Bootstrapping Human-Autonomy Collaborations by using Brain-Computer Interface of SSVEP for Multi-Agent Deep Reinforcement Learning

Joshua Ho[1,2,3], Chien-Min Wang[2], Chun-Hsiang Chuang[3,4], Chung-Ta King[3, 5], Chi-Wei Feng[5], Tun-Hsiang Chou[6], Yen-Min Chen[6], Yu-Hsin Yang[6], and Yi-Cheng Hsiao[5]

[1]Social Networks and Human-Centered Computing Program, Taiwan International Graduate Program
[2]Institute of Information Science, Academia Sinica, Taipei, Taiwan 115
[3]Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu, Taiwan, 30013
[4]College of Education, National Tsing Hua University, Hsinchu, Taiwan, 30013
[5]Department of Computer Science, National Tsing Hua University, Hsinchu Taiwan, 30013
[6]College of EECS, National Taiwan University, Taipei Taiwan, 106
{jho, cmwang}@iis.sinica.edu.tw, ch.chuang@mx.nthu.edu.tw, king@cs.nthu.edu.tw, fenganthony001@gmail.com, {r11922163, b08902132, B09901200}@ntu.edu.tw, s110064508@m110.nthu.edu.tw

***Abstract*— Human-Autonomy Teaming (HAT) has become one of the emerging AI trends due to the advances in sophisticated machine design that allows closer cooperation with humans, while performing moral, reasonable, and applicable tasks as humans' most exemplary assistants. Based on pursuing the collective goal and sharing the authority between humans and machines, adding brain-computer interfaces (BCI) to HAT principles is considered an intuitive and promising approach, enabling HAT to achieve the optimal decision-support systems. This study proposes a BCI-based system with a Reinforcement Learning (RL) algorithm as a *'human-in-the-loop'* teaming integration. The neural responses elicited by the Steady-State Visual Evoked Potential in BCI facilitate the collaboration of learning agents with humans and accomplish this goal in a game simulation environment. The results of our proposed system, NeuroRL, show significant improvement by reducing the non-stationarity of exploitations and explorations. The rewards are optimized during the early investigations to more efficiently achieve the convergence. The novel design proposed in this study can extend the development of the emerging HAT field and *knowledge-based* RL systems for various applications in dynamic and autonomous environments.**

Keywords—human-AI system, reinforcement learning, brain-computer interface, human-in-the-loop

I. Introduction

The method of Reinforcement Learning (RL) has focused on applying Deep Neural Networks to leverage statistical learning approximations and to address the Deep Reinforcement Learning (DRL) methodologies in order to resolve more complicated AI tasks. That is why research and applications based on DRL approaches have become more popular and mature for difficult tasks [1], like robotic control and RL agents for steering vehicles. Intelligent agents can also assist humans in our daily lives to not only help offload the burdens of routine tasks but also reduce the risk of humans being exposed to dangerous conditions. To this end, multi-agent DRL (MADRL) systems have been shown to have comprehensive capabilities to tackle complicated problems [2]. However, the MADRL system may require more efficient communication channels to construct teaming strategies that can validate group membership. The eventual objective for MADRL systems is for DRL agents precisely and efficiently reaching the optimal state by learning better collaborations to achieve the collective goals in dynamic environments.

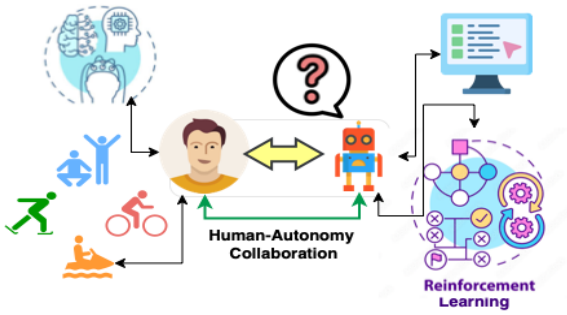

Fig. 1. The proposed HAT system for the BCI-based methodology facilitating RL agents to efficiently achieve accumulated rewards.

Recent developments in Brain-Computer Interfaces (BCIs) have made it easier to transmit human electroencephalogram (EEG) data to computers, which can certainly help physically impaired people, and possibly link our mind and control over remote machines in a more intuitive way. Many companies like Neuralink [24] have devised BCI technologies to decode human neural signals through embedding a device under a person's scalp. On the other hand, other companies like Snap [25] and Facebook (or Meta) [26] focus on non-invasive BCIs by recording and analyzing EEG data through various sensors attached to a person's head. These non-invasive approaches are undoubtedly safer, more natural, and easier to use because they are exoskeletal or helmets with a single or several built-in sensors that measure brain activity for the purpose of brain-machine interaction.

However, human-computer interaction systems often face conflicts and dilemmas that may pose great challenges if humans interact with computer agents less aware of human-centered considerations, like human preferences, behaviors, and intentions. It is essential that there is mutual tacit agreement and trust between humans and machines in an interactive and interpretive human-machine system. For instance, machines may need to learn guiding instructions in order to share authority with humans, which can help prevent potential risks in uncertain conditions. Machines and robots must also be programmed with a moral responsibility to predict safety and learning efficiency for reaching the ultimate

goal of scalability and responsibility. Meanwhile, the instructions from humans or human brains should immediately and precisely guide the robots if the system performance can achieve our requirements, especially for the emerging Human-Autonomy Teaming (HAT) systems and applications.

Our proposed HAT system, shown in Fig. 1, aims to apply human brain signals to enable a MADRL system proposed by HAT principles. The BCI-based system uses human neural responses and interventions to help DRL agents learn the Markov Decision Process (MDP), which will enable the agents to more efficiently achieve the highest cumulative rewards and convergent goals. Based on the neural elicitations of Steady-State Visual Evoked Potential (SSVEP) in BCI [10], the teaming integration has been found to significantly improve the explicit objectives with rapid '*human-in-the-loop'* perceptions in game simulations [4]. The bias-understanding DRL agents can ultimately promote trustworthy AI of accountability [15, 20], efficiency, and transparency while simultaneously building comfortable and manageable mechanisms for humans. For this reason, the integrated framework of *knowledge-based* RL and SSVEP offers a more sophisticated method that bootstraps human-agent collaborations. We anticipate that the proposed human-machine collaborations can thoroughly establish a sophisticated and dependable learning model based on the dynamic accessibility of RL and BCI. This will finally lead to more valuable inputs and greater Human-AI robustness to help agents more harmoniously engage in learning from brain-machine collaborations against adversaries in dynamic game-controlled environments.

The remainder of this work is organized as follows: Section II reviews the related works as the preliminary section. Section III presents the proposed system architecture in our design and Sections IV reports the experiment results. Section V concludes this paper and addresses our plans for future research.

II. Preliminary

A. Multi-Agents DRL

Learning in an RL model in high-dimensional environments is regarded as a challenging task. Recent works have focused more on off-policy and statistical approaches, like DRL based on Deep Neural Networks (DNN) in [1] and Deep Q-Network (DQN) in [2]. They can simplify the parametric estimations and help overcome policy control to improve the latency issues between actions and corresponding rewards. Also, DRL requires high quality environments, better observations, and explicit information on action-space pairs, states, and rewards to manipulate the learning process. In MADRL, the system requires more observations of the RL interactions of multiple agents in the same environment due to dual conditions of cooperation and competition [3]. The work [14] showed an overview of DRL and human guidance in various forms. The action advice [4] and shared knowledge transfer [5, 17] in the MADRL system also play significant roles by advising DRL agents to improve collaborations with humans while achieving accuracy and learning efficiency [15]. These proposed algorithms resolve and incorporate human guidance in the agents' decision-making process. These algorithms imitate the '*human-in-the-loop'* [16] demonstrations to perform future actions and learn an optimal policy to accelerate RL in dynamic environments.

B. Brain-Computer Interface

Over the past few decades, laboratory research has investigated brain communication channels since the first discovery in [22]. They have developed various methods to enhance human peripheral nerves for practical applications in brain-computer interfaces (BCIs) and technologies [11]. BCI aims to access good-quality human-computer interaction data according to several sub-processes, which are often classified as *signal acquisition*, *data pre-processing*, *feature extraction*, and *classification*. In *classification*, the previous works on BCI studied emerging paradigms by utilizing the machine learning approaches based on DNN [6] to achieve a very high system performance for designing SSVEP-based BCIs and their implementations, as shown in [9]. A comparative study [7] performed a canonical correlation analysis (CCA) to improve SSVEP-based BCI detections; however, these were done in the early assessments without calibration and accuracy evaluations for building the classification model. In addition, the multi-target classification in [10] and the user-dependent or independent study in [8] have demonstrated that SSVEP-assisted BCI can be a helpful and practical method for many real-world applications.

C. Human-Autonomy Teaming

HAT has received attention in recent years especially for research works focusing on SSVEP-based BCI, eye-tracking, and force feedback responses [12] to manipulate robot actions in hybrid BCI systems [19]. Related work on human and MADRL interactions [13] presents an explainable and adaptable interface for achieving sampling efficiency to shorten the RL training process. However, even though these works are instrumental in resolving the challenging problems in DRL or BCI, to the best of our knowledge from our literature survey, no work within the DRL for BCI paradigm has considered augmenting AI with thoughtful human neural responses. This promising approach could analyze and manage more dynamic RL models, increase degree of responsive methods and BCIs with more horizons, and also add an essentially *neural-enabled* HAT for an RL agent. In this work, we propose a novel HAT framework, *NeuroRL*, to design a *knowledge-based* RL integrated with an SSVEP-enabled BCI system. We anticipate that *NeuroRL* will bring us one step closer to building a genuinely neural-responsive RL system for '*human-in-the-loop'* integration. It will benefit modern AI development by teaming and augmenting the Human-AI synergy in fast and dynamic learning environments.

III. System Architecture

The proposed system integrates DRL with BCI collected from the experiment subjects for a MADRL environment where the multiple agents work toward the collective goal. SSVEP-based BCI is accurate in elaborating signal-to-noise ratios (SNRs) and achieves robustness for artifacts, which can be very useful for assisting paralyzed people and beneficial for medical and research applications based on the immediate stimuli and *sub-second* response times. The goal of the system architecture for *NeuroRL* (Fig. 2) is to construct an efficient HAT system that results in optimal rewards and prevents non-stationarity. Compared to other human-controlled methods, *NeuroRL* can effectively achieve human neural compliance with *neural-enabled* HAT integration.

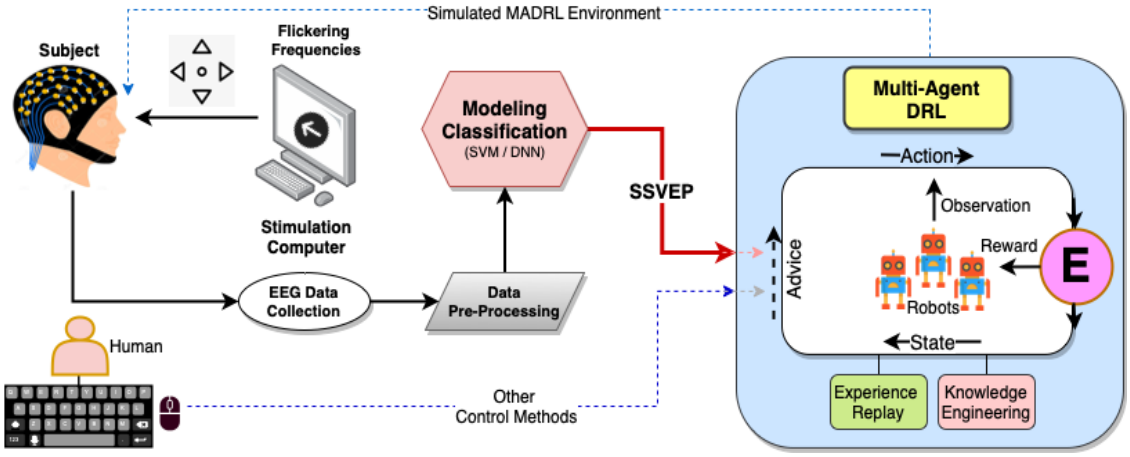

Fig. 2. The overview of system architecture for the proposed *NeuroRL* shows 1) HAT collaborations in the MADRL environment shown as E in the diagram, which aims to improve learning efficiency and prevent non-stationarity for *knowledge-based* RL, and 2) the rapid SSVEP-based BCI integration.

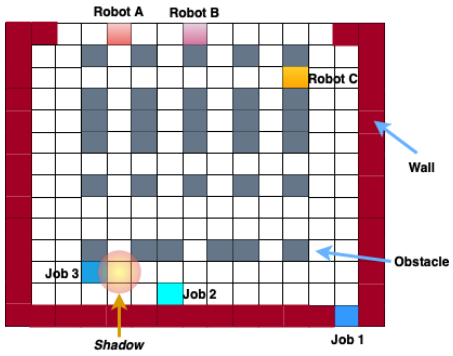

Fig. 3. The *factory-robots* game sets a number of robots with each policy as π guided by *shadow*, and the corresponding jobs in a N x N grid world constrained by the obstacles and walls.

A. Simulated MADRL Environment

Some related works have previously simulated multiple agents in game environments, such as the *hide-and-seek* experiments described in [18]. The simulated game *factory-robots* (Fig. 3) sets several robot agents in an N *x* N grid (cell) of factory as our MADRL environment. The DRL performance is evaluated after the robots finish the jobs. We measured the time spent and the steps counted toward the successful task within a limited number of steps in each episode. Successful task happens in an episode when the each robot has completed its own job. For instance, if robot A, B and C all load to the correct factory destinations, we regard the task as done. Alternatively, if only one or two robots complete the jobs, but the time is up or the third robot is crashed, the game will be reset to another new episode until the total episodes are completed. Any cell should not have two or more agents simultaneously unless the robots may crash into the wall cells and broken within the episode. The wall boundaries limited the robots in the world, and robots should learn how to work together to load the luggage guided by timely inputs of *shadow*. Each episode has a limited 20,000 steps, and if no successful task was done, the current episode should stop to restart a new one till the game set at the 100th episode.

B. DRL, Experience Replay and Knowledge Engineering

The proposed MADRL has multiple cooperative n robot agents denoted by $A = \{0, 1, 2 \ldots . n-1\}$. At time t, each robot agent $a \in A$ observes the state $S_t \in S$ currently in the environment. The taken action $u_t^a \in U$ is stochastically performed based on the policy π'. The final received reward is expressed as $R_t = \{S_t, u_t^a\}$ while the environmental state moves to $S_{t+1} \leftarrow S_t$. Each robot agent's objective is to search a set of policy π for itself to maximize the total overall reward derived as a Q-learning, as per Eq. (1). The probability function shows $R_T = \sum_{j=T}^{\infty} \gamma^{j-T} R_j$, where R_j is the reward gained at time j. The *discount factor* γ (0.95) describes the observability level in the environment. There are two types of method as policy set π: (a) the joint policy for all robots at every π'^j, and (b) a unique policy π'^i for each cop. The general method π'^j considers the collective actions $\prod_{a \in A} U^a$ of all agents working together, whereas the space complexity increases exponentially if the number of agents keeps growing. Our proposed approach, the '*unique policy*' π'^i, is based on each robot finally reducing the complexity of space. Nevertheless, the '*unique policy*' can suffer from the non-stationarity of RL. Therefore, the *Experience Replay* component (Fig. 2) is used for our DQN, which helps to reduce the sample inefficiency and to achieve stationarity, for prioritizes the significantly weighted sample from the memory tree in each iterative update [21].

$$Q_t(s_t, a_t) \leftarrow (1-\alpha) Q_t(s_t, a_t) + \alpha \left[r_t + \gamma \max_{a} Q(s_{t+1}, a) \right], \quad (1)$$

where s_t shows the current state, a_t as the action taken at the current time, r_t is the reward received after performing a_t. $Q_t(s_t, a_t)$ evaluates the action a_t of state s_t. The *learning rate* α and *discount factor* γ are needed in Deep Q-Learning (DQL).

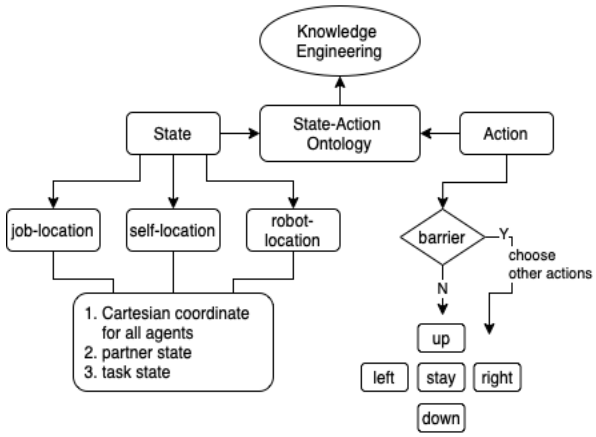

Fig. 4. *Knowledge engineering* and *State-Action* ontology for MADRL to search the state locations based on the performed action.

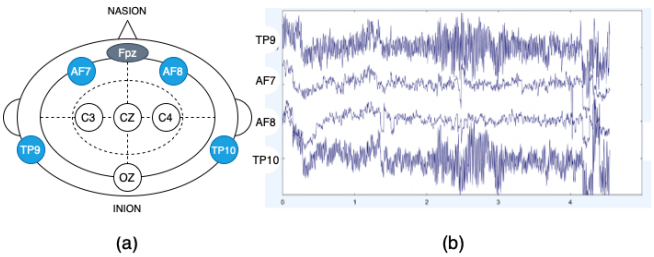

Fig. 5. (a) The four channels of the Muse S band are shown in *blue*; (b) the signals amplitudes in the time domain.

The proposed *NeuroRL* has also focused on the design of *knowledge engineering* (KE) for *State-Action* ontology (displayed in Fig. 4), where the robots can use the *state-action hierarchy-mapping* to choose: (a) Action: the following action like '*up'*, '*down'*, '*left'*, '*right'* or '*stay'*, in order to achieve the targeted task. If the agent encounters an obstacle next to its state, the next action shall be limited by the barriers according to the corresponding coordinates; (b) State: robot can observe its status as *self-location*, and other colleagues' states as *partner-state,* to avoid conflict when they move to the same cell. KE is built upon the knowledge ontology, which offers essentially *knowledge-based* information for *NeuroRL* to achieve its learning objectives more efficiently. Thus, according to experience replay, *NeuroRL* builds KE for the agents to share transferable experiences based on the above *hierarchy-mapping* and weighted memory. These two components significantly contribute to the proposed MADRL, which can explicitly facilitate robot agents quickly grasp information in the dynamic environment, according to episodic memory replays and the *knowledge-based* RL mechanisms within the learning and modeling process.

C. BCI and SSVEP

The proposed system focuses on non-invasive BCI for collecting EEG signals. Based on recorded data, we conducted machine learning (ML) methodologies to build the inference model for predicting the guiding actions advised by the participating subjects. We aimed to simulate the neural control mechanisms according to each participant, who offers '*human-in-the-loop*' instructions to guide the robot agents in the MADRL environment. In the *factory-robots* game setting, the robot agents are required to learn the offline RL strategies to master its job. Each robot agent can stochastically take the actions in the grid world. However, the timely advice for the robots' actions is based on the relative position between the *shadow* position as a '*hint'* and each robot agent's position. Once the *shadow* '*hint'* points to near the *targeted job* cell and the subject can instantly broadcast the advice, the robot agents will be promptly attracted to alter the actions '*advised'* by the human's signals. The subject's controls move the *shadow* to the destination by announcing the '*hint'*. The *shadow* is designed to offer the '*human-in-the-loop'* signals for the robots in MADRL environments (Fig. 5).

Alternatively, *NeuroRL* can simulate the subject's EEG signals assigned to control the *shadow* based on the four directions. So, four flickering frequencies (9.25, 15, 20, and 35 Hz, as described in [7]) are required to build the selective attention and inference model for the directed actions. In the simulated MADRL experiments, the subject stares at each of the four flickering frequencies, which will help build the inference model for moving the *shadow* to the destination of the target's location. Based on the '*hint'* from *shadow*, which is an instant broadcasting signal triggered by the subject's pointing at or near the destination cell of a target job, the robot agent's attention is focused to observe the advice and revise the taken actions, which are continuously guided in the game by the '*human-in-the-loop'* instructions to achieve the optimal reward.

Since *NeuroRL* uses a non-invasive and SSVEP-based BCI method, we chose a Muse S headband device (Fig. 5(a)), for our experiments on *signal acquisition*. The device uses several dry electrodes with a sampling rate of 256Hz for all four channels. There are two *frontal* electrodes, AF7 and AF8, at the forehead and another two *temporal* electrodes, TP9 and TP10, behind the ears to measure the brain activities with a reference channel of Fpz. The device records the streaming EEG data from all channels (Fig. 5(b)) while each targeted flickering frequency triggers the SSVEP-based BCIs for all channels. During the early assessments, we used canonical correlation analysis (CCA) [7], which seeks a pair of linear transformations to maximize the correlations, thus allowing the measuring of the quality of data collected by *signal acquisition* and phase decoding. We also evaluated the data artifacts of the *data pre-processing* by an Independent Component Analysis (ICA) and the noise reduction from Artifact Subspace Reconstruction (ASR) in MATLAB. The range of bandpass filter frequency was set between 5Hz to 48Hz to clean the collected data by removing some noise. In *feature extraction*, an algorithm based on *minimum redundancy - maximum relevance* (MRMR) [23] was applied to calculate the feature ranking scores (Eq. (2)) from the collected data. The classification model was thus trained based on the targeted frequencies from the participating subjects.

$$score_i(f) = \frac{relevence(f \mid target)}{redundancy\ (f \mid fselection)}, \qquad (2)$$

where shows the higher score is obtained of selected feature i per the given target with more relevance and less redundancy; here *fselection* means '*feature selected until i-1'*.

IV. Experimental Results

After the design and implementation of *NeuroRL* was completed, we conducted the experiments in simulating the BCI and DRL that were guided by the '*human-in-the-loop'* instructions by the evaluated machine learning models. We compared the input methods of the control interface to understand how to design the proposed *NeuroRL* that can benefit the learning process in a MADRL environment.

A. BCI Model Evaluation

In the SSVEP-based BCI, we collected 10 trials per targeted frequency and 40 trials in total distributed in 4 different frequencies for each participating subject. Each trial had a length of 80-s streaming data recorded by the Muse S headband device, which was remodeled by a Discrete Fourier Transform (DFT) to construct a transformed vector with 257 points. Based on various degrees of window size to identify the possibly accumulated data block for building an accurate inference model, we evaluated models by *Classification Learner* based on collected EEG data (shown in TABLE I).

TABLE I. The Models of Classification Learner

Evaluated ML Models for SSVEP-based BCI				
SVM	Naïve Bayes	Decision Trees	Nearest Neighbors	Logistic Regression
Ensemble	Neural Networks	Discriminant Analysis	Kernel Approximation	

If we consider the most important 100 features for *user-dependent* modeling, the built models were compared to show that the cubic SVM model can achieve very high accuracy on average. For the window size of *2-sec* window, the best model performance was able to reach 89% accuracy (Fig. 6). Since the high performance of *classification* is evaluated, the elicited SSVEP-based brain signals were simulated to offer the precise '*hint'* by moving the *shadow* guided from humans as the '*human-in-the-loop'* timely input if the subject was identified.

We also evaluated the *user-independent* models of *2-sec* window size (TABLE II). For the *user-independent* models and their performance, the best model, *Ensemble – Bagged Trees*, achieved up to 82.4% accuracy, with 1028 features considered in the construction of all the benchmarked models. Therefore, according to the highly accurate *classification* model as our assumption, the RL agents, or robot agents, are able to collaboratively and efficiently pursue the targets in the proposed MADRL environment according to the effective personal guiding courses by *shadow* of brain signal.

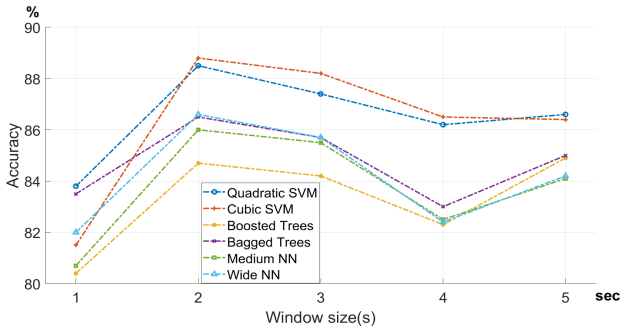

Fig. 6. The performance of user-dependent models.

TABLE II. THE PERFORMANCE OF USER-INDEPENDENT MODELS ON PROPOSED BCIS

Performance of Machine Learning Models				
Ensemble Bagged Trees	Wide Neural Network	SVM Quadratic	SVM Cubic	Ensemble Boosted Trees
82.4%	80.0%	80.1%	62.3%	70.0%

TABLE III. CONTROL INTERFACE FOR HUMAN INPUT METHODS

Keyboard	***SSVEP***
• Press four arrow keys to send the guiding signals of *shadow*. • *Hands-free* may not be achieved; inconvenient for paralyzed users.	• Simulate BCI to send guiding signals of *shadow*. • Neurons control and reply to the robot agents.

B. Control Interface

From the designed *'human-in-the-loop'* instructions, the RL agents receive human guidance to learn the intuitive experiences along with the replay memory through the user's control interfaces in the proposed *NeuroRL*. We observe that the subject can provide the agents with valuable inputs from example methods (TABLE III). One of the ways to deliver the controlling messages of directional arrows is from a keyboard that triggers commands. On the other hand, the simulated SSVEP-based BCI signals can offer an alternative method that can be more convenient for people, especially for those who are restricted by critical conditions or need the *hands-free* requirements. Another strong advantage of using BCI is the capacity to trigger the instructions more promptly. In many scientific reports, neural excitations often occur within a *sub-second* response time. Although we currently achieve a model of *2-sec* window-size, this will still neither limit nor constrain BCI research and future developments to make *NeuroRL* become an essential HAT method.

C. NeuroRL and HAT

The proposed *NeuroRL* aims at giving valuable guidance from SSVEP to the robots, and the robots can avoid non-stationarity during the RL training. Therefore, we consider taking the strategy to timely offer the '*human-in-the-loop'* instructions in each episode when the robots need immediate assistance. This approach allows robot agents to stochastically explore and exploit the environment at the beginning and receive advice when they are close to the targets. In the experiments, we observed that the '*passive'* '*human-in-the-loop*' could adaptively sustain the original RL reward process and to prevent catastrophic failure in the local conditions. Thus, the robots can effectively comprehend the optimal policy sets. The timely advice aims to assist the robots in achieving the highest rewards in the MADRL environment.

The simulated MADRL environment of a walled factory contains obstacles, so the robots were working collaboratively to find the optimal pathways to efficiently load the luggage. In our design, the SSVEP can be triggered to attract the robots that learned to complete the optimal task in their collaboration. Each game has 100 episodes at most to allow the robot agents learn in the MADRL environment. And each episode can only allow limited steps accumulated by 3 robots. If any single job was completed, the job completion rewarded the team with 50 points. Else, a negative reward of -50 points was received if the agent failed to complete the job. Therefore, the maximum reward of the three collaborative robots are with 150 points, and the minimum reward, -150 points, was given to the team if all failure. We experimented the baseline compared to the model using SSVEP-based BCI for inference, which adopts the BCI model's accuracy (about 80%) to generate the 80% probability to direct correctly, and the simulated latency of *2-sec* according to the window size of *2-sec* as the SSVEP-sim. Meanwhile, SSVEP-opt means that the optimal BCI model accuracy (>=95%) with *sub-second* response time was experimented, which might be achieved someday in the future. The report shows that the performance of SSVEP-based approach proposed by *NeuroRL* can achieve better results to efficiently converge to the optimal state (TABLE IV).

TABLE IV. THE COMPARATIVE PERFORMANCE OF SSVEP SIMULATIONS FOR THE PROPOSED NEURORL

Performance Results				
	AVG Reward	**AVG Step**	**AVG Collision**	**Total Time Spent**
Baseline	35.0	3,062.63	28.46	58m
SSVEP-sim	**48.0**	**2,705.53**	**22.85**	**50m**
SSVEP-opt	**71.0**	**2,768.70**	**26.25**	**49m**

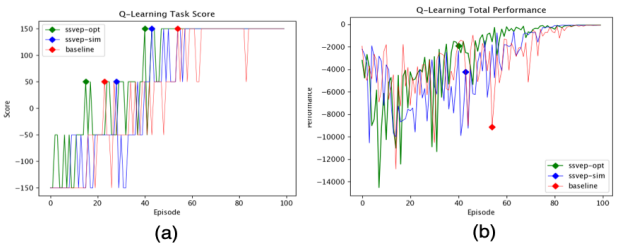

Fig. 7 (a) SSVEP-based methods can quickly achieve the highest rewards at the 40th and 43rd episode compared to the baseline at the 54th episode; (b) the

efficiency of SSVEP was achieved to reach the optimal states, where the marked notations show the first achievement (40th, 43rd vs. 54th episode) was done of the optimal for each approach, while avoiding non-stationarity after the 50th episode for SSVEP-based simulation methods.

TABLE V. THE STATIONARITY OF THE PROPOSED SYSTEM

Stationary Results[a]		
	AVG Steps	**Standard Deviation**
Baseline	1,608.59	1,920.36
SSVEP-sim	**742.43**	**1,323,04**
SSVEP-opt	**444.94**	**545.82**

[a] From the 51st to 100th episode.

The experimental results also show that the SSVEP-based approach can efficiently reach better rewards of 50 points from two robots' success (Fig. 7(a)). In addition, the highest rewards were also first achieved at the 40th, 43rd vs. 54th episode separately for each approach, while non-stationarity of SSVEP can be reduced shown in Fig. 7(b) and TABLE V. Therefore, we anticipate that the proposed *NeuroRL* for HAT system can significantly achieve an integrated collaboration for human perceptions and robot's stationary navigations, to efficiently ensure and consolidate a more natural *neural-enabled* and advanced *decision-making* HAT in the future of dynamic and autonomous environment.

V. CONCLUSION AND FUTURE PLAN

The proposed *NeuroRL* shows that Human-Autonomy proficiency in MADRL is adaptable and reliable for the targeted tasks to achieve efficiency and avoid non-stationarity. Experiment results showed significant improvement while bootstrapping the HAT collaborations. *NeoroRL* can adopt more precise and accurate BCI context models to build a genuinely teaming-oriented, accountable, and flexible human-autonomy system. Thus, it can resolve the lack of human-autonomy design and *neural-enabled* considerations for extremely complicated brain activities for the dynamic RL environments in our daily lives. By integrating the design with DRL and BCI, *NeuroRL* will be an essential and symbolic synthesis, which can efficiently enable fast learning and broaden novel brain formulations of granularity related to robust RL scalability. We foresee that the upcoming HAT in a coherent and autonomous world will leverage moral, reliable, and accounting principles that can more intuitively translate into more efficient and trustworthy behavior. Our further work will be focused more on the *'human-in-the-loop'* algorithms to integrate with our future design. We will design more high-level models for the HAT representations and MADRL environments. Finally, we will turn to multi-party Human-AI models to reinvestigate adaptable and relevant collaborations of a human-machine system in both virtual and practical environments.

REFERENCES

[1] Foerster, Jakob, et al. "Learning to communicate with deep multi-agent reinforcement learning." *Advances in neural information processing systems*. 2016.

[2] Gu, Shixiang, et al. "Continuous deep q-learning with model-based acceleration." *International Conference on Machine Learning*. 2016.

[3] Tampuu, Ardi, et al. "Multiagent cooperation and competition with deep reinforcement learning." *PloS one* 12.4 (2017): e0172395.

[4] S. Frazier and M. Riedl, "Improving deep reinforcement learning in minecraft with action advice," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 15, no. 1, 2019, pp. 146–152.

[5] de Witt, Christian Schroeder, et al. "Multi-agent common knowledge reinforcement learning." *Advances in Neural Information Processing Systems*. (2019).

[6] Thomas, John, et al. "Deep learning-based classification for brain-computer interfaces." *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2017.

[7] Nakanishi, Masaki, et al. "A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials." *PloS one* 10.10 (2015): e0140703.

[8] Ravi, Aravind, et al. "Comparing user-dependent and user-independent training of CNN for SSVEP BCI." *Journal of neural engineering* 17.2 (2020): 026028.

[9] Guney, Osman Berke, Muhtasham Oblokulov, and Huseyin Ozkan. "A deep neural network for ssvep-based brain-computer interfaces." *IEEE Transactions on Biomedical Engineering* 69.2 (2021): 932-944.

[10] Ko, Li-Wei, et al. "SSVEP-assisted RSVP brain–computer interface paradigm for multi-target classification." *Journal of Neural Engineering* 18.1 (2021): 016021.

[11] Wolpaw, Jonathan R., et al. "Brain-computer interface technology: a review of the first international meeting." *IEEE transactions on rehabilitation engineering* 8.2 (2000): 164-173.

[12] Kubacki, Arkadiusz. "Use of force feedback device in a hybrid brain-computer interface based on SSVEP, EOG and eye tracking for sorting items." *Sensors* 21.21 (2021): 7244.

[13] Ho, Joshua, and Chien-Min Wang. "Explainable and adaptable augmentation in knowledge attention network for multi-agent deep reinforcement learning systems." *2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. IEEE, 2020.

[14] Zhang, Ruohan, et al. "Leveraging human guidance for deep reinforcement learning tasks." *arXiv preprint arXiv:1909.09906*(2019).

[15] Ho, Joshua, and Chien-Min Wang. "Human-Centered AI using Ethical Causality and Learning Representation for Multi-Agent Deep Reinforcement Learning." *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*. IEEE, 2021.

[16] Ilhan, Ercüment, Jeremy Gow, and Diego Perez-Liebana. "Teaching on a Budget in Multi-Agent Deep Reinforcement Learning." *2019 IEEE Conference on Games (CoG)*. IEEE, 2019.

[17] Ammanabrolu, Prithviraj, and Mark O. Riedl. "Transfer in deep reinforcement learning using knowledge graphs." *arXiv preprint arXiv:1908.06556* (2019).

[18] Baker, B. (2022, July 20). *Emergent tool use from multi-agent interaction.* OpenAI. Retrieved August 21, 2022, from https://openai.com/blog/emergent-tool-use/

[19] Ma, Jiaxin, et al. "A novel EOG/EEG hybrid human–machine interface adopting eye movements and ERPs: Application to robot control." *IEEE Transactions on Biomedical Engineering*62.3 (2014): 876-889.

[20] Nushi, Besmira, Ece Kamar, and Eric Horvitz. "Towards accountable ai: Hybrid human-machine analyses for characterizing system failure." *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 6. 2018.

[21] Schaul, Tom, et al. "Prioritized experience replay." *arXiv preprint arXiv:1511.05952* (2015).

[22] H. Berger, "Uber das Electrenkephalogramm des Menchen," *Arch Psychiat Nervenkr*, vol. 87, pp. 527–570, 1929.

[23] Peng, Hanchuan, Fuhui Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." *IEEE Transactions on pattern analysis and machine intelligence* 27.8 (2005): 1226-1238.

[24] Neuralink, https://neuralink.com, 2022.

[25] Snap Inc., https://snap.com/en-US, 2022.

[26] Meta, https://about.facebook.com, 2022.