ancient bias toward predictive action, while the interval-matching term captures a more culturally or developmentally acquired sensitivity to metric regularity. We have added a sentence to the Discussion clarifying this distinction.

I would be excited to know how the model learns to synchronize at the neural circuit level. Is the solution similar or different from mechanisms that have been proposed? Perhaps this is planned for future work, so I will just leave this comment as encouragement.

Authors' response: We are grateful for this encouragement. Characterizing the computational mechanism by which the LSTM implements predictive synchronization—and comparing it to proposed circuit-level accounts such as oscillatory entrainment or forward-model-based timing—is an exciting direction we intend to pursue in future work.

Reviewer #2 (Remarks to the Author):
This manuscript addresses what reinforcement (reward) structures could lead to the development of human-like sensorimotor synchronization—specifically, metronome tapping with predictive timing. The authors implement an LSTM-based recurrent network trained with a standard PPO strategy in a tapping task. They compare four reward schemes that combine (i) symmetric vs. asymmetric penalization of tap–cue asynchrony and (ii) an interval-matching term. They report that the asymmetric + interval scheme is most effective and yields several "human-like" signatures: asymmetric error correction matching human phase correction data, emergence of negative mean asynchrony (NMA) under jitter/noise, and low-dimensional neural

trajectories resembling macaque neural activity recordings (circular PCA trajectories with tempo-dependent scaling), plus a "period-2" alternation interpreted as subjective rhythmitization.

The overall idea—using deep RL to compare candidate reinforcement strategies for the development of sensorimotor synchronization—is potentially interesting to computational psychology/neuroscience researchers. However, several central claims currently read as over-generalized relative to what the simulations establish, and some key aspects of statistical validity and robustness are underspecified. Overall, I think it needs additional controls, clearer reporting, and more rigorous quantification in its current form. I list several comments as follows without specific order, and hope some of them will be helpful for the improvement of the manuscript.

Majors

1) As stated in the Abstract, the manuscript's major claim is that "the most successful of these schemes rewards early taps more than late ones and additionally incentivizes tempo matching". But the evidence currently shows one engineered reward combination (scheme) training an LSTM PPO agent to display certain signatures; it does not yet show that these ingredients are necessary, sufficient, or robust to alternative formulations. The authors might want to include more control simulations:
a) vary only the asymmetry while holding the rest fixed and similar for interval term weight/shape, missed-tap penalties etc.
b) Compare against at least one alternative plausible learning signal (e.g., reward on prediction error of cue timing; reward on phase error rather than absolute time).
Without these controls, the conclusion risks being an artifact of reward scaling and exploration dynamics rather than a principled insight about biological reinforcement.

Authors' response: We appreciate this concern and would like to clarify that our experimental design already constitutes a systematic factorial comparison. We trained four agents corresponding to the full 2×2 crossing of (i) symmetric vs. asymmetric asynchrony penalty and (ii) absence vs. presence of an interval-matching term—precisely the manipulation of "asymmetry alone" and "interval term alone" that the reviewer requests. To ensure that no agent was disadvantaged by tuning, hyperparameters were optimized separately for each configuration using an identical search space, budget, and selection criterion (Optuna/TPE; see response to Major Comment 2), and each agent was evaluated across 10 random seeds. As a further control, we additionally retrained every agent under the hyperparameters of the most successful agent and found the qualitative outcomes unchanged. This design allows us to attribute differences in performance and human-likeness to the reward structure rather than to agent-specific tuning. We have revised the Methods and Results sections to make this factorial structure more explicit and to report aggregate statistics across all seeds, so that the conclusions rest on replication rather than example traces.

2) The paper states "To ensure fair comparison across reward schemes, we conducted a guided hyperparameter search…" (Methods: "guided hyperparameter search…") and later lists different final hyperparameters per agent (Table 3). This makes it hard to interpret failures (e.g., symmetric+interval "did not substantively improve") as due to reward structure rather than optimization/tuning interactions. If each reward condition is allowed its own tuning budget, "success" partially reflects tuning success. Conversely, if tuning budgets were unequal or early-stopped, failures may be premature. Thus, the author might want to pre-register (or specify) a fixed tuning protocol and budget per scheme (number of trials, search space, selection criterion) and report distributions across many seeds with identical hyperparameters (or a small shared set), plus a sensitivity analysis showing results are stable across a reasonable hyperparameter neighborhood.

Authors' response: We agree this needed to be stated more precisely, and we have clarified our approach. Each reward scheme was optimized with the same tuning protocol: the same hyperparameter search (Optuna, using a Tree-structured Parzen Estimator) over the same search space (learning rate, batch size, entropy coefficient, and episode horizon) and the same selection criterion (mean evaluation reward), with the resulting per-agent settings reported in Supplementary Table 3. Because every scheme is tuned under an identical protocol, differences in outcome cannot be attributed to one agent being tuned more aggressively than another, nor can a failure such as that of the Symmetric+Interval agent be dismissed as premature. To additionally address the suggestion of results under identical hyperparameters, we then retrained all four agents under a single shared set (those of the best-performing Asymmetric+Interval agent); the qualitative pattern was unchanged, indicating that our conclusions are stable across the hyperparameter neighborhood. Finally, we increased training to 10 random seeds per agent and now report means and standard deviations for all key metrics.

3) The manuscript states "Three random seeds were used per agent" (Methods), but most claims are demonstrated by example traces/figures and not summarized with uncertainty. Given the known instability of RL, three seeds are typically not enough to claim reliable emergence of subtle signatures like asymmetric phase correction, period-2 alternation, or NMA scaling. The authors might want to provide across-seed and across-training-run summary statistics (means, SDs) for key outcomes: synchronization accuracy, NMA under each jitter level, generalization performance, and neural trajectory metrics, etc. And use robust metrics rather than a demonstration of an example trial.

Authors' response: We thank the reviewer for this important methodological point. In the revised manuscript we have increased the number of random seeds from 3 to 10 per agent. We now report mean ± SD across seeds for the primary quantitative outcomes—final cumulative reward, negative mean asynchrony (NMA), and mean absolute asynchrony—in Table 4. Figure 10 additionally visualizes the seed-level distribution of NMA as a function of jitter magnitude, providing a direct view of cross-seed variability. We have also replaced or supplemented several qualitative figure panels with summary statistics where feasible.

4) The manuscript highlights similarity to macaque MPC findings: "trajectories traced out circles… faster tempi producing smaller radius circles… taps occurring consistently in the same region…" (Results). But PCA trajectories from RNNs often appear cyclic in low dimensions for periodic tasks. The author might want to provide quantitative measures, for instance, circularity measure, radius scaling with IOI, phase-locking value of tap times to trajectory angle, and strength of period-2 component, etc.

Authors' response: We can see why the reviewer would be unsatisfied with our loose characterization of the dynamics – looping trajectories are indeed to be expected of a task requiring periodic action. However, the only appropriate measures in the Betancourt manuscript take the form of comparing how velocity vs. amplitude scales with trajectory at every point in the tapping cycle. Defining these measures would take a fair bit of space, would add complexity, and (we believe) would not add to our main point, which is that very basic features of the neural behaviour (specifically, amplitude scaling with tempo) resemble the neural data. But we now appropriately disclaim our comparison in the Discussion, noting that cyclic trajectories are not surprising given the periodic structure of the task.

5) The manuscript repeatedly uses terms like "quantitatively matches human data" (Abstract) and "closely matches" (Results; Fig. 7 discussion), but provides little formal quantification: no confidence intervals, no effect sizes, no seed-to-seed variability. For example, "Testing over a range of perturbation sizes showed…" and "quantitatively compared … with human data" (page 8) but the quantitative metric is not described.

Authors' response: We thank the reviewer for highlighting the need for more rigorous uncertainty quantification. In the revised manuscript, we have added confidence intervals to our simulation metrics in the form of seed-level summary statistics (mean ± SD across 10 seeds, including new measures of the four reward schemes reported in Table 4 and Figure 10. However, we note that the point of the simulation/human comparison is simply to show that the metrics lie within the range of human variability, which can be readily observed in Figures 7 and 9; we do not believe any further formal comparison between simulation and humans is necessary to support this point. We have altered our language throughout to make the nature of this point more clear.

6) Related to the above, statistically, this manuscript currently relies heavily on qualitative demonstrations, example traces, and overlays (e.g., Fig. 7 overlay onto human figure). Key claims (error correction asymmetry, NMA scaling, neural trajectory similarity) should be supported with formal quantification, uncertainty estimates, and seed-level replication.

Authors' response: As discussed above, we have added seed-level replication and uncertainty. Actual statistical comparison with the human data is not important to make the points we are trying to make, which are that 1) qualitative aspects of the human data are reproduced, and 2) simulation results fall quantitatively within the range of human

variability. Both of these points are easily read from the graphs in Fig 7 and 9. We have changed the language where appropriate to be clear that this is the scope of our claims.

Minors

1) In Reward scheme 2: "r=ttap−tnext" is described as penalizing asynchrony relative to next cue (Methods). Since rewards are "negative penalties, with 0 maximal reward," this implies early taps (negative) are "better" than late taps (positive), but the text should explicitly state the optimization direction and clipping/normalization (e.g., is reward summed per tap? per timestep, per episode?)

Authors' response: The optimization direction and accumulation are governed by the PPO algorithm, which maximizes expected cumulative episode reward; since all rewards are non-positive, the optimum of 0 corresponds to perfect synchronization. We believe this is sufficiently well specified in the text.

2) The missed penalty is given as in Eq. 6. IOI is in seconds or ms? Elsewhere, IOI is in ms (400–700 ms). This needs unit consistency and a clear explanation of magnitude relative to asynchrony penalties.

Authors' response: Thanks for noting this error. We now specify all of the time/interval inputs to the reward calculation as seconds. We also now discuss the relative magnitudes of the reward components.

3) The manuscript states that it excluded tempos between 500–600 ms, and performance was "indistinguishable" (Generalization section). Please specify: how many excluded tempi; what metric statistical test supported this claim.

Authors' response: We now specify in the text that this is 11 tempi left out, and we share reward and mean asynchrony metrics from the agent trained in this way.

4) Several places say "perfect" based on example trials (e.g., Asymmetric agent; Asymmetric+interval agent). Please report mean absolute asynchrony (and SD) across many trials and tempi.

We now widely include averages over trials, tempi, and seeds.

5) Phrases like "provide evidence that similar reward-driven processes may shape synchronization circuits across species" (Asymmetric agent section) are a bit strong, given the abstraction level and lack of quantitative neural comparison.
That was indeed an overstatement. We have cut that comment and tried to root out other similar ones.

6) Some inconsistencies: for instance, "IOI 400–700 ms" in page 2; later "450–750ms" training range" in page 6. Please clarify.
Authors' response: The correct figure is 450-750. This has been updated throughout.

7) Figure quality needs to be improved. For instance, in Figure 5e, please denote the labels for each colored trace.
We have made sure the legends for the neural trajectories are clearly visible.

8) Some phrases are not consistent. For example, for the second reward scheme, "Asymmetric" and "Next cue" are used alternatively. Try to be consistent on one.
We have replaced "next cue" with "asymmetric" and "nearest cue" with "symmetric" throughout.

9) In equation 3, please define r_nearest. According to the context, it is referring to scheme 1, but better to spell it out.
We now name the different reward components more explicitly and consistently.

Reviewer #3 (Remarks to the Author):
The paper entitled « Reinforcement-trained recurrent networks reproduce human beat-synchronization dynamics » by authors Yassaman Ommi, Matin Yousefabadi, and Jonathan Cannon presents a neural network model of auditory rhythms synchronization, foundational to music and dance.
The hypothesis is whether reward functions are required to develop tempi in a series of metronomic cues. Several experiments are done, with and without reward, on the beat, before after, or just before. Some results are consistent with human-like tendency to tap ahead of the cues, what is called negative match asynchrony (NMA).
A discussion is then made on biological plausibility, link to human infant and monkey sense of taps in audiomotor synchronization.

The paper is clearly written and well-organized. The introduction explains well the problematics and references are properly added. The figures and conclusion are well also. However, the method and experiments, and conclusion based on them can be argued partly.
For instance, the method section presents shortly the LSTM network. A better presentation of the parameters of the LSTM network should be presented, like the learning rates the difference between the different gates. Giving only the number of units is not enough to understand well the model.

Authors' response: We thank the reviewer for this suggestion and have expanded the description of the LSTM architecture in the Methods section. We now report the number of layers (2), the hidden state size (128 units per layer), the learning rate (3×10^{-4}), and the PPO-specific hyperparameters (GAE $\lambda = 0.95$, discount $\gamma = 0.99$, 10 optimization epochs per rollout, batch size 32). We also briefly describe the standard gating structure of the LSTM (input, forget, output, and cell gates) and cite Hochreiter & Schmidhuber (1997) for full architectural details.

Also, it is not clear how the information is encoded into the Neural Network: the binary metronome signal (0 or 1?) is given to the 128 units of the LSTM? Does it means it is a 128 length binary vector? What is the resolution also of the timesteps for the LSTM: is it in milliseconds?

Authors' response: We apologize for the ambiguity. The metronome signal is a single scalar value (1 at the onset of a beat, 0 otherwise), and this same scalar is passed as a one-dimensional input to the LSTM at every timestep—it is not a 128-dimensional vector. The simulation runs at a temporal resolution of Δt = 10 ms per timestep, so each LSTM step corresponds to 10 ms of simulated time. We have added a clarifying sentence to the Methods section specifying the input dimensionality and temporal resolution.

Moreover, embodiment is not explicitely taken into account and argued in the experiment section or the discussion. The system does not model a true sensorimotor network with identified sensor network and motor network. For instance, the delays due to action is not addressed.

This is important in the tapping experiment because we can observe some psychophysical features proper to human perception. Here is a text transcribed from paper 10.1080/03640210701801941:

(...) Gilden, Thornton, and Mallon (1995) asked people to tap each time they thought that a specific time interval had elapsed (in the range 0.3 sec–10 sec). They then calculated the temporal displacement from a "regular" rhythm from each successive interval. These displacements showed correlations across a wide range of time scales—and indeed, the power spectrum revealed a power law, which is characteristic of scale-invariance (the exponent of this power law was roughly –1, a noise-structure often observed in physical data; Handel & Chung, 1993).

Please refer to these different papers. It would be interesting to study or explain also if a power-law arises or not in your model.

Authors' response: We agree that embodiment deserves explicit discussion and have added a few lines addressing this to the Discussion. In general response to this point, we note that we have focused on *when* to act (which is hypothesized to be the responsibility of premotor networks like SMA) rather that sensory processing (which would require an upstream sensor network) or implementation (which would require a downstream motor network). However, delays involved in sensory processing and implementation are very relevant to determining when to act. We have attempted to capture the neuromechanical delay between the motor decision and the physical tap onset by including a fixed execution delay of 50 ms: once the network emits a tap action at time t, the tap is registered in the environment at time t + 50 ms.

To begin with a minimal sufficient model, we have assumed no neural noise or motor timing noise. This limits our ability to investigate the autocorrelation structure of noisy interval generation – our agents are not inherently noisy, so we instead introduce "apparent" noise in the form of noisy training and testing stimuli (i.e., our timeline t is the subjective time of the agent, so internal noise is replaced by temporally noisy stimuli). We now discuss the introduction of noise as an exciting future direction, citing Gilden.

In the neural modeling, the paper does not attempt also to model specifically the dopamine circuit of the basal ganglia with the reward functions. Some references to those bio-inspired models should be added.

We have now added references to the classic model of dopaminergic reinforcement learning (Schultz 1997) to the discussion, and suggested that the model could be extended with a mechanism for dopaminergic RL in RNNs with several cited examples..

Moreover, some dynamical models based on reservoir computing on this topic of time perception/generation have to be added like:
Laje and Buonomano, "Robust timing and motor patterns by taming chaos in recurrent neural networks". Nature Neuroscience 2013
Y. Kawai, J. Park, I. Tsuda, M. Asada Learning long-term motor timing/patterns on an orthogonal basis in random neural networks Neural Netw., 163 (2023), pp. 298-311
Yuji Kawai, Takashi Morita, Jihoon Park, Minoru Asada, Oscillations enhance time-series prediction in reservoir computing with feedback, Neurocomputing, Volume 648, 2025

Could you describe the difference of your model and these ones? Why not have taken dynamical systems model that are dynamic?

Authors' response: We have referenced these models in the Discussion, noting that it is difficult to say whether our network employed any of the techniques that have proven useful in previous literature since our training algorithm does not directly manipulate dynamic regime or modular architecture. Future explorations could try to map out possible correspondences between what happened in our network and what has proven successful in training other timing tasks.

Considering the four reward functions. They are well described.
One of them, the asymmetric model, appear to replicate well, the hebbian Spike-Timing Dependent Plasticity (STDP) learning mechanism. Could you describe the link/difference between your model and this one?

Authors' response: Interesting thought! We have added this possible connection to the Discussion, though we can only speculate as to whether STDP could be a mechanism supporting learning based on the temporal order of sound and tap events.

One strong argument against your model is the duration of the experiments which is very long >100 hours. These training times are they in line with human performance? Why choosing this time scale? Is it possible to speed up learning convergence?

It appears a strong drawback of the method on synchronization. Some analysis and models with more rapid phase synchronization require less duration. The authors should give argument on that.

Authors' response: Excellent question. Some back-of-the-envelope calculation: According to "The Origins of Dance: Characterizing the Development of Infants' Earliest Dance Behavior" (Kim et al, *Developmental Psychology* 2023), 50% of infants are moving spontaneously to music by 6 months old. According to "Everyday Music in Infancy" (Mendoza et al, *Developmental Science* 2021), a sample of infants between 6 and 12 months spent a median of about 40 minutes per day exposed to music. An infant could get 100 hours of exposure to music accompanied by movement in just the 6 months before their first birthday; even assuming that only a portion of this time is appropriately rhythmic and evokes movement, they could certainly get sufficient training by the age of 4 when synchronization-like behavior begins to emerge. We have added this as a paragraph in the Discussion.

We thank the three reviewers for their work on this manuscript. Reviewers 1 and 3 report no remaining concerns; we address Reviewer 2's three comments below.

Reviewer #1

My concerns have been sufficiently addressed.

Response: Thank you, we are glad the revisions were satisfactory, and grateful for the comments that prompted them.

Reviewer #2

Although the language has been improved, some phrases such as "qualitatively and quantitatively" or "closely resembles human data" may still sound somewhat stronger than the analyses warrant, especially where the comparison is primarily based on overlap with human variability rather than formal model comparison. I suggest slightly softening these statements, e.g., "falls within the range of human data" or "captures key qualitative and quantitative features."

Response: We agree, and have softened these claims wherever possible. We have kept "closely resembles" for the negative mean asynchrony, as the agents' asynchronies and their scaling with tapping noise and with training track the human data closely enough that a weaker phrase would, we think, understate the correspondence.

The authors now appropriately acknowledge that cyclic PCA trajectories are expected in periodic tasks. Still, the text could make it even clearer that the neural comparison is mainly descriptive and hypothesis-generating, rather than a strong mechanistic match to macaque data.

Response: We agree, and have kept the neural comparison descriptive throughout. We do not have access to the underlying macaque recordings, so the comparison can only be made visually; the resemblance is, we think, best read as a starting point for a closer quantitative comparison rather than as a mechanistic match — one we would be glad to attempt should those data become available.

Please check for a few remaining typographical or formatting issues in the revised manuscript and figures, including some figure-label readability and minor wording issues such as "would have the option of continue to tap." These do not affect the scientific conclusions but should be cleaned up before publication.

Response: Thank you for catching these. We have gone through the text and captions and corrected the remaining errors.

Reviewer #3

The authors corrected and improved the manuscript accordingly to the reviewers' questions. I have no other questions.

Response: Thank you for your time across both rounds of review.

We thank the three reviewers for their work on this manuscript. Reviewers 1 and 3 report no remaining concerns; we address Reviewer 2's three comments below.

Reviewer #1

My concerns have been sufficiently addressed.

Response: Thank you, we are glad the revisions were satisfactory, and grateful for the comments that prompted them.

Reviewer #2

Although the language has been improved, some phrases such as "qualitatively and quantitatively" or "closely resembles human data" may still sound somewhat stronger than the analyses warrant, especially where the comparison is primarily based on overlap with human variability rather than formal model comparison. I suggest slightly softening these statements, e.g., "falls within the range of human data" or "captures key qualitative and quantitative features."

Response: We agree, and have softened these claims wherever possible. We have kept "closely resembles" for the negative mean asynchrony, as the agents' asynchronies and their scaling with tapping noise and with training track the human data closely enough that a weaker phrase would, we think, understate the correspondence.

The authors now appropriately acknowledge that cyclic PCA trajectories are expected in periodic tasks. Still, the text could make it even clearer that the neural comparison is mainly descriptive and hypothesis-generating, rather than a strong mechanistic match to macaque data.

Response: We agree, and have kept the neural comparison descriptive throughout. We do not have access to the underlying macaque recordings, so the comparison can only be made visually; the resemblance is, we think, best read as a starting point for a closer quantitative comparison rather than as a mechanistic match — one we would be glad to attempt should those data become available.

Please check for a few remaining typographical or formatting issues in the revised manuscript and figures, including some figure-label readability and minor wording issues such as "would have the option of continue to tap." These do not affect the scientific conclusions but should be cleaned up before publication.

Response: Thank you for catching these. We have gone through the text and captions and corrected the remaining errors.

Reviewer #3

The authors corrected and improved the manuscript accordingly to the reviewers' questions. I have no other questions.

Response: Thank you for your time across both rounds of review.