Simulating Agent Behavior Using Reinforcement Learning

Ofer Drori[1], Nir Sassy[1], Ben Engelhard[2]
[1] Department of Biomedical Engineering, Technion - IIT, Haifa, Israel
[2] Department of Medicine, Engelhard Lab, Technion - IIT, Haifa, Israel

Abstract: This article explores the field of Reinforcement Learning (RL), a subfield of Machine Learning (ML) that concentrates on developing algorithms enabling agents to acquire optimal behavior through interaction with their environment. Both RL and biological learning utilize reward signals as behavior guidelines. RL agents receive feedback from the environment in the form of rewards or punishments, while biological learners experience it through neurotransmitter release, such as dopamine, in response to positive outcomes. The project's primary objective is to extract learning signals from RL agents to make comparisons with biological learners. The methods employed encompass Deep Q-Learning Network (DQN), a value-based RL algorithm that approximates the Q-function using a deep neural network, and Proximal Policy Optimization (PPO), a policy-based 'Actor-Critic' algorithm. The results showcase the learning signals extracted from RL agents during complex tasks. Moreover, a platform has been established to train and evaluate agents using various RL algorithms. This research marks a significant milestone in the broader investigation aimed at understanding biological learning through deep reinforcement learning. The extracted data serves as a valuable baseline for future comparisons with biological learners, and the platform facilitates easy conduction of further RL experiments in the laboratory.

Keywords: Reinforcement Learning, Deep Q-Network, Proximal Policy Optimization, Machine Learning, Biological Learning.

1. Introduction

Reinforcement Learning (RL), a subset of Machine Learning (ML), has garnered substantial attention for its ability to train agents to acquire optimal behaviors by interacting with their environments. This attention stems from its success in tackling intricate tasks, such as game-playing and autonomous vehicle control [1]. The parallels between RL and biological learning processes also offer intriguing insights into the mechanisms underlying human and animal decision-making.

RL has found applications in modeling diverse phenomena in neuroscience and psychology, ranging from basic stimulus-response associations to intricate decision-making employing Markov Decision Processes (MDP) [2]. By utilizing reward signals to steer actions, RL algorithms learn from experience and adapt akin to biological learners. Consequently, RL's potential as a fundamental biological learning mechanism has gained traction [3-5]. Nevertheless, discerning the extent to which RL algorithms mirror the intricacies of biological learning remains an open inquiry. By comparing outputs from deep RL algorithms and biological learning processes, we endeavor to glean deeper insights into the efficacy and limitations of these algorithms, potentially unraveling novel perspectives on the nature of learning and decision-making.

A pivotal parallel between RL and biological learning resides in the use of reward signals to guide behavior. In both realms, agents receive environmental feedback through rewards or penalties, distinguishing favorable from unfavorable actions. In the realm of ML, feedback manifests as reward signals assigned to agent actions, driving the agent to maximize cumulative rewards by learning to discern beneficial from detrimental actions. Similarly, biological learning

leverages neurotransmitter releases, notably dopamine, in response to favorable outcomes [6].

The exploration-exploitation trade off stands as another shared facet between RL and biological learning. In both domains, agents balance the desire to exploit known rewarding actions with the necessity to explore novel actions with potentially greater rewards [7]. This balance gains significance in intricate, uncertain environments where optimal actions aren't readily apparent.

Our study harnesses Machine Learning-trained agents, employing two prominent RL algorithms—Deep Q-Network (DQN) and Proximal Policy Optimization (PPO)—both employing Artificial Neural Networks (ANNs).

ANNs represent a cornerstone in numerous RL algorithms and are thought to mirror aspects of biological learning systems. Inspired by biological neural networks, ANNs are computational models that process input signals, integrate them, and yield output signals using nonlinear activation functions [8].

DQN, a Q-learning algorithm, leverages deep neural networks to approximate optimal value functions in specific environments [9]. By predicting expected rewards for feasible actions within given states and subsequently selecting actions with maximal rewards, DQN engages in iterative learning, updating its action-value function using accumulated rewards [10]. DQN's affinity with biological learning is evident. through its adoption of a memory replay mechanism, archiving past experiences and drawing training samples, akin to the importance of diverse experience accumulation in biological contexts [11]. In contrast, PPO, a policy optimization algorithm, directly refines policies by favoring actions with higher reward probabilities [12]. Through iterative policy enhancements, PPO encourages gradual exploration of slightly varied actions while maintaining continuity with the preceding policy, mirroring trial-and-error learning seen in animals and humans. The agent's ongoing adaptation to environmental feedback in PPO contributes to the refinement of decision-making skills over time [13].

Research avenues for investigating the relationship between ML and biological learning encompass the development of biological learning models influenced by ML algorithms or comparing neural activity during learning tasks to ML algorithm behavior [14, 15]. While potent, these approaches suffer interpretability gaps, impeding comprehensive understanding and comparison [16]. In essence, though RL and dopamine secretion are distinct, they share core traits concerning reward-centric learning. Unveiling these connections, particularly the interplay between ANNs and biological neural systems, promises fresh insights into the learning and decision-making mechanisms in both artificial and biological realms.

2. Methods

Throughout this project, we engaged in extensive research within the domain of Reinforcement Learning (RL) and its corresponding value-based and policy-based algorithms. Our investigation involved a thorough exploration of the foundational theories in Reinforcement Learning, resulting in a profound understanding of its complexities and practical applications.

The initial phase of our endeavor involved establishing a virtual environment for the Preferred Reinforcement Learning library (PFRL). This pivotal stage granted us the essential platform for training RL agents within a controlled and adaptable context. However, accomplishing this task required exploration into the PFRL documentation. Once we had attained a comprehensive grasp of the library's functionalities and capabilities, we proceeded to harness its potential to achieve our experimental objective of extracting the learning signal from RL agents.

Preferred Reinforcement Learning Library (PFRL) [17]:

PFRL is a python based open-source library that is renowned for its high-level implementation of cutting-edge RL algorithms. Notably, among these algorithms are DQN and PPO, both hinging on the employment of ANNs. The selection of PFRL is well-founded as it has garnered widespread adoption within the RL research community, largely attributed to its state-of-the-art benchmark scores achieved through these exemplary implementations. Consequently, the integration of DQN and PPO from the PFRL library presents a robust and established foundation for our investigative pursuits.

Deep Q-Network (DQN) algorithm [18, 19]:

Deep Q-Network (DQN) algorithm is designed to approximate the optimal Q-function, a pivotal metric gauging the anticipated long-term rewards when a specific action is taken in a given state. This Q-function is embodied by a deep neural network, which takes the current state as input and furnishes estimated Q-values for each available action. In essence, the Q-function quantifies the utility of state-action pairings. The algorithm's cornerstone is the Bellman Equation, iteratively updating values through a process that involves a weighted blend of the current value and new information. This update equation, represented in equation (1).

$$(1)\quad Q^{new}_{(s_t,a_t)} = Q(s_t,a_t) + \alpha \cdot \left(r_t + \gamma \cdot \max_a Q(s_{t+1},a) - Q(s_t,a_t) \right)$$

Where $Q^{new}_{(s_t,a_t)}$ is the updated value for a given state-action combination, $Q(s_t,a_t)$ is the current value of a that combination, $\alpha \cdot r_t$ is the reward received from moving from state s_t via action a_t, where α serves as the learning rate. $\alpha \cdot \gamma \cdot \max_a Q(s_{t+1},a)$ represent the projected optimal future value, reflecting the expected cumulative rewards in future states if optimal actions are chosen. γ is the discount factor which discounts the rewards in future states.

The term $r_t + \gamma \cdot \max_a Q(s_{t+1},a) - Q(s_t,a_t)$ embodies the temporal difference, which the algorithm endeavors to minimize during the training phase. This pivotal DQN framework establishes a foundation for approximating optimal Q-values through deep neural networks, capturing the essence of intricate decision-making scenarios.

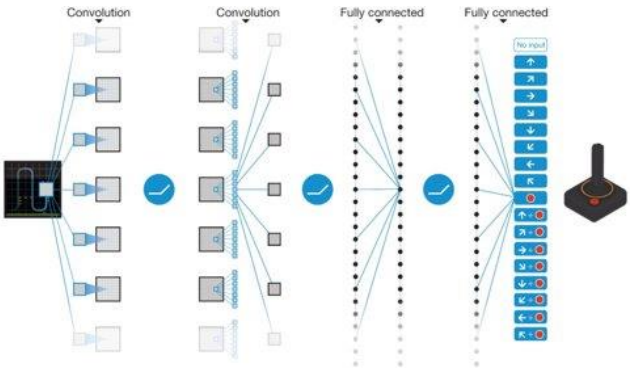

Figure 1 – DQN's ANN architecture.

Proximal Policy Optimization (PPO) algorithm [20, 21]:

The Proximal Policy Optimization (PPO) algorithm addresses reinforcement learning tasks by optimizing a policy, a function that dictates action selection based on the current state. The primary objective of PPO is to enhance the cumulative reward achieved along a trajectory by learning a policy that maximizes this reward expectation. To achieve this, an ANN is employed to represent the policy, transforming states into action probability distributions. The process involves policy updates guided by the minimization of an objective function, quantifying the divergence between the updated policy and its previous version.

The policy is parameterized by θ, signifying the neural network parameters. The optimization task lies in determining the optimal θ, expressed in equation (2).

$$(2)\quad \theta^* = \operatorname*{argmax}_\theta J(\theta)$$

Where $J(\theta)$ is the expected return of the policy $\pi(s|a;\theta)$ starting from the initial state s_0. This term outlines the cumulative reward across the trajectory $s_0 \rightarrow s$.

The maximization problem can be approached using various methodologies, with a prevalent approach being gradient-based iterative techniques. These techniques leverage a fundamental theorem, known as The Policy Gradient Theorem, which establishes a connection between the gradients of the expected return $\nabla J(\theta)$, and the gradients of the policy, $\nabla\pi(a|s;\theta)$. In accordance with this theorem, the relationship is articulated in equation (3):

$$(3)\ \nabla J(\theta) = \sum_{s\in S} \mu(s) \sum_{a\in A} Q^{\pi}(s,a) \cdot \nabla_{\theta}\pi(a|s;\ \theta)$$

Where $\mu(s)$ is the probability of a state given an initial state and current weights θ, and $Q^{\pi}(s,a)$ is the value of a state-action combination which is derived straight from the policy. The theorem fundamentally establishes a gradient-based connection between policy adjustments and expected returns, serving as a foundational principle in the PPO framework.

The actor-critic methodology emerges as a significant enhancement to the PPO algorithm. Within this method, the architecture splits into two core components: the actor and the critic. The actor guides action selection, driven by the policy, while the critic estimates the value of various states. This mutual collaboration enhances the stability of the learning process and culminates in more informed decision-making. The critic's value estimates offer a foundation for evaluating the quality of chosen actions, thereby contributing to the actor's policy adjustments. In harmony, the actor-critic methodology leverages the strengths of both policy-based and value-based methods, augmenting the PPO algorithm's capability to navigate intricate learning landscapes.

3. Results and discussion

Our first primary objective was to establish a functional framework for conducting experimental endeavors utilizing Reinforcement Learning (RL) algorithms within a virtual environment. Following an extensive research phase, the decision was made to adopt OpenAI's PFRL open-source library. This library offers a high-level implementation of a diverse range of RL algorithms.

Having successfully set up the PFRL library alongside its requisite dependencies, our focus shifted to the training of RL agents employing the provided PFRL implementations. These agents were subsequently subjected to performance evaluation through environment rendering (Figure 3). Upon observing satisfactory outcomes, particularly in terms of qualitative assessment through human observation of gameplay, our subsequent objective entailed enabling the extraction of training data from the agents during the training process, thereby establishing a meaningful learning signal.

These acquired learning signals will be the subject of future comparative analysis with the dopamine secretion patterns exhibited by learning mice. Given the inherent methodological diversity among the employed algorithms, it is anticipated that the resulting learning signals will exhibit variations dependent on the specific algorithmic approach and hyperparameters employed in each experiment.

To address this inherent variability, modifications were made to the library to ensure that comprehensive statistical data from the training phase could be stored, which allows for flexibility in terms of signal excretion. Figure 2 visually represents the learning signals derived from the training of two distinct agents: DQN (A, B) trained over 50 million episodes, and PPO (C, D) trained over 10 million episodes. The employed environment is from Atari's collection, with the example being "Breakout" (as depicted in Figure 3).

The illustrated learning signals encompass two facets: the average loss and the average state-action value, both plotted against the progression of episodes. The initial rise in average loss can be attributed to the exploration-exploitation trade-off inherent in RL, whereas the steady increase in the average state-action value underscores the agent's ongoing improvement even when opting for suboptimal actions [22].

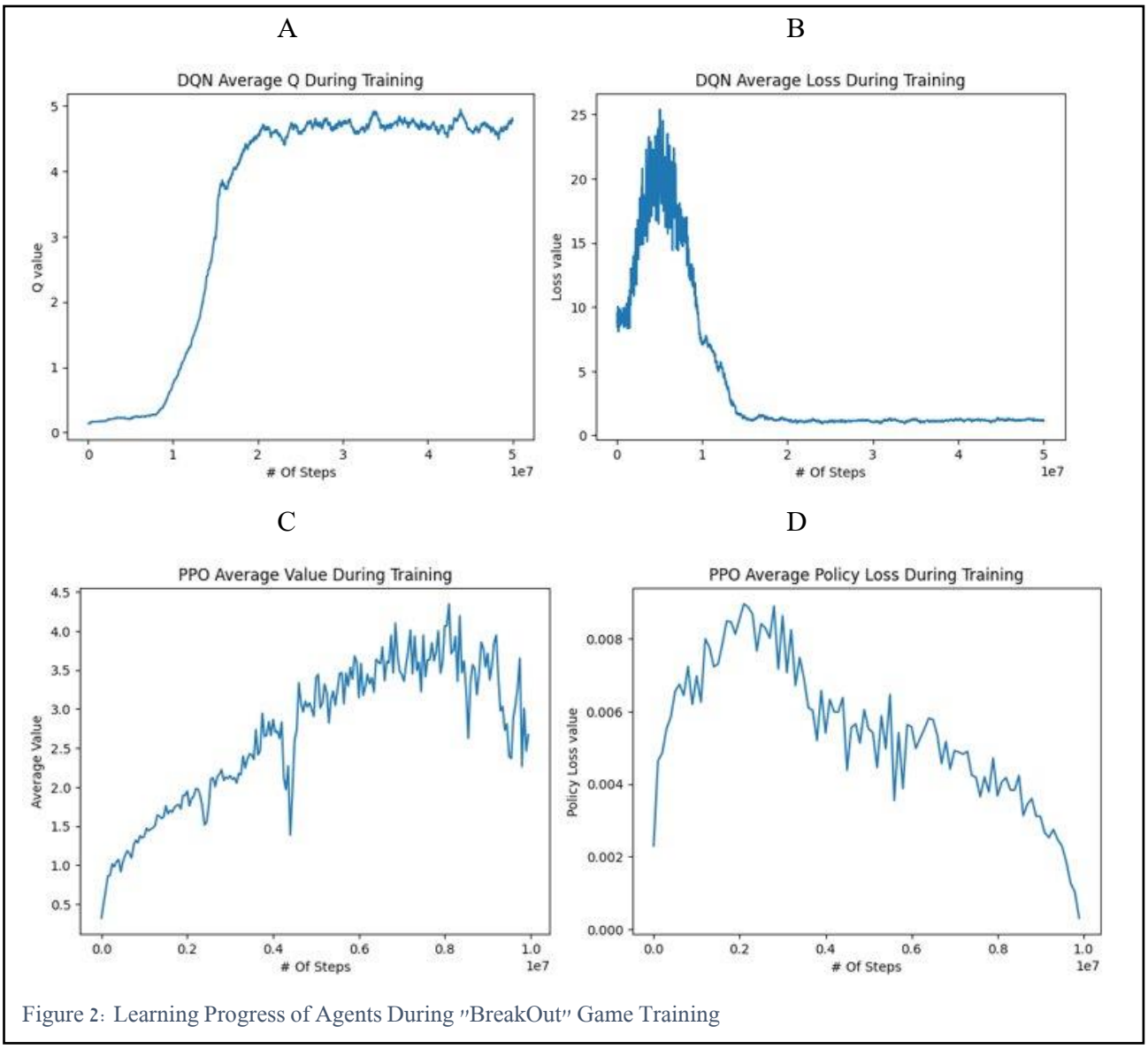

Figure 2: Learning Progress of Agents During "BreakOut" Game Training

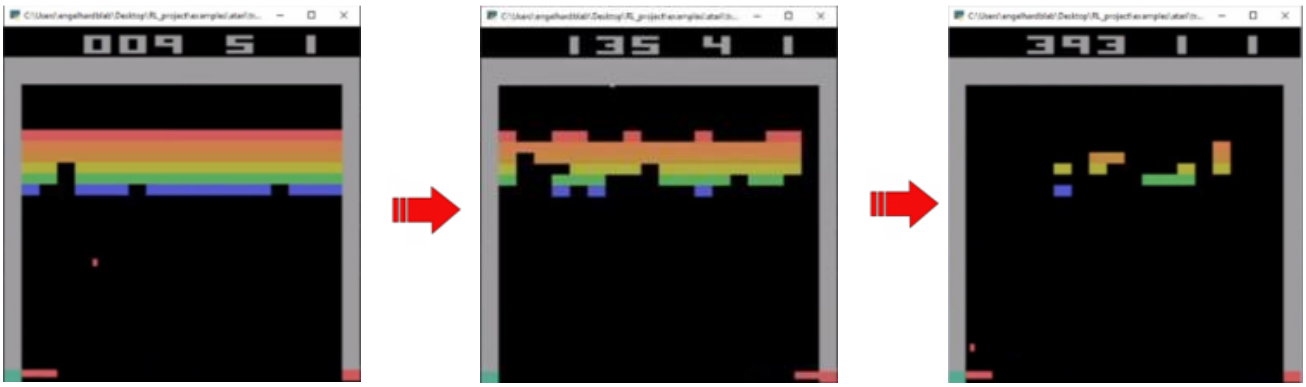

Figure 3 – Atari's 'BreakOut" game played by a trained PPO agent.

4. Conclusions

In this project, we embarked on a comprehensive exploration of RL, a vital subset of Machine Learning, that empowers agents to attain optimal behavior by interacting with their environment. Drawing parallels between RL and biological learning, both of which hinge on reward signals to guide behavior, we delved into the mechanisms underlying these learning processes. Leveraging the power of DQN and PPO algorithms, each driven by ANN, we unveiled profound insights into their capacity to approximate optimal value functions and optimize policies, respectively.

By harnessing PFRL library, we established a robust platform for RL experimentation. Through a meticulous process, we trained and evaluated agents using DQN and PPO algorithms. The extraction of learning signals from RL agents presents a promising baseline for future comparisons with biological learners, fostering a novel perspective on the nature of learning and decision-making. These learning signals, captured in Figure 2, exhibited intriguing patterns, shedding light on the intricate dance between exploration and exploitation and the gradual refinement of state-action values.

In conclusion, this research advances our comprehension of the synergies between RL and biological learning, unveiling the potential of deep reinforcement learning to emulate and enhance learning processes. Our study paves the way for further investigations into the intersection of artificial and biological learning, opening doors to new horizons of understanding in both domains. As we continue to bridge the gap between machine and biological learning, our work contributes to the broader scientific endeavor of deciphering the fundamental mechanisms underlying intelligent decision-making.

5. Acknowledgment

Rotem Shapira

Guy Sassy (A.I Researcher – ECE, Technion)

Uriel Nusbaum (Algorithms Developer – Elbit)

Behrooz Omidvar-Tehrani (Contributor for PFRL library)

6. References

[1] "Reinforcement Learning 101. Learn the essentials of Reinforcement… | by Shweta Bhatt | Towards Data Science." https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292

[2] X. Xu, L. Zuo, and Z. Huang, "Reinforcement learning algorithms with function approximation: Recent advances and applications," Inf Sci (N Y), vol. 261, pp. 1–31, Mar. 2014, doi: 10.1016/J.INS.2013.08.037.

[3] C. B. Delahunt, J. A. Riffell, and J. N. Kutz, "Biological mechanisms for learning: A computational model of olfactory learning in the manduca sexta moth, with applications to neural nets," Front Comput Neurosci, vol. 12, p. 102, Dec. 2018, doi: 10.3389/FNCOM.2018.00102/XML/NLM.

[4] R. S. Lee, B. Engelhard, I. B. Witten, and N. D. Daw, "A vector reward prediction error model explains dopaminergic heterogeneity," bioRxiv, p. 2022.02.28.482379, Mar. 2022, doi: 10.1101/2022.02.28.482379.

[5] K. Doya, "Reinforcement learning: Computational theory and biological mechanisms," HFSP J, vol. 1, no. 1, p. 30, 2007, doi: 10.2976/1.2732246.

[6] R. A. Wise, "Dopamine, learning and motivation," Nature Reviews Neuroscience 2004 5:6, vol. 5, no. 6, pp. 483–494, 2004, doi: 10.1038/nrn1406.

[7] O. Berger-Tal, J. Nathan, E. Meron, and D. Saltz, "The Exploration-Exploitation Dilemma: A Multidisciplinary Framework," PLoS One, vol. 9, no. 4, p. 95693, 2014, doi: 10.1371/journal.pone.0095693.t001.

[8] E. Journal, O. S. Eluyode, and D. T. Akomolafe, "Scholars Research Library Comparative study of biological and artificial neural networks," of Applied Engineering and Scientific Research, vol. 2, no. 1, pp. 36–46, 2013.

[9] V. Mnih et al., "Human-level control through deep reinforcement learning," Nature 2015 518:7540, vol. 518, no. 7540, pp. 529–533, Feb. 2015, doi: 10.1038/nature14236.

[10] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction Second edition, in progress".

[11] E. L. Roscow, R. Chua, R. P. Costa, M. W. Jones, and N. Lepora, "Learning offline: memory replay in biological and artificial reinforcement learning," Trends Neurosci, vol. 44, no. 10, pp. 808–821, Oct. 2021, doi: 10.1016/J.TINS.2021.07.007.

[12] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. K. Openai, "Proximal Policy Optimization Algorithms".

[13] A. Xavier Fidêncio, C. Klaes, and I. Iossifidis, "Error-Related Potentials in Reinforcement Learning-Based Brain-Machine Interfaces," Front Hum Neurosci, vol. 16, p. 806517, Jun. 2022, doi: 10.3389/FNHUM.2022.806517/FULL.

[14] T. Song, P. Zheng, M. L. D. Wong, and X. Wang, "BIO-INSPIRED COMPUTING MODELS AND ALGORITHMS," Bio-Inspired Computing Models and Algorithms, pp. 1–282, Jan. 2019, doi: 10.1142/10119/SUPPL_FILE/10119_PREFACE.PDF.

[15] J. J. Valletta, C. Torney, M. Kings, A. Thornton, and J. Madden, "Applications of machine learning in animal behaviour studies," Anim Behav, vol. 124, pp. 203–220, Feb. 2017, doi: 10.1016/J.ANBEHAV.2016.12.005.

[16] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," Brief Bioinform, vol. 18, no. 5, pp. 851–869, Sep. 2017, doi: 10.1093/BIB/BBW068.

[17] "pfnet/pfrl: PFRL: a PyTorch-based deep reinforcement learning library." https://github.com/pfnet/pfrl.

[18] "Deep Q-Learning Tutorial: minDQN. A Practical Guide to Deep Q-Networks | by Mike Wang | Towards Data Science." https://towardsdatascience.com/deep-q-learning-tutorial-mindqn-2a4c855abffc.

[19] Y. Li, "DEEP REINFORCEMENT LEARNING: AN OVERVIEW,". [Online]. Available: https://arxiv.org/abs/

[20] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. K. Openai, "Proximal Policy Optimization Algorithms," Jul. 2017.

[21] “Proximal Policy Optimization.” https://openai.com/research/openai-baselines-ppo.

[22] “The exploration-exploitation trade-off: intuitions and strategies | by Joseph Rocca | Towards Data Science.” https://towardsdatascience.com/the-exploration-exploitation-dilemma-f5622fbe1e82