BEATS, BOTS, AND BANANAS

BEATS, BOTS, AND BANANAS: MODELING REINFORCEMENT LEARNING OF SENSORIMOTOR SYNCHRONIZATION

By YASSAMAN OMMI,

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Science

McMaster University
MASTER OF SCIENCE (2024)
(Computational Science and Engineering)

TITLE:	Beats, Bots, and Bananas: Modeling reinforcement learning of sensorimotor synchronization
AUTHOR:	Yassaman Ommi BSc Computer Science, Amirkabir University of Technology, Tehran, Iran
SUPERVISOR:	Dr. Jonathan Cannon
NUMBER OF PAGES:	xiii, 77

Lay Abstract

Have you ever wondered how we naturally tap our foot in time with music? This thesis investigates this human ability, known as sensorimotor synchronization, using artificial intelligence. By creating artificial agents that learn to tap along with a steady beat through reinforcement learning—like a person tapping to a metronome—we aimed to understand how the brain acquires this skill.

Our experiments showed that how we define success, significantly affects how the agents learn the skill. Notably, when we rewarded both precise timing and consistent tapping, the agents' behavior closely resembled that of humans. They even exhibited a human-like pattern in error correction, making larger adjustments when tapping too late rather than too early.

This research offers new insights into how our brains process and learn rhythm and timing. It also lays the groundwork for developing AI systems capable of replicating human-like timing behaviors, with potential applications in music technology and robotics.

Abstract

This thesis investigates the computational principles underlying sensorimotor synchronization (SMS) through the novel application of deep reinforcement learning (RL). SMS, the coordination of rhythmic movement with external stimuli, is essential for human activities like music performance and social interaction, yet its neural mechanisms and learning processes are not fully understood.

We present a computational framework utilizing recurrent neural networks with Long Short-Term Memory (LSTM) units, trained via RL, to model SMS behavior. This approach allows for the exploration of how different reward structures shape the acquisition and execution of synchronization skills. Our model is evaluated on both steady-state synchronization and perturbation response tasks, paralleling human SMS studies.

Key findings reveal that agents trained with a combined reward—minimizing next-beat asynchrony and maintaining interval accuracy—exhibit human-like adaptive behaviors. Notably, these agents exhibited asymmetric error correction, making larger adjustments for late versus early taps, a phenomenon documented in human subjects. This suggests that such asymmetry may arise from the inherent reward structure of the task rather than from specific neural architectures. While our model did not consistently reproduce the negative mean asynchrony observed in human steady-state

tapping, it demonstrated anticipatory behavior in response to perturbations. This offers new insights into how the brain might learn and execute rhythmic tasks, indicating that anticipatory strategies in human synchronization could naturally arise from processing rewards and timing errors.

Our work contributes to the growing integration of machine learning techniques with cognitive neuroscience, offering new computational insights into the acquisition of timing skills. It establishes a flexible framework, which can be extended for future investigations in studying more complex rhythms, coordination between individuals, and even the neural basis of rhythm perception and production.

To my family

The steady beat

Guiding me through.

Acknowledgements

Before you delve into the chapters that follow, I wish to express my profound gratitude to the individuals whose support, guidance, and encouragement made this work achievable.

First and foremost, I wish to express my deepest appreciation to my supervisor, Dr. Jonathan Cannon. Thank you for giving me the chance and for believing in my potential when it was just a tiny glimmer. Anyone would be incredibly lucky to have a supervisor as kind and understanding as you. Your insights, patience, and unwavering support have not only shaped this thesis, but have profoundly influenced my growth as a human being.

To my parents, whose unconditional love and sacrifices have been the bedrock of my life, words cannot fully capture my indebtedness. Thank you for nurturing my dreams and instilling in me the values of perseverance and integrity. Your faith in me has been a constant source of strength, and your support has made this accomplishment possible.

To my brother, AmirAli, your friendship and encouragement have been a beacon throughout this journey. Thank you for always being there to listen, to laugh, and to remind me of the joys beyond academia. I wish you endless climbs and new heights, both on the rocks and in life. May you always find the summit you seek, just as you

have helped me reach mine.

I am also indebted to my friends, whose companionship has turned challenges into adventures. Being away from home has not been easy, but your presence has made all the difference. To Matin and Fateme for being my family away from home, and for your unwavering support. To Amir, the rhythm beneath.

To my cat, Tilé, whose luxurious lifestyle and occasional keyboard strolls reminded me daily of cats' superiority in life. Thank you for gracing me during late-night writing sessions and for showing me that a good nap can solve most problems.

And a special thanks to every writer who has ever put pen to paper (or fingers to keyboard), crafting worlds, and spinning tales that transported me far from my desk, offering refuge to a solitary soul. Without them, I might have finished my thesis in half the time.

This thesis stands as a testament to the collective support of all these individuals. Any merits it may have are a reflection of their contributions; its shortcomings are mine alone.

Table of Contents

List of Figures

Chapter 1

Introduction

The remarkable human ability to synchronize bodily movements to an external auditory rhythm, known as sensorimotor synchronization, is a fundamental skill that underlies diverse behaviours from basic bodily movements to complex activities such as speaking and musical performance. Closely intertwined with this is beat perception - the perception and cognitive tracking of the regular pulse underlying periodic rhythmic patterns. These capabilities represent complex integrations of auditory, motor, and cognitive processes that are shaped by multiple factors over the course of human development.

Numerous empirical investigations have shed light on how sensorimotor synchronization and beat perception are influenced by musical training experience [38]. Individuals with extensive formal musical training tend to exhibit significantly enhanced precision and temporal consistency when tapping along with rhythmic auditory stimuli compared to non-musician controls. Musically trained participants also demonstrate improved performance on beat perception tasks that require detecting or discriminating temporal deviations from isochronous sequences. These findings suggest

that long-term engagement with music and its inherent rhythmic structure can fine-tune the neural mechanisms underlying synchronization and beat processing abilities.

Moreover, the cultural context in which one is raised also plays a pivotal role. Different musical systems around the world employ diverse rhythmic structures, meters, and temporal primitives. From a young age, infants are enculturated to the rhythmic patterns and beat biases prominent in their musical environment. Cross-cultural research has revealed that adult listeners' metrical interpretations and beat tracking tendencies are shaped by their familiarity with the rhythmic properties of their culture's music. Sensitivity to canonical rhythmic structures gets reinforced through repeated exposure during development. For instance, individuals raised in Western musical traditions may struggle with complex rhythms common in Balkan or African music, while those from these cultures show enhanced abilities to synchronize with such rhythms [8].

Intriguingly, the development of sensorimotor synchronization and beat perception in humans follows a protracted developmental trajectory. Infants younger than around 2 years old exhibit great difficulty synchronizing body movements to auditory rhythms, failing to consistently match movements to beat locations [56]. This capacity for rhythmic entrainment gradually emerges over late childhood and continues improving into adolescence, suggesting an extended developmental process. In stark contrast, most non-human animal species appear to lack the spontaneous ability for sensorimotor synchronization, although some primates like monkeys have been trained using reinforcement learning techniques to tap in synchrony with an isochronous metronome beat. Moreover, cockatoos and parrots have demonstrated the capacity to spontaneously synchronize with musical beats [33].

These observations indicate that while sensorimotor synchronization and beat perception likely have innate biological foundations, their full-fledged expression critically depends on learning processes that occur over an extended period through interaction with rhythmic stimuli present in the individual's environment and culture. The neural substrates underlying sensorimotor synchronization and beat perception involve a distributed network of brain regions. Neuroimaging studies have consistently implicated the basal ganglia, cerebellum, premotor cortex, supplementary motor area, and auditory cortex in these processes [39]. Of particular interest is the role of the basal ganglia in beat perception and the cerebellum in precise timing. Furthermore, dopaminergic activity in the basal ganglia has been associated with both reward processing and temporal prediction [17], providing a potential neural link between the rewarding nature of rhythmic synchronization and the underlying timing mechanisms.

Crucially, the act of moving in synchrony with rhythmic auditory stimuli appears to be intrinsically rewarding, engaging neural reward circuitry in the brain. Studies have found that the subjective experience of "groove" when synchronizing body movements to music involves the intersection of reward processing and the motor processes underlying beat perception, driven by rhythmic expectation mechanisms [25]. This suggests that sensorimotor synchronization abilities could be acquired incrementally through a reinforcement learning process, where successful synchronization is rewarding and reinforces the appropriate temporal predictions and motor patterns. The reinforcement learning paradigm therefore provides a compelling framework for conceptualizing how these rhythmic skills develop based on the consequences of synchronization attempts and the rewarding experiences that arise from accurate synchronization. With each iteration of rewarding synchronization experiences, the

neural mechanisms for temporal prediction, error correction, and motor entrainment could become gradually refined and entrenched through this reinforcement process. The intrinsically rewarding nature of groove may represent a pivotal driver of the protracted learning trajectory observed in humans as they capitalize on these rewarding synchronization experiences over development.

Reinforcement learning models are neurally plausible accounts of how learning can occur in the brain. Numerous neuroscientific studies have identified neural correlates of reward prediction errors - the core signals driving learning in reinforcement models. Dopaminergic neurons in the basal ganglia have been found to encode just such reward prediction error signals [45, 17, 3]. This neural evidence lends plausibility to the idea that the brain's sensorimotor timing capacities could be tuned over development using reinforcement learning mechanisms.

Importantly, reinforcement learning has already proven successful in training non-human primates like monkeys to synchronize movements to an isochronous metronome beat, providing a compelling existence proof for this learning approach in sensorimotor timing acquisition [5, 50]. In [5], they explored how monkeys can be trained to synchronize tapping movements to rhythmic visual or auditory cues. By recording neural activity in the medial premotor cortex (MPC), the researchers discovered neurons in this region create circular timing patterns that reset with each tap, functioning like an internal clock for beat synchronization. The speed and rhythm of the tapping behaviour was reflected in the size and timing of these neural patterns, even across different sensory modalities. This work provides insight into the brain mechanisms for processing and synchronizing movements to rhythmic beats, which we aims to capture computationally with the reinforcement learning model proposed in the following.

To create an empirical model capturing these learning mechanisms for sensorimotor synchronization, we implement a deep reinforcement learning agent using recurrent neural networks (RNNs), particularly those employing long short-term memory (LSTM) architectures. RNNs represent neurally-plausible computational models well-suited for this purpose. They have proven remarkably effective at learning and encoding time-dependent patterns, rhythmic grammars, and sequence predictions in a manner analogous to how the brain may engage in temporal information processing for time keeping, beat induction, and rhythmic prediction. Moreover, the rhythmic synchronization task we use to train our model agent is directly inspired by experimental paradigms from the neuroscience literature investigating the mechanisms of temporal processing and timing behaviour. We are simulating a synchronization task, where the agent is being rewarded for synchronized actions, initiated in anticipation of the metronome cues. By implementing goal-directed reinforcement learning in an LSTM agent operating within such a rhythmic synchronization task environment, we can create an empirical model that captures the incremental acquisition of sensorimotor synchronization that occurs through repeated interaction with rhythmic stimuli and the consequences of synchronization accuracy.

To further explore the potential of reinforcement learning in modeling sensorimotor synchronization, we will train our RL agent using different reward policies. Specifically, we will investigate the effects of calculating rewards based on two distinct approaches: (1) comparing the agent's taps to the next beat versus the nearest beat, and (2) rewarding for asynchrony only versus rewarding for both asynchrony and interval errors. By implementing these varied reward structures, we aim to understand how different feedback mechanisms influence the learning and performance

of sensorimotor synchronization tasks. This approach allows us to model and compare different hypothetical learning scenarios that may occur during human development or in experimental settings.

After training, we will test our agent on perturbation sensorimotor synchronization (SMS) tasks. These tasks, which involve introducing unexpected changes in the timing of auditory stimuli, have been crucial in understanding error correction mechanisms in human sensorimotor synchronization [37]. By examining the agent's performance in these perturbation tasks, we can assess how well our model captures the flexibility and error correction capabilities observed in human sensorimotor synchronization. This novel approach of combining varied reward policies with perturbation testing will provide insights into the learning mechanisms underlying sensorimotor synchronization and may offer new perspectives on how humans develop and refine these critical timing skills.

In summary, this thesis aims to develop a novel computational model of sensorimotor synchronization using deep reinforcement learning with LSTM networks. By training this model on a tapping task and comparing its performance to existing behavioral data, including responses to perturbations, we hope to gain new insights into the learning mechanisms underlying this fundamental human ability. This approach bridges neuroscience, psychology, and machine learning, offering a unique perspective on how the brain learns to move in time with music.

The subsequent chapters will be structured as follows:

- Chapter 2 - Literature Review: provides a review on the foundational concepts and existing data in sensorimotor synchronization, reinforcement learning, timing models of brain, and recurrent neural networks.

- Chapter 3 - Methodology: details the computational models used and experimental setup.
- Chapter 4 - Results and Discussion: presents findings from the simulations and comparative analysis for existing empirical data.
- Chapter 5 - Conclusion and Future Work: summarizes the research findings, discusses their implications, and outlines potential areas for further research.

Chapter 2

Literature Review

2.1 Learning to Synchronize

Synchronizing movement to external rhythmic patterns is one of the most remarkably complex capabilities of the human mind and body. This fundamental skill, known as sensorimotor synchronization (SMS), enables activities ranging from musical performance and dance to the rhythmic patterning of speech and coordinated physical actions. At its core lies beat perception – the cognitive ability to extract and internally model the underlying periodic pulse from a rhythmic sensory stream. Beat perception allows us to leverage the temporal regularities within complex auditory signals like music and speech, enabling us to prospectively guide synchronized motor outputs. These inextricably intertwined abilities emerge from the convergence of multiple intricate processes spanning perception, action, reward, cognition, and temporal prediction. Their development follows an extended trajectory, gradually unfolding through childhood and adolescence. While they likely have innate neurobiological foundations, their full-fledged behavioural expression critically depends on learning

and enculturation processes that occur through immersion in the structured rhythmic patterns present within one's cultural and musical environments. Extensive musical training further refines and optimizes these capabilities by progressively sculpting and enhancing their neural underpinnings.

Sensorimotor synchronization and beat perception are not innate; they develop over time, shaped by cultural exposure and practice. Children begin exhibiting basic rhythmic abilities early on, but mastering these skills is a protracted process that continues into adolescence. Cultural context plays a crucial role in shaping how individuals perceive and synchronize with musical rhythms. Different musical traditions emphasize unique rhythmic patterns and complexities, which influence synchronization abilities. For example, the structured meters of Western classical music might foster different synchronization skills compared to the intricate rhythms found in African drumming traditions. Cross-cultural studies have shown that while most adults can synchronize movements with simple, regular patterns, the ability to accurately follow complex rhythms is significantly affected by one's musical enculturation. The precision of this synchronization is linked to a listener's musical experience; long-term exposure to specific musical cultures shapes rhythm perception and creates expectations that make familiar rhythms easier to predict and synchronize with. This phenomenon is observed in finger-tapping tasks, where participants from India [53], Turkey [15], Mali [34], and other non-Western musical cultures exhibit remarkable consistency in synchronizing with the complex rhythms common in their traditions. This contrasts sharply with participants from Western cultures, where such rhythmic patterns are relatively rare. Listeners from non-Western musical backgrounds can accurately tap along to the intricate rhythmic nuances of their native music, a

task that may be challenging for those from Western musical backgrounds. These findings underscore how extensive exposure to the structured rhythmic complexities embedded within a culture's musical environment critically refines and influences sensorimotor synchronization skills specific to those rhythms. Therefore, while basic synchronization to simple beat patterns emerges through typical development, achieving precision with highly complex rhythms appears to depend crucially on sustained learning and immersion in the rhythmic nuances particular to one's cultural and musical context. Extensive musical training also acts as a catalyst, enhancing sensorimotor synchronization and beat perception abilities. Musicians often demonstrate superior timing accuracy and an exceptional capacity to synchronize with complex rhythmic structures. Training progressively refines the neural mechanisms that underpin these abilities, making motor responses more precise and auditory processing more acutely sensitive to rhythmic nuances.

Recent research by Betancourt et al. (2023) [5] provides compelling evidence for the learned nature of rhythmic skills and sensorimotor synchronization. In their study, they trained macaque monkeys to perform a synchronization tapping task (ST) with both auditory and visual metronomes. The monkeys were able to learn to tap in synchrony with isochronous stimuli at different tempos (450 ms and 850 ms intervals), demonstrating that even non-human primates can acquire these abilities through training and reinforcement learning. The study found that the monkeys could accurately produce intervals with errors close to zero, showing slight underestimation in the auditory condition. Importantly, the monkeys exhibited an error correction mechanism, particularly strong for visual metronomes. This suggests that

the monkeys learned not only to synchronize but also to continually adjust their timing based on feedback, a hallmark of a sophisticated learned skill. Furthermore, the researchers observed that the monkeys' performance showed characteristics similar to human sensorimotor synchronization, such as the scalar property of timing (increased variability for longer intervals) and differences in performance between auditory and visual modalities. These findings support the idea that rhythmic skills, while potentially built on innate neural substrates, are significantly shaped by learning and experience.

2.1.1 Neural Basis of Beat Perception

The neural basis of beat perception involves a complex interplay between auditory and motor regions, particularly the motor cortico-basal-ganglia-thalamo-cortical (mCBGT) circuit [26]. Key structures in this circuit, such as the supplementary motor area (SMA) and putamen, are consistently active during beat perception tasks [10]. Recent research emphasizes the critical role of motor systems in beat perception, even without overt movement. Cannon and Patel (2021) [9] proposed that beat anticipation relies on "proto-actions" in the SMA, orchestrated by the dorsal striatum. These neural processes provide temporal structure without specifying particular movements.

The tight coupling between auditory and motor systems is further evidenced by increased functional connectivity between these areas during beat perception, particularly for musicians [10]. This connectivity is facilitated by oscillatory activity, especially in the delta (1–3 Hz) and beta (15–30 Hz) frequency bands, which aligns motor predictions with auditory inputs, enabling the synchronization of movements

with perceived beats [18], and underlies the ability to imagine and process hierarchical timing [11]. Thus, motor regions are not merely involved in movement execution but also play a fundamental role in the perception and cognitive processing of rhythmic patterns [26].

The sensation of groove – the pleasurable urge to move with music – also engages motor regions in tandem with reward systems. Rhythms of medium complexity, which maximize groove, activate motor areas like the SMA and premotor cortex, as well as reward centers like the nucleus accumbens. This interplay suggests that the rewarding aspects of rhythm enhance synchronization by motivating engagement with the beat, reinforcing the deep connection between music, movement, and pleasure [25].

2.1.2 Laboratory Studies of SMS

Laboratory studies of sensorimotor synchronization (SMS) frequently employ simple yet insightful paradigms, such as finger tapping tasks with auditory sequences of tones or clicks. However, the experimental landscape is rich and diverse, with numerous variants arising from different forms of movement (e.g., tapping on a surface versus finger flexion or limb movement without contact), different modalities of sensory stimulation (e.g., auditory or visual rhythms), and different coordination patterns (e.g., in-phase or antiphase synchronization) [37, 39]. Among these paradigms, perturbation studies, introduced by Michon (1967) [27], have played a pivotal role in elucidating the underlying mechanisms of SMS. In these experiments, researchers intentionally introduce unexpected timing deviations or disruptions into the rhythmic sequences to probe how the sensorimotor system responds and maintains synchronization. Three main types of perturbations are commonly used: phase shifts, event onset

shifts, and tempo changes. These studies have revealed two primary error correction mechanisms: phase correction and period correction. Phase correction is a rapid, automatic process that adjusts the timing of individual movements, while period correction involves a slower, more cognitive adjustment of the internal timekeeper.

A particularly illuminating paradigm involves event onset shifts, wherein the timing of a single tone within the sequence is abruptly advanced or delayed. Despite explicit instructions to ignore such timing shifts, participants exhibit an involuntary phase correction response, automatically adjusting their tapping to realign with the new rhythm. This robust phenomenon not only reveals the exquisite sensitivity of the SMS system to timing deviations but also provides a window into the implicit temporal processing mechanisms involved in beat tracking and synchronization. Furthermore, research has uncovered an intriguing asymmetry in the error correction process during SMS tasks. Tomyta et al. (2023) [51] demonstrated that the error correction rate is larger when the asynchrony is positive (taps following the metronome) compared to when it is negative (taps preceding the metronome). This asymmetric error correction may contribute to the well-known negative mean asynchrony (NMA) phenomenon, where tapping onset tends to precede metronome onset by a few tens of milliseconds.

It is worth noting that while humans excel at SMS, this ability is relatively rare in the animal kingdom [7]. Some bird species, particularly those capable of vocal learning, have demonstrated SMS abilities, but evidence in non-human primates is limited. This species difference has led to intriguing hypotheses about the evolutionary origins of rhythm perception and its potential links to vocal learning and language

[32]. In conclusion, the human capacity for sensorimotor synchronization with auditory rhythms emerges from an intricate tapestry of perception, action, reward, and prediction systems. Cultural exposure, musical training, and continuous feedback and error correction are crucial for developing and refining these abilities. Perturbation studies reveal the system's remarkable flexibility and precision in maintaining synchronization, highlighting the importance of implicit and explicit error correction processes. The rewarding sensation of groove further motivates and reinforces rhythmic engagement, underscoring the profound interconnectedness between music, movement, and the human experience. Future research in this field may benefit from integrating insights from neuroscience, psychology, and computational modeling to further unravel the complexities of this fundamental human ability.

2.2 Inner Timekeepers

Understanding how the brain perceives, maintains, and utilizes time intervals is crucial for a wide range of cognitive functions, from motor control to decision-making. To take full advantage of the rewards in our environment, we rely on relationships between causes and effects that exhibit precise temporal dependencies. For instance, to evade an incoming threat, one must gauge its velocity by tracking its movement over a fixed time period, project its future location during another time interval, and time an escape maneuver based on the estimated moment of potential impact. This capability to measure time and use it to guide behaviour is essential and ubiquitous in both biological organisms and artificial agents. However, due to fundamental differences in their underlying implementations, artificial intelligence (AI) and biological

systems have distinct relationships with the concept of time. Nonetheless, examining the temporal measurement challenge across these domains may yield valuable cross-disciplinary insights.

2.2.1 Recurrent Neural Networks

Neural Networks are computational models inspired by the structure and function of biological neural networks in the human brain. They consist of interconnected nodes that transmit signals between each other, mimicking the behavior of neurons. The connections between these artificial neurons are modulated by numeric weights, analogous to the strengths of synaptic connections in the brain. Neural networks excel at finding intricate patterns in data, making them powerful tools for various tasks. Importantly, their architecture and mechanisms have enabled researchers to draw insights from neuroscience literature and develop models that simulate brain functions, bridging the gap between artificial and biological neural networks.

Feedforward neural networks are the most basic type, where the connections between nodes do not form any cycles or loops, and information flows in only one direction from input to output. These networks process each input independently, lacking the capability to maintain and utilize information from previous inputs. However, human cognition is marked by the persistence of thoughts and integration of information across time, where our understanding of the present is shaped by our prior experience and context. Recurrent Neural Networks (RNNs) address this shortcoming by introducing recurrent connections that allow information to flow through the network over time. Through these looped connections, RNNs can maintain an internal state that captures and integrates sequential information, imitating the persistent

nature of human thinking and reasoning (Figure 2.1).

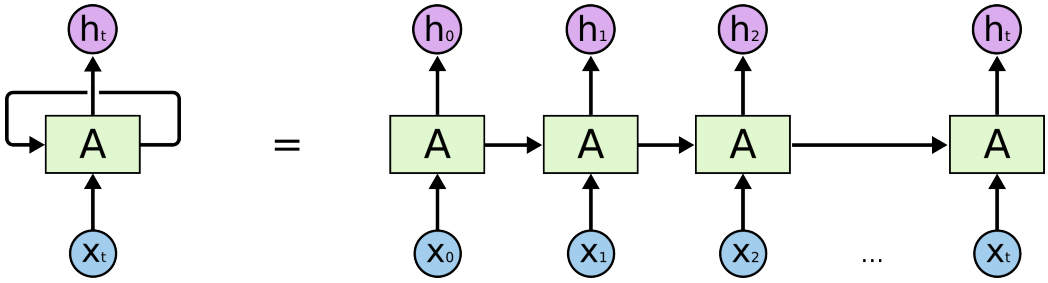


Figure 2.1: Recurrent Neural Networks have loops. This sequential structure shows that recurrent neural networks are inherently connected to sequences. (Figure from [31])

Due to these recurrent connections, RNNs are well-suited for processing sequential data, such as time series or natural language, and have been extensively used to model temporal processing in the brain. The basic architecture of an RNN can be described by the following equations:

$$h_t = f(W_{hx}.x_t + W_{hh}.h_{t-1} + b_h), \tag{2.2.1}$$

$$y_t = g(W_{yh}.h_t + b_y). \tag{2.2.2}$$

Where:

- x_t is the input at time step t
- h_t is the hidden state at time step t
- y_t is the output at time step t
- W_hx, W_hh, W_yh are weight matrices
- b_h and b_y are bias vectors

- f and g are activation functions

At each time step t, the network computes a new hidden state h_t based on the current input x_t and the previous hidden state h_{t-1}. This recurrent computation, represented by the term $W_{hh}.h_{t-1}$, allows the network to maintain and update information over time. The function f is typically a nonlinear activation function such as $tanh$ or $ReLU$, which introduces nonlinearity into the network's computations. The output y_t is then computed from the current hidden state h_t. RNNs are particularly well-suited for modeling timing processes in the brain for several reasons:

- **Temporal Integration:** RNNs can integrate information over time, mimicking the brain's ability to accumulate and process temporal information.
- **State-Dependent Processing:** The current output of an RNN depends on both its current input and its internal state, reflecting how neural responses in the brain depend on both current stimuli and recent history.
- **Flexible Timescales:** RNNs can learn to represent and process information across various timescales, from milliseconds to seconds or longer, matching the brain's ability to handle diverse temporal tasks.
- **Emergent Timing:** Instead of relying on an explicit clock mechanism, timing in RNNs emerges from the network dynamics, aligning with theories that suggest timing in the brain arises from the intrinsic properties of neural circuits.

Learning to Forget: Long Short-Term Memory

However, traditional RNNs suffer from the vanishing and exploding gradient problems, which can make it difficult to learn long-term dependencies. The vanishing gradient problem occurs when gradients become extremely small as they are propagated back through time during training, effectively preventing the network from learning connections between temporally distant events. This limitation is particularly problematic for tasks requiring long-term memory, such as understanding context in language or recognizing patterns over extended time periods in timing tasks.

To address this, LSTM units were introduced by Hochreiter and Schmidhuber in 1997 [16]. LSTMs maintain a cell state that is modulated by gates (input, output, and forget gates), enabling them to selectively remember or forget information over long periods. Gates function as selective information filters, consisting of a sigmoid layer (σ) and pointwise multiplication. The sigmoid layer outputs values between 0 and 1, determining the degree of information transmission for each component. This mechanism enables precise control over information flow within the network. The LSTM architecture and the interaction between the its layers are shown in Figure 2.2.

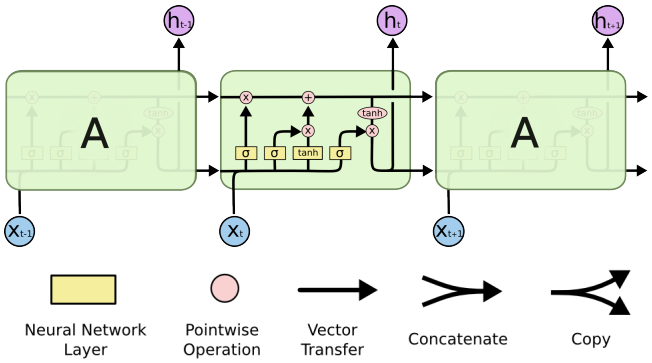
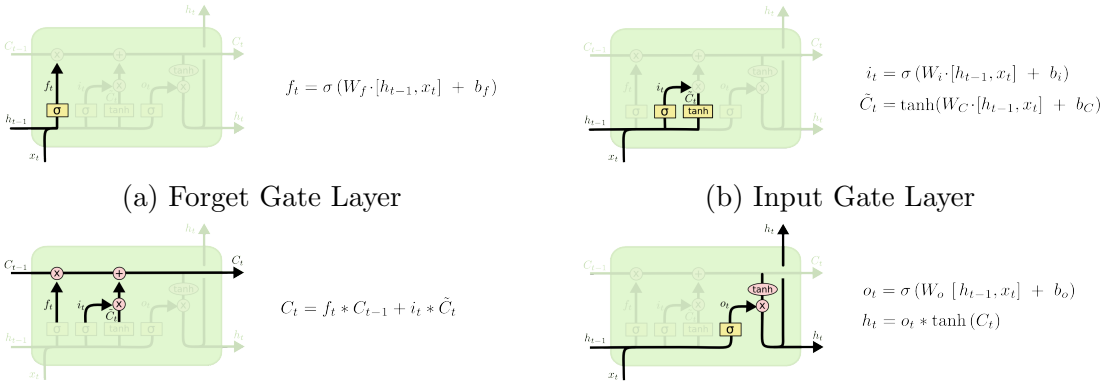


Figure 2.2: The repeating module in an LSTM contains four interacting layers.

A step-by-step overview of how an LSTM works and updates its cell state, C, is shown in 2.3. For a more detailed review on LSTM and its equations, see [16, 47] .



(a) Forget Gate Layer

(b) Input Gate Layer

(c) Cell State Update

(d) Output Gate Layer

Figure 2.3: The equation governing each layer of an LSTM unit. Notations are the same as mentioned in section 2.2.1. 2.3a The forget gate decides what information to keep and what to forget. 2.3b The input gate decides what new information is going to be stores in the cell state. 2.3c Cell state is updated by forgetting some $(f_t * C_{t-1})$ and learning some $(i_t * \tilde{C}_t)$. 2.3d The output gate filters the cell state and generates the output.

RNNs for Neural Dynamics Modeling

The use of RNNs in modeling neural processes, particularly those involved in timing and temporal processing, has gained significant traction in recent years. This approach is rooted in the idea that the brain's ability to process temporal information emerges from the dynamics of recurrently connected neural networks, rather than relying on a centralized clock mechanism. The biological plausibility of RNNs stems from the recurrent connectivity observed in cortical circuits. In the brain, neurons are extensively interconnected, with feedforward and feedback connections forming complex networks. This architecture allows for the persistence and integration of information over time, a key feature that RNNs aim to emulate. Mante et al.

(2013) [23] demonstrated that the prefrontal cortex implements context-dependent processing through the dynamics of recurrent networks. Their findings showed striking similarities between the neural trajectories observed in monkey prefrontal cortex and those produced by trained RNNs, providing strong support for the use of these models in understanding cortical computation. Another crucial aspect of neural processing captured by RNNs is state-dependent computation. In the brain, the response of neural circuits to incoming stimuli depends not only on the current input but also on the network's internal state, which is shaped by recent history. This property is essential for tasks requiring temporal integration or working memory.

In the domain of timing and temporal processing, RNNs have proven particularly valuable. The ability of these networks to maintain and manipulate temporal information aligns well with theories of how the brain represents time. Laje and Buonomano (2013) [21] proposed that the brain tells time through the evolution of neural trajectories in recurrent networks. They demonstrated that RNNs can be trained to generate complex temporal patterns, mimicking the ability of cortical circuits to encode and produce precisely timed sequences.

Moreover, RNN models are able to capture the heterogeneity and mixed selectivity observed in real neural populations. In the brain, individual neurons often respond to multiple task variables and exhibit complex, time-varying patterns of activity. Rigotti et al. (2013) [41] showed that this mixed selectivity is a crucial feature of prefrontal cortex, enabling flexible cognitive behavior. RNNs naturally develop similar properties when trained on complex tasks, providing a computational framework for understanding how mixed representations support cognitive flexibility. Recent work has focused on establishing more direct links between the computations performed

by RNNs and potential neural mechanisms. For instance, Masse et al. (2019) [24] developed a method for mapping RNN units onto biologically plausible neural circuits, providing a way to generate testable predictions about the implementation of complex computations in the brain.

RNN Models of Timing in the Brain

Recent theories suggest that timing may emerge from the intrinsic dynamics of neural networks rather than from a dedicated clock mechanism. This aligns well with the aforementioned properties of RNNs, which can maintain temporal information across variable lengths of time. Jazayeri and Shadlen (2015) [19] conducted a pivotal study recording neural activity in the lateral intraparietal cortex (LIP) of monkeys performing a time reproduction task. They found that neural activity patterns during the measurement phase predicted the timing of subsequent actions. This study provided compelling evidence that the brain encodes time prospectively, integrating sensory inputs with motor plans to achieve precise timing. Building on this work, Jazayeri and colleagues have developed several RNN-based models to explain various aspects of timing in the brain. In a 2018 study, Wang et al. [54] trained RNNs to perform interval timing tasks and found that the networks developed representations similar to those observed in neural recordings. The RNNs exhibited ramping activity and neural trajectories that closely matched empirical data from the prefrontal and parietal cortices. Remington et al. (2018) [36] further extended this approach by using RNNs to model flexible timing behavior. They trained networks to produce different time intervals based on contextual cues and found that the models captured key features of neural dynamics observed in primate supplementary motor area and prefrontal cortex

during similar tasks. More recently, Sohn et al. (2019) [46] used RNNs to investigate how the brain might represent time across multiple scales. Their model, trained on tasks requiring both precise timing and temporal abstraction, developed hierarchical representations that mirror those found in the cortex and striatum, validating the use of RNNs in modeling complex timing behaviors. The study demonstrated how prior beliefs about time intervals can be incorporated into neural representations, allowing for Bayesian integration of sensory inputs and prior knowledge. This work revealed that recurrent interactions among neurons can mediate Bayesian inference, with the network's activity manifesting as a low-dimensional curved manifold that warps neural representations to reflect prior statistics.

2.3 Learning through Reward

The ability to make decisions and learn from their outcomes is a fundamental aspect of intelligent behavior, crucial for survival and success in complex, dynamic environments. This process, known as reinforcement learning (RL), allows organisms to adapt their behavior based on the consequences of their actions, maximizing rewards and minimizing punishments over time. Decision-making, as an integral part of RL, is essential for navigating the complexities of life. Effective decision-making enables organisms to choose actions that lead to favorable outcomes, thereby increasing their chances of survival and reproduction.

Reinforcement learning is ubiquitous in nature. From single-celled organisms navigating chemical gradients to humans making complex decisions, the principles of trial-and-error learning and reward maximization are at play. Evolution has shaped our brains to be highly adept at this form of learning, as it confers significant adaptive

advantages. In the natural world, animals face a constant stream of decisions: where to forage, how to avoid predators, whom to mate with, and how to allocate limited resources. Those individuals better at learning which actions lead to positive outcomes and which to negative ones have a clear evolutionary advantage. This selective pressure has resulted in sophisticated neural mechanisms for reinforcement learning.

For humans, reinforcement learning extends far beyond basic survival. It underpins our ability to acquire new skills, form habits, navigate social interactions, and make countless daily decisions in personal and professional contexts. From a child learning to ride a bicycle to an adult mastering a new language or optimizing their investment strategy, reinforcement learning processes are at work. Decision-making is not only vital for immediate survival but also for long-term success and fulfillment. It influences every aspect of our lives, from daily routines to major life choices, shaping our experiences and determining our paths.

The study of reinforcement learning in the brain lies at the intersection of neuroscience, psychology, and computational modeling. Computational models of reinforcement learning have proven invaluable in formalizing hypotheses about neural processes and generating testable predictions. These models provide a normative framework for understanding behavior—that is, they describe how an agent should behave to maximize rewards, given certain assumptions and constraints. In recent years, the field of deep reinforcement learning has made significant strides, with algorithms like Deep Q-Networks (DQN) and Proximal Policy Optimization (PPO) achieving superhuman performance in complex tasks. However, bridging the gap between these artificial systems and biological learning remains a significant challenge.

In the following sections, we will explore the theoretical background of reinforcement learning models and their neural correlates, focusing on how these computational ideas have illuminated our understanding of brain function, adaptive behavior, and the learning of rhythmic and synchronized motor skills.

2.3.1 Adaptive Minds: Theoretical Background

Reinforcement learning (RL) emerged as a powerful paradigm in machine learning through the convergence of two distinct research streams. The first, pioneered by Richard Sutton and Andrew Barto, drew inspiration from psychology and artificial intelligence. Sutton, with his background in psychology, and Barto, a computer scientist, developed core RL algorithms and concepts based on insights from Pavlovian and instrumental conditioning [1, 49]. Their work laid the groundwork for fundamental RL principles and methods, and is what considered today as the core concepts of RL. Concurrently, a second line of research evolved from the fields of operations research and optimal control. Engineers like Dimitri Bertsekas and John Tsitsiklis approached the problem from a mathematical perspective, developing stochastic approximations to dynamic programming methods. They termed this approach "neurodynamic programming," which led to reinforcement learning rules that closely paralleled those developed by Sutton and Barto [4]. The fusion of these two research streams proved transformative. It couched the behaviorally-inspired heuristic reinforcement learning algorithms in more formal terms of optimality, providing a rigorous mathematical foundation for what were initially intuitive approaches. This synthesis not only validated the practical effectiveness of RL algorithms but also provided powerful tools for analyzing their convergence properties in various situations. This interdisciplinary

origin of RL contributes to its strength and versatility, combining insights from psychology, computer science, and control theory. As a result, modern RL offers a robust framework for developing adaptive learning systems capable of making sequential decisions in complex, uncertain environments.

Rescorla-Wagner Model

The Rescorla-Wagner model [40], proposed in 1972, was a groundbreaking contribution to our understanding of associative learning. It formalized the intuition that learning occurs when events violate expectations, providing a simple yet powerful mathematical framework for predicting the strength of associations between stimuli. The model rests on two fundamental assumptions: (1) learning occurs exclusively when events are unpredicted, and (2) predictions arising from various stimuli are combined to generate the overall prediction in a trial. These assumptions enabled the model to account for several previously puzzling phenomena in classical conditioning, including blocking, overshadowing, and conditioned inhibition.

Consider a conditioning trial with some conditional stimuli CS (like a tone, or light), and an affective unconditional stimulus US (like food). The Rescorla-Wagner model mathematically expresses:

$$V_{new}(CS_i) = V_{old}(CS_i) + \eta\big[\lambda_{US} - \sum_i V_{old}(CS_i)\big]. \tag{2.3.1}$$

In this error-correcting learning rule, $V(CS_i)$ denotes the associative strength of a conditional stimulus CS_i. The alteration in associative strength is prompted by the difference between the predicted outcome $\Sigma_i V_{old}(CS_i)$, where i represents all CSs present in the trial, and the actual outcome λ_{US}, whose magnitude corresponds to

the significance of the unconditional stimulus. η is a learning rate that can vary depending on the salience of both the unconditional and conditional stimuli being associated.

These rules and assumptions allowed the model to parsimoniously explain several anomalous features of animal learning. It clarified why a previously predicted unconditioned stimulus (US) does not support conditioning of a new conditional stimulus (CS), a phenomenon known as blocking. It also elucidated how conditional stimuli of varying salience, when presented together, might form different associations with an unconditioned stimulus, explaining overshadowing. Additionally, it accounted for inhibitory conditioning, where a stimulus that predicts the absence of an expected unconditioned stimulus gains a negative associative strength.

However, despite its successes, the Rescorla-Wagner model had several limitations. It was temporally insensitive, treating conditioning trials as discrete events and failing to account for the effects of different temporal relationships between CS and US within a trial. It also lacked an explanation for second-order conditioning, which is when stimulus B predicts an affective outcome, and stimulus A predicts stimulus B, then stimulus A also gains reward predictive value. Moreover, the model assumed a constant learning rate, whereas evidence suggests that learning rates can change over the course of training. Lastly, it did not address how animals learn about the timing of events, which is crucial in many forms of conditioning. These limitations motivated the development of more sophisticated models, including the temporal difference learning model, which we'll explore next. The temporal difference model addresses these limitations by taking into account the timing of different events, allowing it to account for higher-order conditioning and making it sensitive to temporal

relationships within learning trials.

Temporal Difference Learning

Temporal Difference (TD) learning, introduced by Sutton and Barto (1990) [48], addressed many of the limitations of the Rescorla-Wagner model by incorporating the concept of estimating future rewards. The key innovation was to frame the learning problem as one of predicting the total expected future reward from any given state, rather than just the immediate outcome. In TD learning, the goal is to estimate the value $V(S_t)$ of a state S_t, defined as the expected sum of all future rewards when starting from that state:

$$\begin{aligned} V(S_t) &= E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ...|S_t] \\ &= E\left[\sum_{i=t}^{\infty} \gamma^{i-t} r_i \middle| S_t\right]. \end{aligned} \tag{2.3.2}$$

Where r_t is the reward at time t and $\gamma \leq 1$ is a discount factor that reduces the weight of future rewards. The discount rate γ was initially introduced to ensure that the total of future rewards remains finite. Additionally, it aligns with the observation that humans and animals favor immediate rewards over delayed ones. This exponential discounting corresponds to the assumption of a constant 'interest rate' per unit time on obtained rewards or a consistent probability of leaving the task per unit time. This formulation allows TD learning to consider long-term consequences of actions and states.

The core of TD learning is the temporal difference error, which measures the difference between the predicted value of the current state and the sum of the immediate

reward plus the discounted value of the next state:

$$\delta_t = r_t + \gamma V(S_{t+1}) - V(S_t). \tag{2.3.3}$$

This error signal is crucial for learning and represents a form of prediction error. It can be interpreted as the difference between the "actual" return $(r_t + \gamma V(S_{t+1}))$ and the predicted return $V(S_t)$. The "actual" return includes the immediate reward r_t and the discounted estimate of future rewards $\gamma V(S_{t+1})$, while $V(S_t)$ represents the current estimate of the total expected future reward from the current state. If the prediction is accurate, the error will be zero. A positive error indicates that the outcome was better than expected, suggesting that the value of the current state should be increased. Conversely, a negative error means the outcome was worse than expected, and the value of the current state should be decreased. This error signal drives the learning process by constantly refining the value estimates based on experienced outcomes. We can then use the temporal difference error as a measure of 'surprise' in the Rescorla-Wagner learning rule:

$$V(S_t)_{new} = V(S_t)_{old} + \eta.\delta_t. \tag{2.3.4}$$

Where η is a learning rate that determines how quickly new information overrides old information. This update rule allows the agent to incrementally improve its value estimates based on experienced transitions and rewards, gradually converging on accurate predictions of future rewards. While TD learning shares some similarities with the Rescorla-Wagner model, such as using a learning rate and being driven by discrepancies between expected and actual outcomes, it introduces several key

innovations. TD learning explicitly represents time within a trial, allowing learning to occur at every time point, rather than treating each trial as a discrete unit. It also extends predictions to include not only immediate rewards but also future predictions from stimuli that will still be present in subsequent time steps. This allows TD learning to account for delayed rewards and long-term consequences. Furthermore, TD learning updates its predictions continuously, rather than only at the end of a trial, capturing the temporal dynamics of conditioning more accurately. These advancements allow TD learning to address major shortcomings of the Rescorla-Wagner model, such as its inability to explain second-order conditioning and its lack of sensitivity to temporal relationships between stimuli within a trial. TD learning is temporally sensitive, operating in continuous time (or fine-grained discrete time steps), allowing it to account for the effects of different $CS-US$ intervals. It naturally explains second-order conditioning and provides a mechanism for learning about the timing of events, as the value function can represent expectations of reward at different future time points.

Action Selection and Actor/Critic Models

The above applies when the probabilities of transitioning between different states of the environment are constant, such as in Pavlovian conditioning (where the animal cannot influence events through its actions). However, how do we improve action selection to obtain more rewards, as in instrumental conditioning? Since the environment rewards actions rather than predictions (even accurate ones), it can be argued that the ultimate goal of prediction learning is to assist in selecting actions. In reinforcement learning, a fundamental challenge is the problem of credit assignment -

determining which actions in a sequence led to a particular outcome, especially when rewards are delayed. This issue is crucial for improving the behavioral policy, as actions leading to rewards should be reinforced while those leading to punishments should be avoided. The credit assignment problem becomes particularly complex in scenarios where actions have long-term consequences or when multiple actions are required to achieve an outcome. For instance, in a game of chess, the final outcome (win or loss) is the result of a long sequence of moves, and it is not immediately clear which moves were critical in determining the result.

Actor-critic methods offer an elegant solution to this problem by using temporal difference learning to estimate state values and guide action selection. This approach was first introduced by Barto, Sutton, and Anderson in 1983 [1], inspired by neural network models of learning. In this framework, an "adaptive critic element", Critic, learns to estimate state values $V(S)$ using TD learning, while an "associative search element", Actor, learns and maintains a policy $\pi(S, a)$ - a probability distribution over actions for each state.

The key insight of the actor-critic model is that even when external reinforcement is delayed, the TD prediction error can provide a useful learning signal at every time step. This prediction error, given by $\delta_t = r_t + \gamma V(S_{t+1}) - V(S_t)$, serves as a surrogate reinforcement signal for the Actor. A positive prediction error indicates that the chosen action has led to a better state than expected, suggesting that this action should be repeated in similar situations in the future. Conversely, a negative prediction error suggests that the action should be chosen less often. Moreover, in the case of no external reinforcement (ie $r_t = 0$), the prediction error will be $\gamma V(S_{t+1}) - V(S_t)$, which is the comparison of two consecutive states, and can provide

data on whether the chosen action has landed us in a higher-value state than the previous one or not. Using δ_t, the Actor's *policy* - a probability distribution over all available actions at each state $\pi(S,a)_{new} = p(a|S)$, is updated according to the following rule:

$$\pi(S,a)_{new} = \pi(S,a)_{old} + \eta_\pi \delta_t, \tag{2.3.5}$$

where η_π is the policy learning rate and δ_t is the TD prediction error. This allows the Actor to learn to select actions that lead to higher-valued states, effectively solving the credit assignment problem. Actor-critic methods have been closely linked to the function of the basal ganglia in the brain, particularly in relation to instrumental action selection and Pavlovian prediction learning. The Critic is often associated with the ventral striatum and related structures, which are thought to be involved in reward prediction and evaluation. The Actor, on the other hand, is linked to the dorsal striatum and motor cortical areas, which play a crucial role in action selection and execution. This mapping aligns with our understanding of how the basal ganglia contribute to both learning and action selection in the brain, and has been supported by various neurophysiological and neuroimaging studies. We will elaborate on the neural correlates of RL in the next section. While actor-critic models have provided valuable insights into basal ganglia function, some researchers have raised important questions about their anatomical plausibility. Joel et al. (2002) [20] provide a critical review of several actor-critic models of the basal ganglia, highlighting discrepancies between model assumptions and known neuroanatomy. They argue that many implementations of the critic in basal ganglia circuitry rely on anatomical assumptions that are not well-supported, particularly in primates. As an alternative, they propose a "reinforcement driven dimensionality reduction" (RDDR) model that aims

to better account for basal ganglia anatomy and physiology. This work underscores the importance of grounding computational models in biological constraints and suggests promising directions for refining actor-critic architectures to more closely match brain structure and function. Even though actor-critic methods have shown success in many applications, they are not guaranteed to converge to an optimal policy in all cases. Nevertheless, they remain one of the strongest links between reinforcement learning theory and neurobiological data on decision-making in animals and humans.

An alternative to actor-critic methods is to learn state-action values, denoted as $Q(S, a)$, which represent the expected future reward of taking a specific action in a given state. This approach, known as $\mathcal{Q}$-learning, was introduced by Watkins in 1989 [55]. $\mathcal{Q}$-learning allows for direct action selection by choosing the action with the highest $\mathcal{Q}$-value in each state. The $\mathcal{Q}$-values are updated using the following rule:

$$Q(S_t, a_t)_{new} = Q(S_t, a_t)_{old} + \eta\delta_t, \tag{2.3.6}$$

where the prediction error δ_t is computed slightly differently:

$$\delta_t = r_t + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, a_t), \tag{2.3.7}$$

where the max operator indicates that the temporal difference is calculated based on the action believed to be the optimal choice at the next state S_{t+1}. This "off-policy" method considers the best possible future action when computing the prediction error, even if this is not the action that will actually be taken.

These state-action value methods have also found support in neuroscience research. Recent studies in non-human primates and rats suggest that dopaminergic

neurons may be conveying prediction errors based on state-action values, rather than state values as in the actor-critic model. Some evidence supports a $\mathcal{Q}$-learning-like prediction error, while other findings point to alternative formulations of state-action value-based learning. The diversity of these models and their neural correlates highlights the complexity of reinforcement learning in the brain and suggests that multiple mechanisms may be at play in different contexts or brain regions. It is possible that the brain uses a combination of these strategies, perhaps employing different approaches depending on the specific task, the level of uncertainty in the environment, or the stage of learning. Understanding how these different mechanisms interact and are implemented in neural circuits remains an active area of research in both neuroscience and artificial intelligence.

2.3.2 The Rewarding Brain: Reinforcement Learning in The Brain

The application of reinforcement learning (RL) models to neuroscience has provided crucial insights into the neural mechanisms underlying learning and decision-making. Of particular significance is the discovery of neural correlates of RL processes, especially in the dopamine system. Groundbreaking research by Wolfram Schultz and colleagues in the 1990s revealed that dopaminergic neurons in the midbrain exhibit firing patterns remarkably consistent with RL theory, specifically temporal difference (TD) learning [44, 42]. These neurons were found to encode a 'prediction error' signal - a key concept in RL. Initially, they responded strongly to unexpected rewards. However, as animals learned to associate specific cues with subsequent rewards, the

neuronal response shifted to the predictive cue itself. Importantly, if a predicted reward was omitted, these neurons showed a dip in their firing rate at the expected time of reward delivery. This pattern closely mirrors the TD error signal in computational RL models, suggesting that dopamine neurons might broadcast a teaching signal used to update value estimates throughout the brain. These findings not only challenged the simplistic "dopamine equals reward" hypothesis, but also provided a neural substrate for RL processes. Subsequently, similar RL-like signals have been identified in other brain areas, including the striatum, prefrontal cortex, and habenula, forming a distributed network involved in value-based learning and decision-making. This convergence of computational theory and neurophysiology has significantly advanced our understanding of how the brain implements reinforcement learning mechanisms.

Dopamine as a Neural Currency of Prediction

The reward prediction error hypothesis of dopamine, proposed by Montague, Dayan, and Sejnowski (1996) [28], is one of the most successful applications of computational theory to neuroscience. This hypothesis posits that the phasic activity of midbrain dopamine neurons encodes a reward prediction error signal analogous to the temporal difference (TD) prediction error.

Extensive and compelling evidence supports this hypothesis. Dopamine neurons exhibit phasic activation to unexpected rewards early in learning, and as learning progresses, this activation shifts to the earliest reliable predictor of reward, precisely as TD learning predicts. Dopamine neurons increase their firing rate for positive prediction errors and decrease their firing rate below baseline for negative prediction errors, with the magnitude of responses scaling with the size of the prediction error.

When cues predict delayed rewards, the magnitude of the dopamine response to the cue decreases with longer delays, consistent with temporal discounting in TD learning. The evolution of dopamine responses over the course of learning closely matches the dynamics predicted by TD learning models. The temporal precision of dopamine signals further bolsters their consistency with TD learning, as dopamine neurons respond to reward prediction errors with remarkably short latencies, often within 50-100 ms after stimulus onset. This rapid signaling is crucial for real-time learning and decision-making, allowing swift updates of value estimates in dynamic environments.

Causal manipulations of dopamine signaling also provide compelling evidence for its role in learning. Advanced techniques like optogenetics have shown that activation of dopamine neurons can effectively substitute for actual rewards in classical conditioning paradigms. Conversely, inhibiting dopamine neurons at the time of expected reward significantly impairs learning. These findings underscore the causal relationship between dopamine signaling and reinforcement learning processes.

Niv's review on RL in brain [30], highlights several key experiments that have solidified our understanding of dopamine's role in reward prediction. For instance, studies by Bayer and Glimcher [3] have shown that the contribution of previously experienced rewards to the current dopaminergic response follows an exponentially weighted average of past experiences, aligning perfectly with TD learning principles. Furthermore, investigations into probabilistic rewards have revealed that dopamine responses to predictive cues scale with reward probability, while responses to the rewards themselves inversely scale with probability [13]. This nuanced encoding of uncertainty and expectation in dopaminergic signaling provides a neural substrate for

the complex computations required in reinforcement learning.

Actor/Critic Models and The Basal Ganglia

The reward prediction error hypothesis naturally led to the proposal that the basal ganglia implement an actor-critic architecture for reinforcement learning, as detailed by Joel et al. (2002) [20]. In this framework, the ventral striatum, particularly the nucleus accumbens, acts as the "critic," learning to predict future rewards based on current states. The dorsal striatum, on the other hand, functions as the "actor," learning to select actions that maximize predicted rewards. Dopamine signals from the midbrain serve as the common currency for updating both critic and actor components, providing a unified mechanism for learning and action selection.

This model is supported by converging evidence from multiple research approaches. Anatomical studies have revealed that the ventral and dorsal striatum receive topographically organized inputs from cortical and limbic areas, allowing them to represent states and actions, respectively. This anatomical organization aligns perfectly with the proposed functional division in the actor-critic model. Functional imaging studies, particularly fMRI investigations in humans, have shown that the ventral striatum responds robustly to reward prediction errors, while the dorsal striatum is predominantly involved in action selection and learning. These findings provide a direct link between the computational principles of reinforcement learning and the functional organization of the basal ganglia. Lesion studies have further corroborated this framework. Damage to different parts of the basal ganglia produces dissociable deficits in learning and decision-making that are consistent with the actor-critic model. For instance, lesions to the ventral striatum impair the ability to learn reward

predictions, while dorsal striatal lesions specifically affect action selection and skill learning.

Rewarding Synchronization

Recent work has begun to explore the role of reward and reinforcement processes in sensorimotor synchronization (SMS). The study by Matthews et al. (2020) [25] provides intriguing insights into the neural correlates of groove, defined as the pleasurable desire to move to music. Their findings reveal that medium-complexity rhythms, which elicit the strongest groove sensation, activate both motor regions involved in beat perception and reward-related areas, including the nucleus accumbens and medial orbitofrontal cortex. This research suggests that the rewarding aspects of SMS may engage similar neural circuits as those involved in other forms of reward-based learning. The authors propose a model in which different cortico-striatal circuits interact to support groove. A "motor" circuit involving the putamen and supplementary motor area is thought to be responsible for internal generation of the beat. A "cognitive" circuit including the caudate and prefrontal cortex is implicated in updating beat-based expectations. Finally, a "limbic" circuit involving the nucleus accumbens and orbitofrontal cortex is proposed to assign affective value to rhythmic patterns. This framework integrates reinforcement learning principles with the sensorimotor aspects of rhythm perception and production. It suggests that the pleasure derived from synchronizing with a beat may serve as an intrinsic reward signal, potentially driving learning and refinement of SMS skills. This intrinsic reward could explain why humans are motivated to engage in rhythmic activities and continually improve their synchronization abilities.

While the work of Matthews et al. (2020) [25] focuses on the rewarding aspects of SMS, it is crucial to consider how the brain represents and processes time intervals, which is fundamental to accurate synchronization. In this context, the work of Gershman et al. (2014) [14] provides valuable insights into the role of interval timing within the framework of reinforcement learning (RL) and semi-Markov processes. They proposed that the basal ganglia could model interval timing using state transitions that occur at variable intervals, similar to the dynamics observed in RNNs. Gershman and colleagues argued that the basal ganglia's ability to represent time intervals as states in a semi-Markov process aligns well with the properties of RNNs, which can maintain temporal information across variable lengths of time. This perspective highlights the flexibility of the basal ganglia in adapting to different timing tasks through state-dependent processes. The interaction between reinforcement learning algorithms and state-dependent timing could provide insights into the neural mechanisms underlying both the motor and reward aspects of SMS.

By integrating these two lines of research, we can begin to construct a more comprehensive model of SMS that accounts for both the rewarding nature of rhythmic synchronization and the precise temporal computations required to achieve it. The cortico-striatal circuits proposed by Matthews et al. (2020) could potentially implement the state-dependent timing processes described by Gershman et al. (2014), with the basal ganglia playing a central role in both timing and reward processing. This integrated view offers a promising direction for future research, potentially bridging the gap between traditional RL models and more dynamic, state-dependent models of timing in the context of sensorimotor synchronization. It suggests that the pleasure

derived from synchronizing with a beat may serve as an intrinsic reward signal, driving learning and refinement of SMS skills through reinforcement learning mechanisms. Simultaneously, the flexible timing abilities emerging from neural network dynamics could explain the remarkable adaptability of human synchronization abilities across various tempos and rhythmic complexities.

Chapter 3

Methodology

This chapter outlines the methodology employed in our study of sensorimotor synchronization (SMS) using deep reinforcement learning. We present a novel approach to modeling time-related behaviors, specifically metronome tapping synchronization, using recurrent neural networks (RNNs) and reinforcement learning (RL). Our method is inspired by neuroscience literature and aims to bridge the gap between computational models and observed human behaviors in SMS tasks.

The study of SMS is crucial for understanding how humans and artificial agents perceive and interact with rhythmic stimuli. By applying deep RL techniques to this domain, we aim to shed light on the underlying mechanisms of temporal processing and motor coordination. This approach not only allows us to model SMS behavior but also provides insights into the learning processes that might be at play in biological systems. Our choice of using RNNs with Long Short-Term Memory (LSTM) units in our RL agents is based on theoretical reasoning, which is consistent with contemporary neuroscientific insights. This approach offers several benefits for modeling timing and decision-making processes:

- LSTMs can learn and remember dependencies over long sequences, crucial for tasks involving timing and sequential decision-making. This aligns with the dynamic state representation observed in state-dependent network models of timing, where time is encoded in the evolving states of neural circuits.

- RNNs can adapt to various timing tasks without explicit timing mechanisms, mirroring the flexible timing observed in the brain.

- In RL, LSTM units can help assign credit to actions based on their long-term consequences, improving learning efficiency in environments with delayed rewards.

- RNNs with LSTM units can maintain an internal state that captures relevant history, making them ideal for non-Markovian environments where the current state does not fully capture the interaction history.

Empirical studies support this approach. Bi and Zhou (2020) [6] demonstrated that RNNs could effectively model the computation of time in neural networks. Jazayeri and Shadlen (2015) [19] showed that the brain uses a form of preplanning to coordinate sensorimotor functions, a process that can be mirrored in RL agents using RNNs with LSTM to plan and execute timed actions. Deverett et al. (2019) [12] found that recurrent agents with LSTM units demonstrated near-perfect accuracy in interval timing tasks and could generalize timing rules to new intervals. The use of LSTM-based agents is thus not merely a computational convenience, but a theoretically motivated choice that aligns with current neuroscientific understanding.

Moreover, our methodological approach draws partial inspiration from the groundbreaking work of Betancourt et al. (2023) [5] on training monkeys in to synchronize.

Their study utilized a synchronization-continuation task (SCT) where monkeys synchronized taps to auditory or visual cues with IOI ranging from 450 ms to 850 ms, using juice to reward them for maintaining a specified interval error and asynchrony thresholds. They demonstrated that non-human primates could learn to synchronize tapping movements with both auditory and visual metronomes through reinforcement. Building on these insights, we developed our computational framework that employs a similar task environment to model the learning processes underlying sensorimotor synchronization.

The following sections provide a detailed description of our task design, proposed model architecture, environment setup, agent specifications, reward function formulations, training procedures, and the RL algorithm used.

3.1 Task Design

Our study focuses on a metronome tapping synchronization task, a paradigm widely used in SMS research. In this task, the agent receives a series of beats and is required to synchronize its tapping with the metronome after the third beat. The taps during the first three beats are not recorded or rewarded, allowing the agent to acclimate to the rhythm before being evaluated. The task mechanism operates on a fine-grained timestep resolution of 0.01 seconds, allowing for precise temporal control and measurement. At the beginning of each episode, an inter-onset interval (IOI) is randomly selected from a range of 400ms to 700ms, determining the rhythm of the metronome beats for that particular trial. The metronome input is presented to the agent as a binary signal, where a value of 1 represents a beat and 0 indicates the absence of a

beat. In response, the agent can choose between two actions at each timestep: initiating a tap (represented by 1) or not taking any action (represented by 0). To simulate the realistic delay between a decision to tap and the actual motor execution, the task incorporates a fixed delay between the agent's action initiation and the resulting tap execution. This design closely mimics the temporal dynamics and decision-making processes involved in human sensorimotor synchronization tasks, providing a robust framework for investigating timing-related behaviors in artificial agents.

This design incorporates key features of human SMS behavior, including the need for temporal prediction and the inherent delay between motor command and action execution. The range of IOIs (400-700ms) is chosen to reflect a tempo range commonly used in SMS studies. This allows for direct comparisons between our model's performance and existing data. The fixed delay between action initiation and tap execution is a crucial feature of our task design. It simulates the neural and biomechanical delays present in human motor systems, adding a layer of realism to our model. This delay challenges the agent to not only predict when a beat will occur but also to initiate its action in advance to achieve synchronization.

3.2 Proposed Model

Our model utilizes a recurrent neural network architecture to capture the temporal dependencies inherent in the SMS task. Specifically, we employ a single-layer Long Short-Term Memory (LSTM) network. This choice is motivated by the LSTM's ability to learn and remember long-term dependencies, which is crucial for temporal tasks like metronome synchronization. The model's architecture is as follows:

- **Input layer:** Accepts the binary input representing metronome beats

- **LSTM layer:** A single layer of LSTM units (various unit numbers were tested during training)
- **Output layer:** A fully connected layer that produces the binary action output

The model's input (state) is a binary one-dimensional array representing the metronome beats. The output (action) is also a binary one-dimensional array indicating the agent's chosen action (tap or no tap). The LSTM layer is key to our model's ability to process temporal sequences. Unlike feedforward networks, the LSTM can maintain information about past inputs, allowing it to detect and utilize temporal patterns in the metronome sequence. This is particularly important for SMS tasks, where the timing of future beats must be predicted based on the pattern of past beats.

3.2.1 The Environment

The environment is designed to simulate a metronome tapping task with several key characteristics. Each episode has a variable length, which is determined by the selected inter-onset interval (IOI) for that particular trial (8 times the IOI). Throughout the episode, the metronome stimulus is presented to the agent as a binary input at every 0.01-second timestep, providing a high-resolution temporal framework. The tapping mechanism incorporates a realistic fixed delay between the agent's action selection and the actual tap execution, mimicking the neural and biomechanical delays present in biological systems. Importantly, rewards are not provided at the moment of action selection, but rather at the moment when the tap actually lands. The delayed reward mechanism is a crucial feature of our environment. By providing rewards at the moment of tap landing rather than at action selection, we create a more challenging and realistic learning scenario. This design choice forces the agent to learn to

anticipate the consequences of its actions, much like humans must do in real-world SMS tasks. This design creates a challenging and dynamic environment that closely mirrors the conditions of sensorimotor synchronization experiments conducted with human participants. The design of our environment reflects the non-Markovian nature of many real-world timing tasks, where the information from the current state that can presently be directly observed alone is not sufficient for optimal decision-making. This aligns with our use of RNNs with LSTM units, which can maintain an internal state capturing relevant history.

3.2.2 The Agent

Our agent is implemented as a recurrent neural network, specifically utilizing a single-layer Long Short-Term Memory (LSTM) architecture. The agent receives input in the form of a binary one-dimensional array representing metronome beats, or in an extended configuration, a two-dimensional array that includes both beats and action feedback. The output of the network is a binary one-dimensional array indicating whether to tap or not. The agent operates in a discrete action space, choosing between tapping and not tapping. Its state space is also discrete and binary, representing the presence or absence of beats at each timestep. The agent's primary task is to learn a policy that effectively maps these discrete sequences of metronome beats to appropriate tapping actions. Despite the discrete nature of the input, the LSTM architecture allows the agent to maintain an internal representation of the beat sequence, facilitating anticipation of future beats and precise timing of actions. Figure 3.1 includes a scheme of the whole RL model.

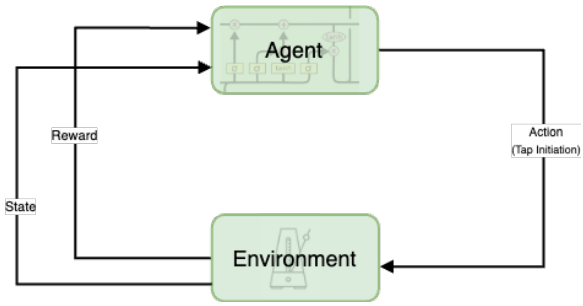


Figure 3.1: Reinforcement learning framework for sensorimotor synchronization. The agent receives state information and rewards from the environment, and initiates tapping actions. The environment processes these actions and provides updated state and reward signals, creating a closed-loop interaction between the agent and its sensorimotor task environment

3.2.3 The Reward Function

Our study investigates various reward policies to explore their impact on the agent's learned behavior in the sensorimotor synchronization task. These policies are designed to test different aspects of synchronization and timing accuracy, allowing us to compare how various reward structures influence the agent's performance and strategy development.

It is important to note that in our RL framework, the 'rewards' are calculated to be negative values, with 0 representing the optimal outcome. While this formulation might more intuitively be described as a 'penalty' system, we maintain the terminology of 'rewards' to align with standard RL conventions. In RL literature, 'reward' is the conventional term used to describe the signal provided to the agent, regardless of whether the values are positive or negative. Hence, this choice of terminology ensures clarity and coherence with established RL practices.

- The first policy, the *next beat* reward policy, focuses on the asynchrony between

the agent's tap and the next metronome beat, calculated as the negative absolute time difference between these two events. This approach encourages the agent to anticipate upcoming beats and tap slightly ahead of them, mirroring the anticipatory behavior often observed in human participants. A perfect tapping behaviour results in 0 as reward.

$$reward = t_{tap} - t_{next\_beat} \tag{3.2.1}$$

- The second policy, namely the *nearest beat* reward policy, considers the asynchrony relative to the nearest beat, rather than the next beat. This allows for a more flexible synchronization strategy, potentially accommodating both anticipatory and reactive tapping behaviors.

$$reward = -|t_{tap} - t_{nearest\_beat}| \tag{3.2.2}$$

- Our third policy, *next/nearest beat + interval*, combines asynchrony with interval accuracy. In addition to rewarding synchronization with reference beats (next vs. nearest), this policy provides an additional reward if the agent's inter-tap interval closely matches the metronome's inter-onset interval (within a 16% error margin). This dual reward structure encourages both precise synchronization and consistent tempo matching, addressing two key aspects of skilled sensorimotor synchronization.

$$reward = reward_{asynchrony} + reward_{interval} \tag{3.2.3}$$

- To discourage inaction and promote consistent engagement with the task, we implement two penalty mechanisms. First, a missed tap incurs a penalty proportional to the current inter-onset interval. This scaling ensures that the penalty is contextually appropriate across different tempi, reflecting the increased difficulty of maintaining synchronization at slower tempi.

$$Missed\ tap\ penalty = -\frac{interval \times (interval - 1)}{2} + 1 \tag{3.2.4}$$

  Second, a fixed huge penalty is applied if the agent makes no taps at all during an episode, strongly discouraging complete inaction.

Importantly, rewards are not provided at the moment of action selection, but at the moment of tap landing. This delay in reward delivery creates a more challenging learning environment that more closely mimics the temporal dynamics of real-world sensorimotor tasks. By comparing the agent's performance and learned strategies across these different reward policies, we aim to gain insights into how various aspects of reward structure influence the development of sensorimotor synchronization skills. For instance, the comparison between next beat and nearest beat asynchrony allows us to explore whether the agent learns to anticipate or react to beats. The inclusion of interval accuracy in one reward function enables us to study how the agent balances synchronization and regularity in its tapping behavior. The results may provide valuable insights into the reward mechanisms that might underlie human sensorimotor learning and performance, potentially informing both artificial intelligence design and our understanding of human motor control and timing.

3.3 Training

3.3.1 Reinforcement Learning Algorithm

We employ a recurrent version of the Proximal Policy Optimization (PPO) algorithm for training our agent, specifically the implementation provided in the stable-baselines3 contrib version. [35]

PPO [43] is an on-policy algorithm that belongs to the family of policy gradient methods in reinforcement learning. It aims to improve the stability and sample efficiency of policy gradient methods by limiting the size of policy updates. The key idea behind PPO is to ensure that the new policy doesn't deviate too far from the old policy, which helps prevent catastrophic drops in performance during training. This is achieved through a clipped objective function:

$$L^{CLIP}(\theta) = \hat{E}_t\Big[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\Big], \tag{3.3.1}$$

where θ represents the policy parameters, $r_t(\theta)$ is the probability ratio between the new and old policy: $\frac{\pi(a_t|s_t)}{\pi_{\theta\ old}(a_t|s_t)}$, $\hat{A}_t$ is the estimated advantage function (which is the value for a selected action, used by the agent to choose the better one), and ϵ is a hyperparameter, typically set to 0.2. The clip function restricts $r_t(\theta)$ to the interval $[1-\epsilon, 1+\epsilon]$, which prevents excessively large policy updates. This clipping mechanism is a key innovation of PPO, providing a simple yet effective way to constrain policy updates. By limiting the magnitude of policy changes, PPO helps maintain stability during training and avoids catastrophic performance drops that can occur with unconstrained policy updates. The algorithm alternates between sampling data through interaction with the environment and optimizing this surrogate objective

function using stochastic gradient ascent:

$$\theta_{k+1} = \max_{\theta} \hat{E}_t[L^{CLIP}(\theta)]. \tag{3.3.2}$$

This iterative process allows the agent to gradually improve its policy based on its experiences in the environment. The use of the clipped objective ensures that these improvements are made conservatively, striking a balance between exploration and exploitation. PPO also typically includes an entropy bonus to encourage exploration:

$$L^{CLIP} + S(\theta) = \hat{E}_t[L^{CLIP}() + \beta S\pi_\theta] \tag{3.3.3}$$

Where S denotes the entropy of the policy and β is a coefficient. This entropy term helps prevent premature convergence to suboptimal deterministic policies by encouraging the agent to maintain a level of stochasticity in its actions. This is particularly important in complex environments where exploration is crucial for discovering optimal strategies.

The PPO algorithm offers several advantages that make it suitable for our task. Its stability, provided by the clipping mechanism, ensures more consistent training compared to other policy gradient methods. This is particularly valuable in complex environments where training instability can be a significant challenge. Moreover, PPO's compatibility with recurrent policies is crucial for our work in sensorimotor synchronization. The ability to easily adapt PPO to work with recurrent neural networks allows us to capture and learn from the temporal dependencies inherent in our task. This is essential for an agent that must understand and respond to rhythmic patterns and timing cues.

3.3.2 Training Configurations

The reinforcement learning environment for our sensorimotor synchronization task was implemented using the Gymnasium library [52], which provides a flexible and standardized interface for defining and interacting with RL environments. This allowed for seamless integration with our chosen other libraries. For the implementation of the reinforcement learning algorithms, we utilized the stable-baselines3 contrib library [35]. This library offers implementations of state-of-the-art RL algorithms, including the PPO algorithm used in our study, with support for recurrent policies. The RL agent's neural network architecture consisted of a single LSTM layer with a hidden size of 128 units, allowing the agent to capture and utilize temporal dependencies in the sensorimotor synchronization task. All experiments were run using Python 3.9.13.

Training was conducted on a MacBook Air 2022, equipped with an Apple M2 Chip and 16 GB of memory. This setup provided adequate computational power to train the model efficiently while maintaining a relatively low hardware requirement. Training times varied depending on the complexity of the reward policy, ranging from approximately 36 hours for the simplest policy to 100 hours for the most complex. Convergence was determined based on the stability of the reward curve and the agent's performance on test episodes.

Hyperparameter optimization was crucial to enhance the performance of the RL agent. The hyperparameters for each reward policy were individually optimized to achieve the best performance. Key hyperparameters included:

- **Entropy coefficient:** Adjusted to balance exploration and exploitation
- **Learning rate:** Tuned to ensure stable and efficient learning

- **Batch size:** Optimized for each policy to balance computational efficiency and learning stability

The specific values for these hyperparameters varied across the different reward policies and were determined through manual tuning.

Chapter 4

Results and Discussion

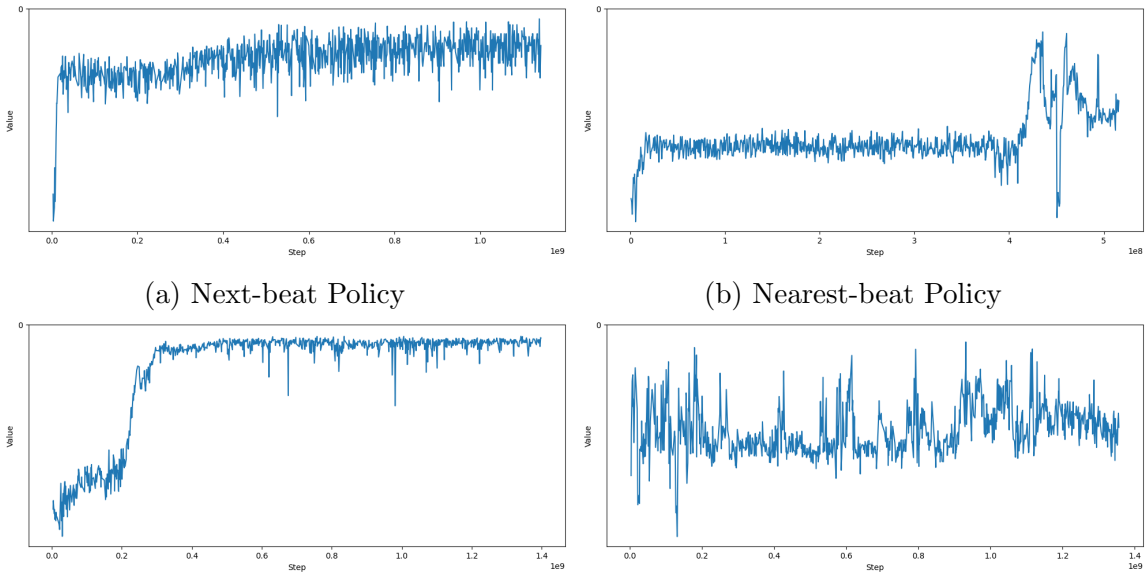
This chapter presents the results of training reinforcement learning (RL) agents using different reward policies for sensorimotor synchronization tasks. The agents were evaluated on their ability to synchronize with a steady metronome as well as their responses to perturbations in the form of event onset shifts. The findings provide insights into how different reward structures influence the development of synchronization behavior in artificial agents, with implications for understanding human sensorimotor synchronization.

4.1 Training Process

As mentioned before, four different reward policies were investigated:

- Next-beat asynchrony reward
- Nearest-beat asynchrony reward
- Next-beat asynchrony + interval accuracy reward

- Nearest-beat asynchrony + interval accuracy reward



(a) Next-beat Policy

(b) Nearest-beat Policy

(c) Next-beat + Interval Policy

(d) Nearest-beat + Interval Policy

Figure 4.1: Learning Curves For Each Trained Agent

Figure 4.1 shows the learning curves for agents trained with each reward policy, as the episodes' mean reward over the total number of steps taken in all episodes.

The next-beat asynchrony reward policy (Figure 4.1a) resulted in the fastest learning and most stable performance during training. The agent learned to minimize asynchrony relative to the next metronome beat. In contrast, the nearest-beat asynchrony reward (Figure 4.1b) led to slower and less stable learning, with the agent converging on a sub-optimal solution. Encouraging exploration also won't make the agent move from that point, and makes the learning process even more unstable.

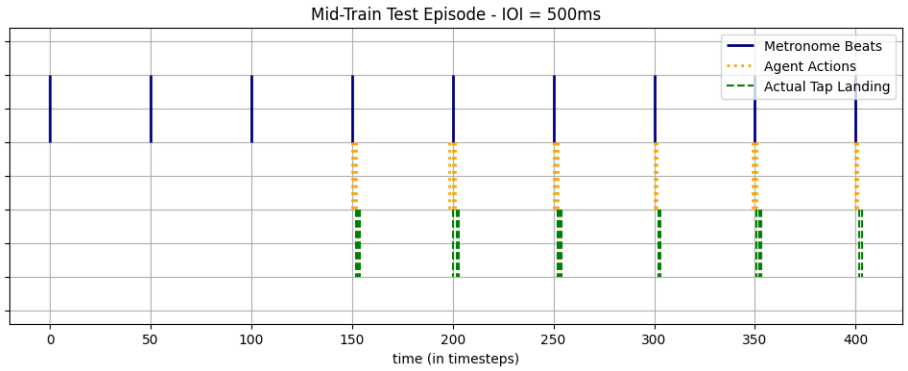
Adding an interval accuracy component to the reward function (policies 3 and 4) increased the complexity of the learning task. The next-beat + interval policy (Figure 4.1c) showed intermediate learning speed and stability, and eventually converges to

the optimal solution. The nearest-beat + interval policy (Figure 4.1d) had the slowest learning progression and highest variability in performance during training. Due to its poor performance, we decided to abandon this policy for further tests and continue with well-trained agents with acceptable performances in the tasks.

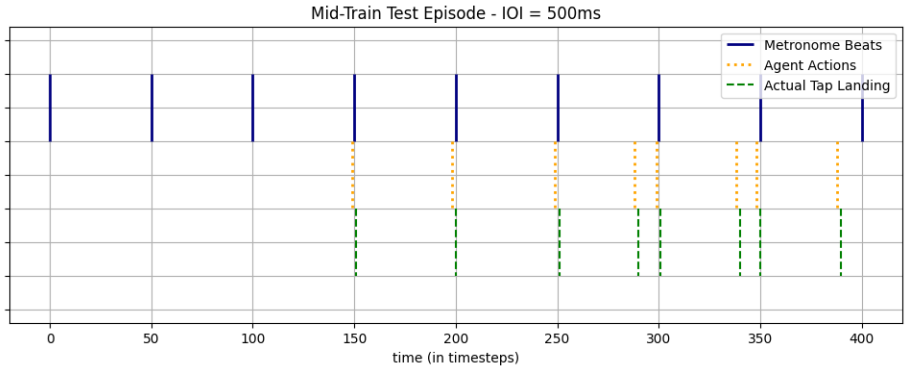
These results highlight how the structure of the reward signal significantly impacts the learning dynamics and ultimate behavior of the RL agents. The simpler, more directed next-beat reward facilitates rapid acquisition of basic synchronization, while more complex reward schemes pose greater challenges for the learning algorithm. However, it's important to note that while we aim to avoid unnecessary complexity in reward structures, we must also ensure that the reward policy includes all features required for robust and adaptable performance across various tasks. As we will see in subsequent tests, although the next-beat policy leads to faster learning and better performance in the basic synchronization task, the next-beat+interval policy results in a more robust agent capable of adapting to different scenarios. This underscores the delicate balance between simplicity for efficient learning and complexity for comprehensive skill development in reinforcement learning contexts.

4.2 Metronome Synchronization Task

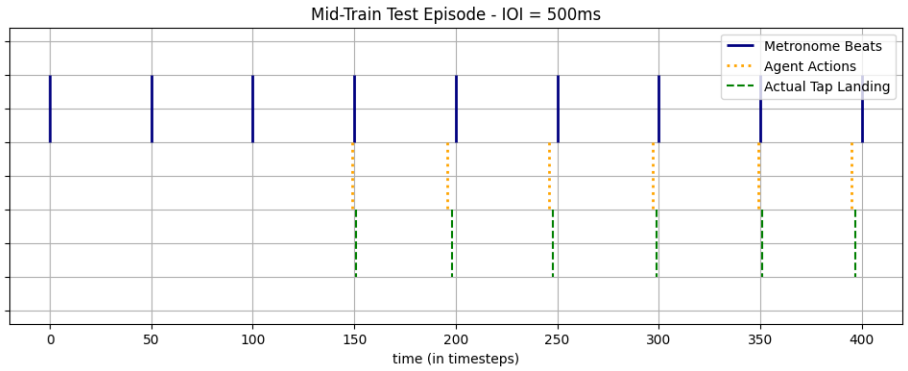
During and after training, the agents were tested on their ability to synchronize with a steady metronome at the training tempo. Figures 4.2 and 4.3 show example episodes of metronome synchronization for each agent, mid-training process and after training respectively.



(a) Nearest-beat Policy

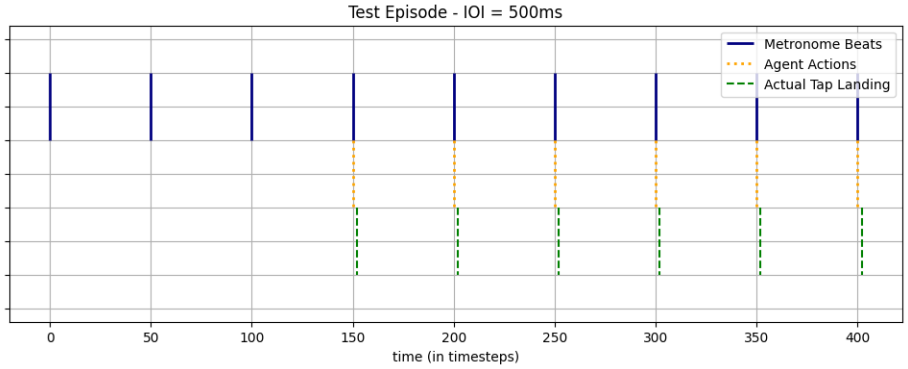


(b) Next-beat Policy

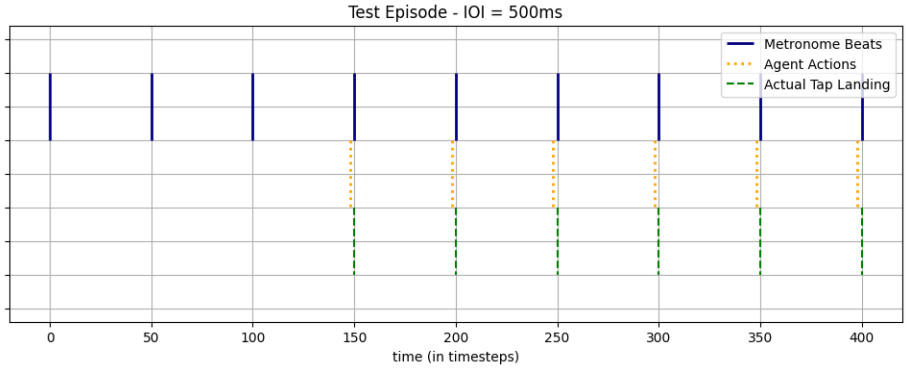


(c) Next-beat + Interval Policy

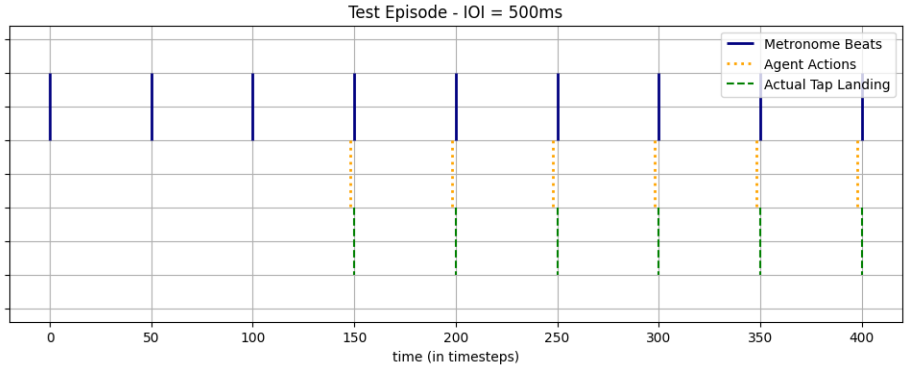
Figure 4.2: Semi-Trained Agents' Performances in a Synchronization Task with IOI=500ms. (4.2a) Nearest-beat Policy: The agent displays sub-optimal synchronization, behaving more like reacting to the beat and making multiple taps. (4.2b) Next-beat Policy: The agent inclines toward tapping ahead of the beat but still makes errors, likely due to ongoing exploration mid-training. (4.2c) Next-beat + Interval Policy: This agent, with extra reward for interval accuracy, demonstrates lower variability in interval consistency compared to the other policies.



(a) Nearest-beat Policy



(b) Next-beat Policy



(c) Next-beat + Interval Policy

Figure 4.3: Fully-Trained Agents' Performances in a Synchronization Task with IOI=500ms. (4.3a) Nearest-beat Policy: This agent has converged on a sub-optimal solution of reacting to the beat rather than anticipating it. (4.3b) Next-beat Policy: The agent has successfully learned to anticipate and synchronize with the metronome's beats. (4.3c) Next-beat + Interval Policy: Similar to (4.3b), this agent demonstrates successful anticipation and synchronization, potentially with improved interval consistency and robustness due to the additional reward component.

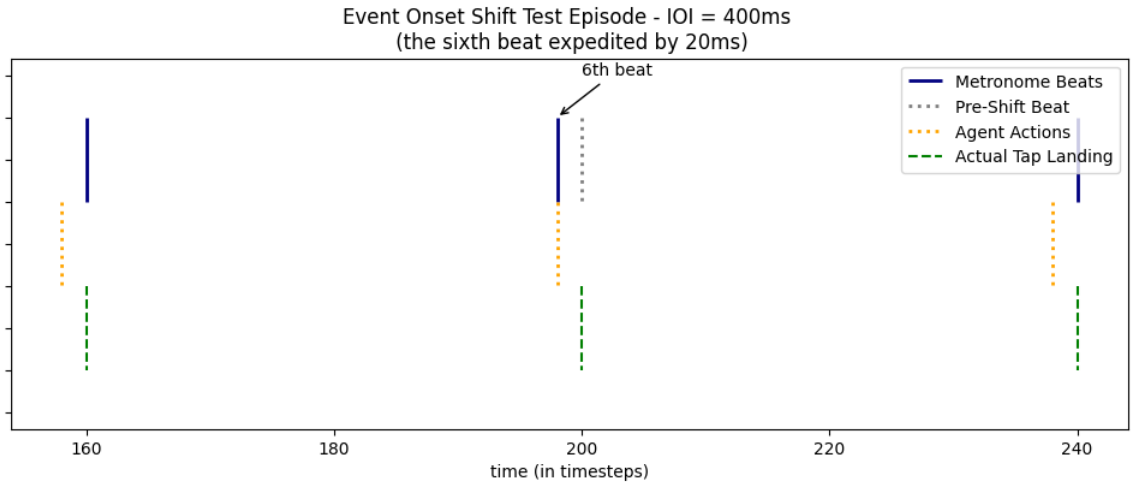
4.3 Event Onset Shift Task

Based on the initial training results and steady metronome synchronization performance, we made a strategic decision to focus our subsequent analyses on the most promising agents. Specifically, we continued our investigation with the next-beat asynchrony agent (1) and the next-beat + interval agent (3), as these demonstrated stable learning and effective synchronization behavior. The nearest-beat asynchrony agent (2) and the nearest-beat + interval agent (4) were excluded from further testing due to their sub-optimal performance and inconsistent behavior.

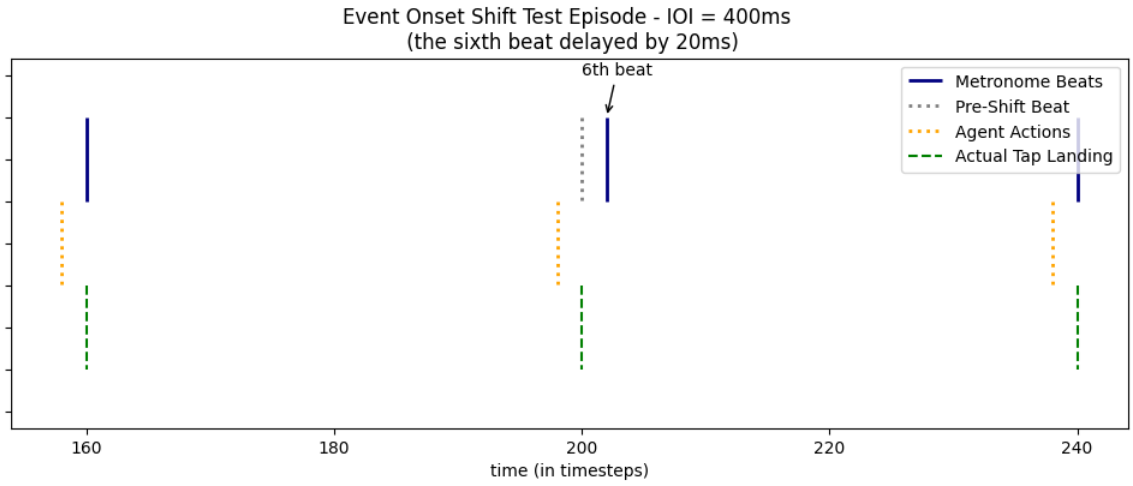
To assess the agents' ability in recovery of synchronization following a perturbation, we tested their responses to event onset shifts in the metronome sequence. Shifts of varying magnitudes ($\pm 5\%$, $\pm 10\%$ of the inter-onset interval) were introduced, and the agents' tap timing adjustments were analyzed. Figures 4.4 and 4.5 shows example perturbation responses for each agent.

The next-beat asynchrony agent (Figure 4.4) showed no adjustment to its tapping in response to event onset shifts. This inflexibility suggests the agent learned to produce taps at a fixed interval matching the initial tempo, rather than truly synchronizing with external events. While effective for steady sequences, this strategy fails to adapt to timing changes.

Most notably, the next-beat + interval reward agent (Figure 4.5) demonstrated adaptive responses to event onset shifts that closely resemble human behavior. This agent made appropriate corrections to its tap timing following perturbations, with an asymmetry in the magnitude of corrections that mirrors patterns observed in human subjects. Specifically, the agent made larger corrections when its taps were late (positive asynchrony) compared to when they were early (negative asynchrony).

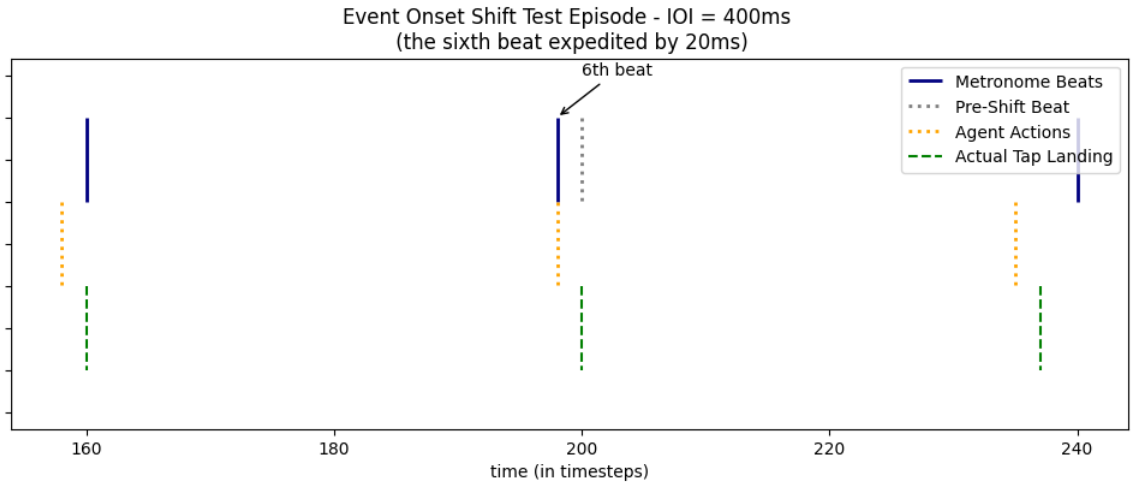


(a)

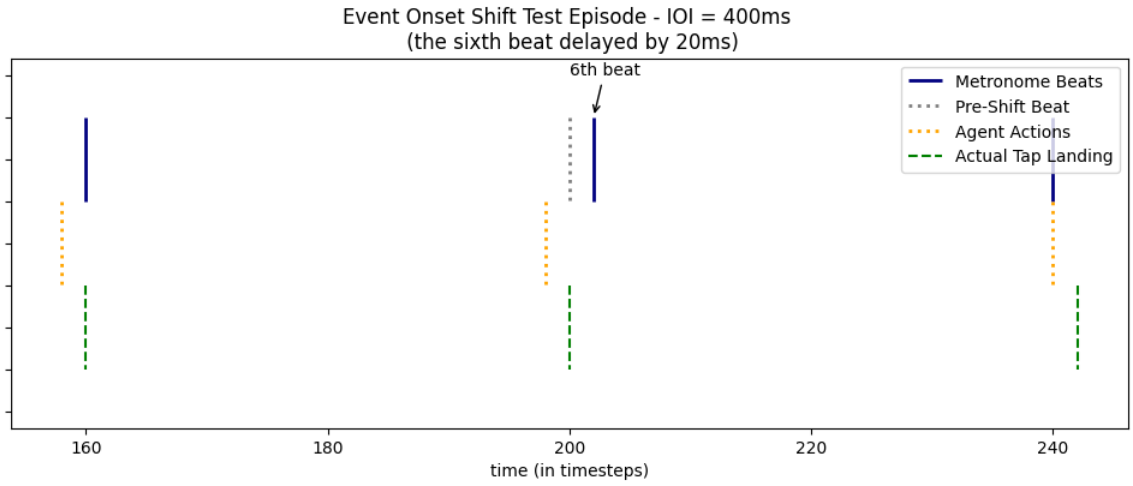


(b)

Figure 4.4: Next-beat Agent's Event Onset Shift Test Episodes. In (4.4a) the sixth metronome beat is expedited by 20ms and in (4.4b), it's delayed by 20ms. The pre-shift position of the shifted beat is denoted by the grey dotted line. (Only the shifted beat and its adjacent beats are shown for better clarity.)



(a)



(b)

Figure 4.5: Next-beat + Interval Agent's Event Onset Shift Test Episodes. In (4.5a) the sixth metronome beat is expedited by 20ms and in (4.5b), it's delayed by 20ms. The pre-shift position of the shifted beat is denoted by the grey dotted line. (Only the shifted beat and its adjacent beats are shown for better clarity.)

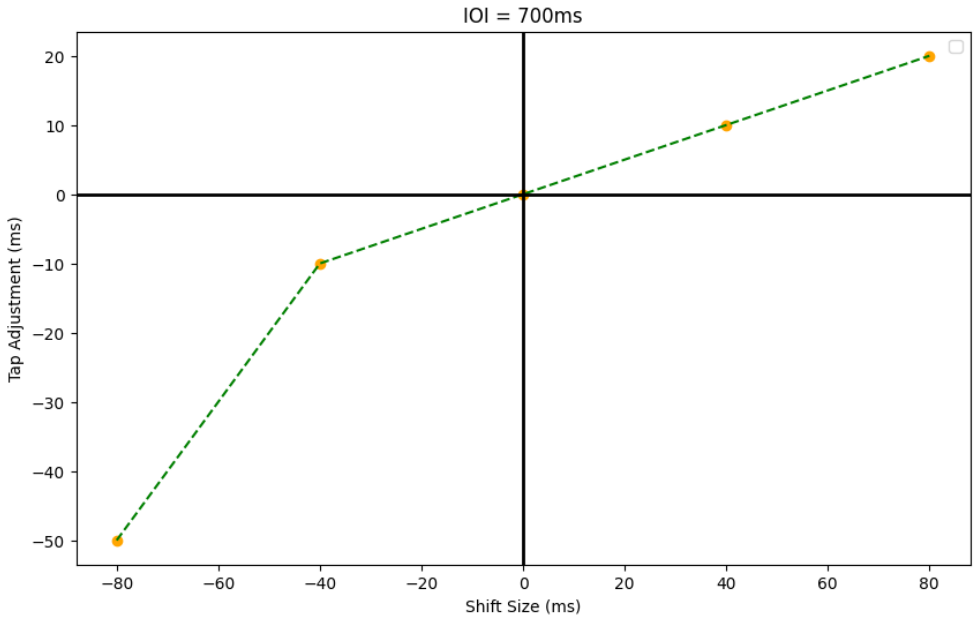


Figure 4.6: Asymmetric Error Correction in the Next-beat + Interval RL Agent at IOI=700ms. This plot demonstrates the agent's tap adjustment as a function of shift size in the metronome beat. The x-axis represents the shift size in milliseconds, with negative values indicating early shifts (expedited beats) and positive values indicating late shifts (delayed beats). The y-axis shows the agent's corresponding tap adjustment in milliseconds. The asymmetric response is evident, with the agent making larger corrections when it is late compared to when it is early.

Figure 4.6 quantifies this asymmetry in error correction across different perturbation magnitudes in a sample test episode. This asymmetric error correction is a documented phenomenon in human sensorimotor synchronization [2, 51, 29]. The emergence of this behavior in our RL agent suggests that the combination of next-beat asynchrony and interval accuracy rewards captures important aspects of the human synchronization learning process. To further validate our model's behavior, we compared our results to empirical data from human subjects performing similar sensorimotor synchronization tasks. Figure 4.7 shows data from a study [29] that demonstrates asymmetric error correction in human participants.

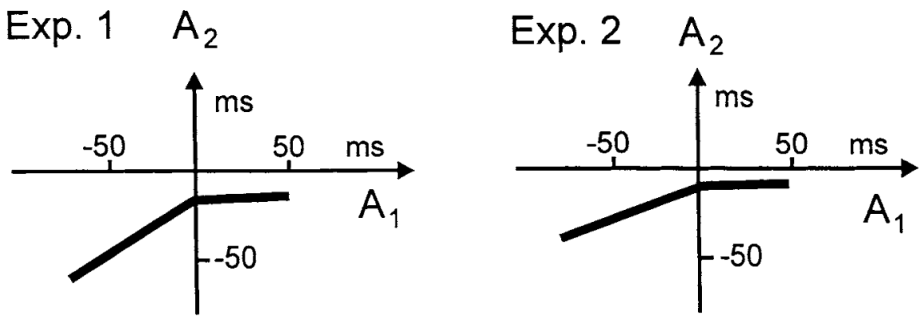


Figure 4.7: Reprinted from [29]. Negative asynchronies are usually left alone, while positive ones are quickly adjusted. In both studies, the data plots meet precisely at zero asynchrony, indicating the system can accurately differentiate between early and late timing.

4.4 Discussion

The results of this study offer valuable insights into the development of sensorimotor synchronization through reinforcement learning. The stark differences in performance across reward policies underscore the critical role that reward structure plays in shaping synchronization behavior. Simple, directed rewards can facilitate rapid learning of basic synchronization, while more complex reward schemes may better capture nuanced aspects of human performance.

One of the most intriguing findings is the emergence of human-like adaptation in the agent trained with the next-beat + interval reward policy. This agent demonstrated adaptive behavior closely resembling human sensorimotor synchronization, including the phenomenon of asymmetric error correction. The agent made larger corrections when its taps were late compared to when they were early, mirroring a documented pattern in human subjects. This asymmetry likely emerges from the reward structure itself, where late taps risk missing the next beat entirely and thus incur a more severe penalty.

The appearance of asymmetric error correction in our model suggests that human synchronization abilities may arise from a similar combination of timing goals – minimizing asynchrony with upcoming events while maintaining a consistent interval between actions. It aligns with theories proposing that humans may experience different subjective costs for early vs. late synchronization errors [22, 29]. In real-world scenarios, early actions may often be less costly than late ones, potentially leading to the development of timing strategies biased towards anticipation rather than reaction.

Interestingly, while our model didn't explicitly demonstrate the negative mean asynchrony (NMA) commonly observed in human subjects during steady-state synchronization, it did exhibit behavior that implies a similar underlying mechanism. The next-beat + interval agent, when faced with event onset shifts, showed a tendency to "play it safe" by tapping earlier rather than later. This bias towards earlier responses, while not manifesting as a consistent NMA in steady-state tapping, suggests a similar anticipatory strategy to that seen in humans.

The relationship between this anticipatory bias and the asymmetric error correction observed in our model provides a novel perspective on these phenomena. Both may emerge from the same underlying reward structure that more harshly penalizes late responses. In humans, this anticipatory strategy manifests as the NMA during steady synchronization. In our model, it appears more prominently in adaptive situations, such as responding to perturbations. This finding suggests that the NMA observed in humans might not be a necessary feature of the synchronization process itself, but rather a byproduct of an anticipatory strategy that becomes more evident in challenging or unpredictable timing contexts. Our model thus offers a new computational framework for understanding how these anticipatory behaviors might develop

through experience and feedback, even if they don't always manifest in the same way as in human subjects.

The failure of agents trained with nearest-beat rewards to develop stable, adaptive synchronization highlights the importance of temporal direction in error signals for sensorimotor learning. Human learners likely benefit from clear feedback distinguishing early from late timing errors, allowing for appropriate adjustments to their internal timekeeping processes.

Chapter 5

Conclusion and Future Directions

This thesis has explored the application of deep reinforcement learning to model sensorimotor synchronization (SMS), a fundamental human ability that underlies activities ranging from musical performance to everyday social interactions. By implementing recurrent neural network agents trained with various reward policies, we have demonstrated that reinforcement learning can capture key aspects of human synchronization behavior, including adaptive responses to perturbations and asymmetric error correction.

Our results highlight the critical role of reward structure in shaping synchronization behavior. The emergence of human-like adaptation in agents trained with a combined next-beat and interval accuracy reward suggests that similar reward mechanisms may underlie the development of SMS skills in biological systems. The observed asymmetry in error correction, mirroring patterns seen in human subjects, provides a computational framework for understanding how anticipatory strategies might arise from experience-dependent learning processes.

Importantly, our model offers new perspectives on established phenomena in SMS

research. While we did not observe a consistent negative mean asynchrony (NMA) during steady-state synchronization, our agents exhibited anticipatory behavior in response to perturbations. This suggests that the NMA observed in humans may be a byproduct of a more general anticipatory strategy that becomes evident in challenging timing contexts, rather than a necessary feature of the synchronization process itself.

5.1 Summary of Contributions

The key contributions of this thesis are as follows:

- Development of a novel deep reinforcement learning framework for modeling SMS, using recurrent neural networks with LSTM units to capture temporal dependencies in rhythmic tasks.
- Demonstration that different reward structures lead to distinct synchronization behaviors, with combined next-beat and interval accuracy rewards producing the most human-like performance.
- Emergence of asymmetric error correction in our model, providing a computational account for this documented phenomenon in human SMS.
- New insights into the relationship between anticipatory bias and asymmetric error correction, suggesting a common underlying mechanism driven by the reward structure of the task.
- A flexible computational framework for studying SMS that can be extended to investigate more complex rhythmic behaviors and interpersonal coordination.

5.2 Future Directions

While our work has provided valuable insights into the computational principles underlying SMS, several avenues for future research remain:

- **Increased Action-Tap Interval:** Investigate the impact of varying delays between the agent's action selection and the resulting tap execution. This could involve: (a) Systematically increasing the delay between action choice and tap landing during training to examine how agents adapt their strategies. (b) Exploring how different delay durations affect the agent's ability to synchronize and its error correction mechanisms.

- **Robustness and Noise:** Our agents were trained and tested in noise-free environments, which may have contributed to their performance. Future work should investigate the robustness of these models by introducing various forms of noise during training and testing. This could include: (a) Motor noise: Implementing a jittered metronome during training to simulate the variability inherent in biological motor systems. (b) Sensory noise: Adding uncertainty to the agent's perception of beat timings. (c) Neural noise: Incorporating stochasticity into the neural network's activations.

- **Sensory Feedback:** Introducing feedback from the agent's own taps and manipulating the delay between tap execution and the agent's perception of the feedback could reveal how agents adjust to sensorimotor feedback delays. Humans experience delays in sensory processing, and studying how agents cope with feedback delays would provide deeper insights into error correction mechanisms.

- **Expanded Temporal Range:** Train agents on a wider range of tempi to investigate how synchronization strategies generalize across different time scales. This could reveal insights into the scalar property of timing and the limits of SMS abilities.

- **Complex Rhythmic Patterns:** Extend the model to handle more complex rhythmic patterns with metrical hierarchies. This could shed light on how agents learn to extract and utilize hierarchical temporal structures, a key aspect of musical rhythm perception.

- **Interpersonal Synchronization:** Develop multi-agent models to study the emergence of interpersonal synchronization. This could involve: (a) Training multiple agents to coordinate with each other, simulating ensemble performance. (b) Investigating how different reward structures influence the development of leader-follower dynamics. (c) Exploring how individual differences in timing abilities affect group synchronization.

- **Neural Activity Analysis:** Investigate the internal representations and trajectories of the trained agent's neural network and compare them to existing data from non-human primates performing similar SMS tasks.

- **Synchronization-Continuation Task:** Extend the model to perform synchronization-continuation tasks, where the agent must continue tapping at the learned tempo after the external metronome stops.

In conclusion, while this thesis lays the groundwork for using RL to model sensorimotor synchronization, future work should focus on training agents in more complex, noisy, and interactive environments to fully understand the mechanisms behind robust

synchronization behaviors. By incorporating these additional research directions, we can further bridge the gap between computational models and neurobiological findings, potentially leading to more comprehensive theories of temporal processing in both artificial and biological systems. These investigations could also provide valuable insights into the neural basis of rhythm perception and production, with implications for understanding both normal timing behavior and timing-related disorders.

Bibliography

[1] A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5):834–846, 1983.

[2] M. L. Bavassi, E. Tagliazucchi, and R. Laje. Small perturbations in a finger-tapping task reveal inherent nonlinearities of the underlying error correction mechanism. *Human Movement Science*, 32(1):21–47, 2013.

[3] H. M. Bayer and P. W. Glimcher. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1):129–141, 2005.

[4] D. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

[5] A. Betancourt, O. Pérez, J. Gámez, G. Mendoza, and H. Merchant. Amodal population clock in the primate medial premotor system for rhythmic tapping. *Cell Reports*, 42(10):113234, Oct. 2023. ISSN 22111247. doi: 10.1016/j.celrep.2023.113234. URL `https://linkinghub.elsevier.com/retrieve/pii/S2211124723012469`.

[6] Z. Bi and C. Zhou. Understanding the computation of time using neural network

models. *Proceedings of the National Academy of Sciences*, 117(19):10530–10540, 2020.

[7] F. L. Bouwer, V. Nityananda, A. A. Rouse, and C. Ten Cate. Rhythmic abilities in humans and non-human animals: A review and recommendations from a methodological perspective. *Philosophical Transactions of the Royal Society B*, 376(1835):20200335, 2021.

[8] D. J. Cameron, J. Bentley, and J. A. Grahn. Cross-cultural influences on rhythm processing: reproduction, discrimination, and beat tapping. *Frontiers in Psychology*, 6:366, 2015.

[9] J. J. Cannon and A. D. Patel. How Beat Perception Co-opts Motor Neurophysiology. *Trends in Cognitive Sciences*, 25(2):137–150, Feb. 2021. ISSN 13646613. doi: 10.1016/j.tics.2020.11.002. URL `https://linkinghub.elsevier.com/retrieve/pii/S1364661320302746`.

[10] J. L. Chen, V. B. Penhune, and R. J. Zatorre. Moving on time: brain network for auditory-motor synchronization is modulated by rhythm complexity and musical training. *Journal of cognitive neuroscience*, 20(2):226–239, 2008.

[11] T.-H. Z. Cheng, S. C. Creel, and J. R. Iversen. How do you feel the rhythm: Dynamic motor-auditory interactions are involved in the imagination of hierarchical timing. *Journal of Neuroscience*, 42(3):500–512, 2022.

[12] B. Deverett, R. Faulkner, M. Fortunato, G. Wayne, and J. Z. Leibo. Interval timing in deep reinforcement learning agents.

[13] C. D. Fiorillo, P. N. Tobler, and W. Schultz. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614):1898–1902, 2003.

[14] S. J. Gershman, A. A. Moustafa, and E. A. Ludvig. Time representation in reinforcement learning models of the basal ganglia. *Frontiers in Computational Neuroscience*, 7, 2014. ISSN 1662-5188. doi: 10.3389/fncom.2013.00194. URL `http://journal.frontiersin.org/article/10.3389/fncom.2013.00194/abstract`.

[15] E. E. Hannon, G. Soley, and S. Ullal. Familiarity overrides complexity in rhythm perception: a cross-cultural comparison of american and turkish listeners. *Journal of Experimental Psychology: Human Perception and Performance*, 38(3):543, 2012.

[16] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[17] J. R. Hollerman and W. Schultz. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature neuroscience*, 1(4):304–309, 1998.

[18] J. R. Iversen and R. Balasubramaniam. Synchronization and temporal processing. *Current Opinion in Behavioral Sciences*, 8:175–180, 2016.

[19] M. Jazayeri and M. Shadlen. A Neural Mechanism for Sensing and Reproducing a Time Interval. *Current Biology*, 25(20):2599–2609, Oct. 2015. ISSN 09609822. doi: 10.1016/j.cub.2015.08.038. URL `https://linkinghub.elsevier.com/retrieve/pii/S096098221501009X`.

[20] D. Joel, Y. Niv, and E. Ruppin. Actor–critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*, 15(4-6):535–547, June 2002. ISSN 08936080. doi: 10.1016/S0893-6080(02)00047-3. URL `https://linkinghub.elsevier.com/retrieve/pii/S0893608002000473`.

[21] R. Laje and D. V. Buonomano. Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nature neuroscience*, 16(7):925–933, 2013.

[22] J. Mansuri, H. Aleem, and N. M. Grzywacz. Systematic errors in the perception of rhythm. *Frontiers in Human Neuroscience*, 16:1009219, 2022.

[23] V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013.

[24] N. Y. Masse, G. R. Yang, H. F. Song, X.-J. Wang, and D. J. Freedman. Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature neuroscience*, 22(7):1159–1167, 2019.

[25] T. E. Matthews, M. A. Witek, T. Lund, P. Vuust, and V. B. Penhune. The sensation of groove engages motor and reward networks. *NeuroImage*, 214: 116768, July 2020. ISSN 10538119. doi: 10.1016/j.neuroimage.2020.116768. URL `https://linkinghub.elsevier.com/retrieve/pii/S105381192030255X`.

[26] H. Merchant, J. Grahn, L. Trainor, M. Rohrmeier, and W. T. Fitch. Finding the beat: a neural perspective across humans and non-human primates. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1664):20140093,

Mar. 2015. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2014.0093. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2014.0093.

[27] J. A. Michon and P.-B. Instituut voor Zintuigfysiologie RVO-TNO (Soesterberg. *Timing in temporal tracking*. Institute for Perception RVO-TNO Soesterberg, The Netherlands, 1967.

[28] P. R. Montague, P. Dayan, and T. J. Sejnowski. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of neuroscience*, 16(5):1936–1947, 1996.

[29] K. Müller, G. Aschersleben, R. Koch, H.-J. Freund, and W. Prinz. Action timing in an isochronous tapping task: Evidence from behavioral studies and neuroimaging. In *Advances in psychology*, volume 129, pages 233–250. Elsevier, 1999.

[30] Y. Niv. Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154, June 2009. ISSN 00222496. doi: 10.1016/j.jmp.2008.12.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0022249608001181.

[31] C. Olah. Understanding lstm networks, 2015. URL https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[32] A. D. Patel. Vocal learning as a preadaptation for the evolution of human beat perception and synchronization. *Philosophical Transactions of the Royal Society B*, 376(1835):20200326, 2021.

[33] A. D. Patel, J. R. Iversen, M. R. Bregman, and I. Schulz. Experimental evidence for synchronization to a musical beat in a nonhuman animal. *Current biology*, 19(10):827–830, 2009.

[34] R. Polak, N. Jacoby, T. Fischinger, D. Goldberg, A. Holzapfel, and J. London. Rhythmic prototypes across cultures: A comparative study of tapping synchronization. *Music Perception: An Interdisciplinary Journal*, 36(1):1–23, 2018.

[35] A. Raffin, A. Hill, M. Ernestus, A. Gleave, A. Kanervisto, and N. Dormann. Stable baselines3. `https://github.com/DLR-RM/stable-baselines3`, 2019.

[36] E. D. Remington, D. Narain, E. A. Hosseini, and M. Jazayeri. Flexible sensorimotor computations through rapid reconfiguration of cortical dynamics. *Neuron*, 98(5):1005–1019, 2018.

[37] B. H. Repp. Sensorimotor synchronization: A review of the tapping literature. *Psychonomic Bulletin & Review*, 12(6):969–992, Dec. 2005. ISSN 1069-9384, 1531-5320. doi: 10.3758/BF03206433. URL `http://link.springer.com/10.3758/BF03206433`.

[38] B. H. Repp. Sensorimotor synchronization and perception of timing: effects of music training and task experience. *Human movement science*, 29(2):200–213, 2010.

[39] B. H. Repp and Y.-H. Su. Sensorimotor synchronization: A review of recent research (2006–2012). *Psychonomic Bulletin & Review*, 20(3):403–452, June 2013. ISSN 1069-9384, 1531-5320. doi: 10.3758/s13423-012-0371-2. URL `http://link.springer.com/10.3758/s13423-012-0371-2`.

[40] R. Rescorla and A. Wagner. *A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement*, volume Vol. 2. 01 1972.

[41] M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, and S. Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013.

[42] R. Romo and W. Schultz. Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. *Journal of neurophysiology*, 63(3):592–606, 1990.

[43] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[44] W. Schultz, P. Apicella, and T. Ljungberg. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of neuroscience*, 13(3):900–913, 1993.

[45] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

[46] H. Sohn, D. Narain, N. Meirhaeghe, and M. Jazayeri. Bayesian computation through cortical latent dynamics. *Neuron*, 103(5):934–947, 2019.

[47] R. C. Staudemeyer and E. R. Morris. Understanding lstm–a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*, 2019.

[48] R. S. Sutton and A. G. Barto. Time-derivative models of pavlovian reinforcement. 1990.

[49] R. S. Sutton and A. G. Barto. Reinforcement learning: An introduction. *Robotica*, 17(2):229–235, 1999.

[50] R. Takeya, M. Kameda, A. D. Patel, and M. Tanaka. Predictive and tempo-flexible synchronization to a visual metronome in monkeys. *Scientific reports*, 7 (1):6127, 2017.

[51] K. Tomyta, H. Ohira, and K. Katahira. Asymmetric error correction in the synchronization tapping task. *Timing & Time Perception*, 1(aop):1–10, 2023.

[52] M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. De Cola, T. Deleu, M. Goulão, A. Kallinteris, M. Krimmel, A. KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

[53] S. Ullal-Gupta, E. E. Hannon, and J. S. Snyder. Tapping to a slow tempo in the presence of simple and complex meters reveals experience-specific biases for processing music. *PLoS One*, 9(7):e102962, 2014.

[54] J. Wang, D. Narain, E. A. Hosseini, and M. Jazayeri. Flexible timing by temporal scaling of cortical responses. *Nature neuroscience*, 21(1):102–110, 2018.

[55] C. J. Watkins and P. Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.

[56] M. Zentner and T. Eerola. Rhythmic engagement with music in infancy. *Proceedings of the National Academy of Sciences*, 107(13):5768–5773, 2010.