


Article

New Approach in Human-AI Interaction by Reinforcement-Imitation Learning

Neda Navidi *  and Rene Landry, Jr.

LASENA Laboratory, École de Technologies Supérieure (ÉTS), Montreal, QC H3C 1K3, Canada; renejr.landry@etsmtl.ca

* Correspondence: Neda.Navidi@lassena.etsmtl.ca

Abstract: Reinforcement Learning (RL) provides effective results with an agent learning from a stand-alone reward function. However, it presents unique challenges with large amounts of environment states and action spaces, as well as in the determination of rewards. Imitation Learning (IL) offers a promising solution for those challenges using a teacher. In IL, the learning process can take advantage of human-sourced assistance and/or control over the agent and environment. A human teacher and an agent learner are considered in this study. The teacher takes part in the agent's training towards dealing with the environment, tackling a specific objective, and achieving a predefined goal. This paper proposes a novel approach combining IL with different types of RL methods, namely, state-action-reward-state-action (SARSA) and Asynchronous Advantage Actor-Critic Agents (A3C), to overcome the problems of both stand-alone systems. How to effectively leverage the teacher's feedback—be it direct binary or indirect detailed—for the agent learner to learn sequential decision-making policies is addressed. The results of this study on various OpenAI-Gym environments show that this algorithmic method can be incorporated with different combinations, and significantly decreases both human endeavors and tedious exploration process.



Citation: Navidi, N.; Landry, R., Jr. New Approach in Human-AI Interaction by Reinforcement-Imitation Learning. *Appl. Sci.* **2021**, *11*, 3068. <https://doi.org/10.3390/app11073068>

Academic Editors: Jean Baratgin, Thierry Karsenti and Alain Jaillet

Received: 15 February 2021

Accepted: 23 March 2021

Published: 30 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: reinforcement learning; imitation learning; human-AI collaboration

1. Introduction

Reinforcement Learning (RL) in various decision-making tasks provides effective and powerful results with learning for an agent from the stand-alone reward function. However, it suffers with large amounts of environment states and action spaces, and with high implicitness concerning rewards for real complex environments. The complexity, which is due to the high dimensionality and continuousness of real environments, leads to the RL needing a large number of learning trials in order to understand and learn the environment [1]. A promising solution for the limitation is addressed by Imitation Learning (IL) and exploiting teacher feedback. In IL, the learning process can take advantages of human assistance and control over the agent and the environment. In this study, the human is considered as a teacher who teaches a learner to deal with the environment and to tackle a specific object.

A teacher can express his feedback to improve a policy using two main methods, namely, direct dual feedback and indirect detailed feedback. While in the first method the teacher evaluates the agent's actions by sending back rewards (positive or negative), in the second method he can demonstrate the way to complete a task to the agent by actions [2,3]. One of main limitations of existing IL approaches is that they may expect extensive demonstration information in long-horizon problems. Our proposed approach leverages integrated RL-IL structures (see Figure 1) to overcome the RL and IL limitations simultaneously. Moreover, the approach considers both cases where the agent does or does not need human feedback. Our key design principle is a cooperative structure, in which feedback from the teacher is used to improve the learner's behavior, improve the sample efficiency and speed up the learning process by IL-RL integration (See Figure 2).

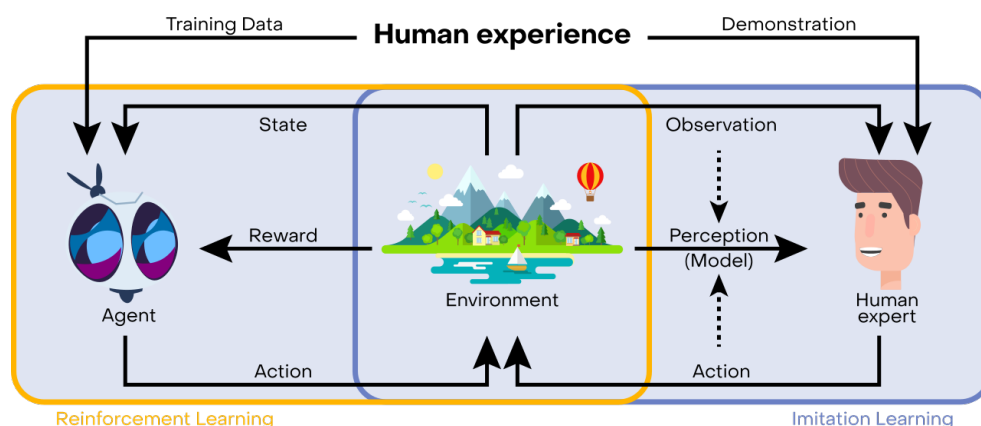


Figure 1. Reinforcement Learning–Imitation Learning (RL-IL) integration structure.

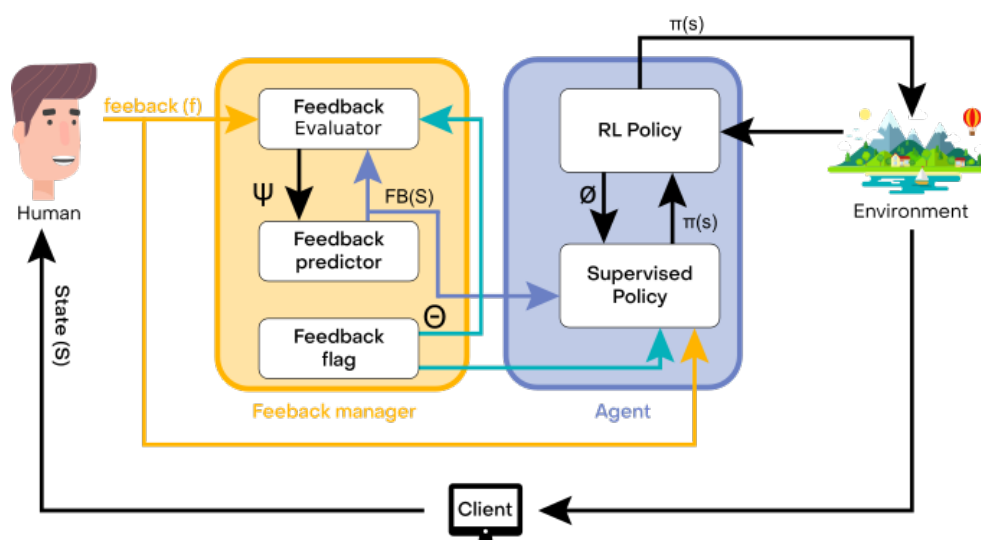


Figure 2. Proposed structure for human–AI interaction.

Teacher assistance considers both direct dual feedback, with positive and negative reward, and indirect detailed feedback, with access to action domain feedback using an online policy IL process. Management of teacher’s feedback in the “feedback management” block is one of the features of the structure (see Figure 2). Moreover, this structure reflects the online teacher feedback as soon as the learner takes action and deals with the quantity of teacher’s feedback.

This paper begins by overviewing the related work on RL and IL in Section 2. It is continued by formalizing the problem of imitation learning and the details of the proposed structure (Section 3). The proposed structure is validated and compared with RL stand-alone in Section 4. Experimental validation and analysis of the results are concluded in Section 5.

2. Related Works

Having an agent learn behavior from a standalone reward, which is the main concept of Reinforcement Learning (RL), is particularly difficult in a complicated environment. The main problems are high dimensionality of environment spaces in challenging tasks. Moreover, the definition of reward function in real-word applications is very complicated and implicit. Contribution of humans and agents in the form of using human knowledge in the training loop by Imitation Learning (IL) is a promising solution to improve the data efficiency and to gain a robust policy [4].

In IL, the agent observes, trusts and replicates the teacher’s behavior. For a typical method of IL, which is presented as Behavioural Cloning (BC) or Learning from Demon-

stration (LfD), the goal is to train a classifier-based policy to predict the teacher's actions. In BC, features are lists of an environment's features and the labeled data are actions performed by the teacher. However, the statistical learning assumption affected by ignoring the relationship of current action and next states during execution of the learned policy causes poor performance concerning this method [5,6].

The forward training algorithm in IL was introduced to train one policy at each time step to achieve a non-stationary policy. In the training algorithm, the agent learns how to imitate the teacher's behaviour in the states generated by the previous learned policies [7]. The main disadvantage of the forward training algorithm is that it requires an investigation of the environment over all periods, regardless of the horizon size. In fact, considering the non-stationary policy in this model causes its impracticality in real-world applications with a large horizon.

Search-based structured prediction (SEARN) learns a classifier to choose the search optimally. This model outperforms the traditional search-based methods which first learn some sort of global model and then start exploring. SEARN follows the teacher's action at the beginning of a task. Then it aggregates more demonstrative data iteratively to obtain an updated policy. It generates new episodes to create a combination of previous models and teacher behaviour [8]. However, the optimistic prediction, as a result of the difference between its initial value and the optimal policy, is the main drawback of this learning method.

Stochastic Mixing Iterative Learning (SMILe) has been proposed to improve the forward training algorithm using SEARN's benefits with a straightforward implementation and less dependency on interaction with a teacher. After several iterations, the method utilizes a geometric policy by training a stochastic policy [7]. While the training process can be interrupted at any time, it suffers from instability in the model because of the stochastic policy assumption.

Two popular IL algorithms called dataset aggregation (DAGGER) and Generative Adversarial Imitation Learning (GAIL) [4,9] introduce new approaches for incorporating teacher experience. These papers proposed iterative algorithms for an online learning approach to train for a stationary deterministic policy. It was proved that the combination of the algorithms with reduction-based approaches outperforms the policy findings in sequential settings, thanks to reuse of existing supervised learning algorithms.

DAGGER performs well for both complex and simple problems; however, the information may not be intuitive from the states [10]. Moreover, GAIL presented considerable achievement on imitation learning, especially for complex continuous environments. However, it suffers from the requirement of a huge amount of interactions during the training. Furthermore, it is very time-consuming in real-world applications, where there needs to be more interactions between an agent and the environment to achieve an appropriate model [11].

Reference [12] shows that Reductions-based Active Imitation Learning (RAIL) consists of N iterations, where each iteration has a specific stationary policy over time steps and has a significant difference with the previous iteration. This method provides a small error at the expert actions prediction, considering the state distribution of the former iteration. Nevertheless, the results in [12] can be faulty and impractical in some cases due to the unlabeled state distributions in the previous iterations.

As has been presented in the research studies above, all IL methods mostly suffer from instability in the model because of the stochastic policy assumptions. Moreover, the labeled-information needs make the presence of an expert human necessary in order to annotate the dataset. These two main drawbacks prevent the use of IL for high-dimensional high-frequency real-world applications. Fortunately, a promising solution is the integration of IL with RL to overcome these aforementioned limitations.

The idea of exploiting IL to increase the speed of convergence of RL has been considered in [13]. However, that study considers the stochastic policy and uses IL as a "pre-training" solution to speed up the convergence. IL has been considered as a pre-

training step for reward reshaping [14], policy reshaping [15] and knowledge transfer [16] with teacher feedback.

Reference [17] uses a reward shaping method which is one of the significant aspects of RL. Reference [18] describes that this method is an accepted way for human feedback in RL, but it causes some issues where human feedback signals may contain inconsistency, infrequency, ambiguity and insufficiency [17]. As an example, translating statements into rewards may be difficult and unclear; accordingly [19], tried to solve this problem considering a drift parameter to reduce the dependency on human feedback signals. To overcome some of the aforementioned limitations, reference [20] proposed an UNDO function as a policy feedback which contains a reversible feedback signal for agents. The results in [18] demonstrate that the human feedback signals can improve RL algorithms by applying them in the process of action selection. Some recent studies use the human feedback as an optimal policy instead of a reward shaping method like agent's exploration seed [21] and inverse RL control [22] and even as a substitute of exploration [16,17].

The core of this paper provides an accessible and effective structure for the agent to become an expert with teacher help and advice. It also addresses a set of generic questions, namely, what should be imitated, how the agent may imitate, when is the time to imitate and who is trustworthy to imitate. Moreover, it addresses when teachers are available and how their feedback can be the most effectively leveraged.

3. Teacher Assistance Structure

The proposed structure exploits teacher feedback as a rectification of the action domain of the learner; as soon as an action is performed by the agent, this teacher feedback improves the online policy. It can also infer the policy of the agent from infrequent teacher feedback. Four main characteristics of human teacher feedback and their related effects are considered and formulated; namely, duality, reaction time delay, contingency and instability.

While several studies consider a range for teacher feedback, like very bad, bad, good, very good or -100 , -50 , 50 , 100 , giving feedback by humans from a range is very complicated and requires very good knowledge of the task and environment. This study takes advantage of duality feedback when a human teacher is satisfied with the decision made by agent or not. This kind of feedback can be sent by expert or non-expert human due to its simplicity concerning knowledge transferral.

The next feature that is considered is the reaction time delay of the human to send feedback. Several studies like [23] present the sample efficiency for training neural reinforcement when it is pre-trained by an expert using the supervised learning method. In fact, in different IL algorithms like DAGGER and GAIL, which are based on offline learning, there is no need to consider the reaction time delay of the human in the model. In the mentioned algorithms just an expert should prepare a time-consuming metadata before starting the training process. Moreover, unprofessional feedback from a non-expert teacher can ruin the training process.

The proposed structure for human–AI interaction presents a methodology to enable AI agents to interact with the human (teacher, expert or not) completely online; we should deal with the delay in the reaction of humans [24,25]. Using the teacher feedback in online training without recognizing that delay can make the training process impractical. However the reaction time delay is not constant and it would vary depending on teacher personality, teacher knowledge, complexity of the environment and ambiguity of actions.

In addition to the reaction time delay, the contingency of the human teacher feedback as a feature of reactive manner is dealt with. Due to limited patience, mostly the human teacher stops to send positive feedback while the agent carries out actions correctly. Moreover, the frequency of releasing the feedback can vary based on human preference [26]. So the proposed methodology considers a module named “feedback predictor” (See Figure 2) to present the contingency and stochasticity of correct feedback which is sent in a specific timestamp.

The details of the structure are explained as follows (See Figure 2):

- Feedback predictor: This section gathers the previous feedback and advice, and predicts the next action to be taken by the human.
- Supervised policy: This module can improve the variables of the RL policy;
- Feedback evaluator: This module can upgrade the variables of the supervised policy module;
- Feedback flag: This module can manage and control the time-lags and postponement of human response;
- RL policy: This considers an on-policy algorithm of Q-learning and a one policy-based RL algorithm.

As soon as the agent picks an action, the supervisor can observe the outcome of that action on the environment and send his feedback. This feedback (f) is a positive, negative or neutral value to intimate that the last performed action should be modified increasingly or decreasingly. The neutral value is considered when the teacher is not available or he prefers not to expose his idea. The environment state (S), performed action, $A(S)$, and teacher feedback (f) are sent back to the “feedback evaluator” and “supervised policy” sections to update the Φ and Ψ .

3.1. Feedback Predictor

$FB(s)$ shows the policy learned in the “feedback manager” box and is able to predict the next feedback of the teacher by observing the current state and the agent’s action. The dual teachers feedback is in the range $[-1, 1]$; -1 shows that the teachers are not satisfied with the action taken by the agent, so they send a request to the agent to stop or reduce it. On the other hand, they send back $+1$ whenever the taken action is convincing, so the teachers encourage the agent to continue it. Moreover, an adjustable learning rate considered to improve the online and offline models on the online training dataset is monitored by the learning algorithm and the learning rate can be adjusted in response. The policy is formulated by Equation (1):

$$FB(O, A) = \psi^T \theta(O, A) \quad (1)$$

where $FB(O, A)$, ψ and $\theta(O, A)$ are the teacher feedback policy, the parameters vector and the probability density function delay of the human’s feedback signal, respectively. Details about these parameters are explained in the next sections.

3.2. Supervised Policy

The policy can be updated and modified directly using a supervised policy based on supervised learning methods. In fact, the agent can change its actions based on state-action training pairs provided real-time by the supervisor, without considering the value of the training data. This element can improve the model parameters, using state-action-reward-state-action (SARSA) from value-based RL algorithms or Asynchronous Advantage Actor–Critic Agent (A3C) from policy-based algorithms.

RL algorithms are required for the optimization process, whereas the teacher helps the agent to gain a level of skill while the RL algorithms provide poor estimation of value functions. The supervised policy module provides both sorts of error information for the agent as long as the actions are for the environment.

The agent receives evaluations of its behaviour that can help in carrying out the given task. When the agent becomes a professional concerning the task, the teacher gradually withdraws the additional feedback to shape the policy toward optimality for the true task. The error of the prediction is not clear because of the uncertainty of the quantified human feedback. This is considered in Equation (2):

$$e(t) = r(t) * k \quad (2)$$

where $e(t)$, $r(t)$ and k present the prediction error, error sign extracted from human feedback and constant error value predefined by the user, respectively.

3.3. Feedback Evaluator

The responsibility of the feedback evaluator is to update the parameters vector (ψ) of the teacher feedback policy ($FB(s)$). $FB(s)$ can be calculated by multiplying the teacher feedback and the parameters vector. This can be rewritten as Equation (3):

$$\Delta\psi = \gamma(f - FB(s)) \cdot \left(\frac{\delta(FB(s))}{\delta\psi} \right) \quad (3)$$

where γ is the adjustable learning rate and can be observed by:

$$\gamma(s) = \|FB(s)\| + b \quad (4)$$

where b shows the predefined value of the learning rate as the bias of the model. The variation between the actual teacher feedback and the predicted teacher feedback (calculated from the teacher feedback policy) is considered as the prediction error.

3.4. Feedback Flag

The action spaces of the control system are generally dichotomized in continuous and discrete modes. Continuous control systems mostly are designed to deal with continuous action space, especially in high-frequency environments like video games. Speed of system, time delay of call response and non-constancy of the human-response rate make communication between the system and human very difficult in these environments. To deal with this problem, The “feedback flag” module is presented to bufferize and integralize the several past state-action tuples. Each past state-action tuple is weighted with the corresponding probability that characterizes the delay in teacher response and is shown by $RD(t)$.

It is used by the “feedback evaluator” and the “supervised policy”. The teacher feedback function is a linear model (Equation (1) and Figure 2). The uncertainty of the feedback’s receiving time, time t , is defined as $t_1 < t < t_n$, and directly affects the agent that is trying to attach the reward signal to a specific action in time. This feedback could in fact be attached to any prior action at time $t - 1, t - 2, \dots, t - n$. This is why we need to use (Θ) to define the delay of the human’s response signal:

$$\Theta_{t_0} = \int_{t-t_0}^{t_0} RD(t) \quad (5)$$

where Θ is the density of the continuous human’s response.

3.5. Reinforcement Learning (RL) Policy

In RL, we consider a predefined environment. In such an environment, an agent performs actions and reactions sequentially to complete a task, using its observations of the environment and the rewards it gets from it. The agent can choose an action from an action space, $A_s = A_1, A_2, \dots, A_n$. That action passes to the stochastic environment and return a new observation space, $O_s = O_1, O_2, \dots, O_n$, and reward, $R_s = R_1, R_2, \dots, R_n$. At each step, the agent observes the current state game state and cannot understand the whole task by observing just the current game state. Moreover, the Markov Decision Process (MDP) is a fundamental of RL. MDP can be used in a cooperative structure for the decision-making tasks to partly or completely control and balance agents.

The first RL algorithm in this study is SARSA and its details are presented in Algorithm 1. SARSA is very similar to Q-learning. The key difference between them is an on-policy, whereas Q-learning is a class of off-policy Temporal Difference (TD). This implies that the SARSA learning process deals with the actions taken by the current policy instead of the greedy policy. So the SARSA update (see Equation (6)) does not consider the maximum value and greedy policy (Equation (2)).

$$Q(O_n, A_s) = Q(O_n, A_s) + \alpha[R_{n+1} + \gamma Q(O_{n+1}, A) - Q(O_n, A_s)] \quad (6)$$

On the other hand, Q-learning consists of a multi-layer neural network (NN), where its inputs would be states of an environment and the outputs would be the action value, $Q(s, \theta)$, where θ is the parameters of NN. In fact, Q-learning updates for the parameter after taking action A_n after observing the state O_n , and receive an immediate reward R_n and Q_{n+1} . So the update equation is given by:

$$Q(O_n, A_s) = Q(O_n, A_s) + \alpha[R_n + \gamma \max Q(O_{n+1}, A) - Q(O_n, A_s)] \quad (7)$$

This equation shows that the policy considered to select an action is a greedy policy calculated by Equation (8):

$$\max Q(O_{n+1}, A) - Q(O_n, A_s) \quad (8)$$

The second RL algorithm in this study is Asynchronous Advantage Actor–Critic Agents (A3C), a policy gradient algorithm with a special focus on parallel training. In A3C, the critic part learns the value function while the actor part is trained in parallel and becomes synchronized with the global parameters sequentially. In A3C, there is a loss function for state value to minimize the Mean Square Error (MSE) (Equation (9)) and this is the baseline in the policy gradient update. Finally, the gradient descent can be applied to find the optimal value. For more details about the A3C see the Algorithm 2.

$$J_V(\omega) = (G_t - V_\omega(s))^2 \quad (9)$$

Algorithm 1 State–action reward–state–action (SARSA)

Require:

Observations: $O_s = O_1, O_2, \dots, O_n$,

Actions: $A_s = A_1, A_2, \dots, A_n$,

Reward Function: $R_s = R_1, R_2, \dots, R_n$,

Transition: $T_s = T_1, T_2, \dots, T_n$,

Initialization:

Learning rate: $\alpha \in [0, 1]$, initialized with = 0.1

Discount factor: $\gamma \in [0, 1]$

Balancing rate: $\lambda \in [0, 1]$; Trade off between Temporal-Difference and Monte-Carlo

Process Q-learning ($O, A, R, T, \alpha, \gamma, \lambda$)

while Q is not converged **do**

 select $[O, A] \in O_s$

while O is not terminated **do**

$R \leftarrow R(O, A)$

$O' \leftarrow T(O, A)$

$A' \leftarrow \phi(O')$

$e(O, A) \leftarrow e(O, A) + 1$

$\sigma \leftarrow \lambda Q(O_{n+1}, A_{n+1}) - Q(O_n, A_n)$

for $[O', A'] \in O_s$ **do**

$Q(O', A') \leftarrow Q(O', A') + \alpha * \gamma * \sigma * e(O', A')$

$e(O', A') \leftarrow \gamma * \lambda * e(O', A')$

$O' \leftarrow O$

$A' \leftarrow A$

end for

end while

end while

return Q

Algorithm 2 Asynchronous Advantage Actor-Critic Agents (A3C)**Require:**

requirements of Algorithm 1,

Initialization:

Meta parameters: θ, ω **while** $T < T_{max}$ **do** $\theta \leftarrow 0$ and $\omega \leftarrow 0$ **while** O is not terminated **do** $A' \leftarrow \phi(O')$ $R \leftarrow R(O, A)$ $O' \leftarrow T(O, A)$ **for** $i = t - 1 : t_{start}$ **do**update $R, d\theta$ and $d\omega$ **end for****end while****end while****return** Q **4. Experiments and Results**

The performance of the proposed Algorithms 3 and 4 is evaluated on two separate use-cases:

- Continuous classic cart-pole OpenAI-Gym environment
- Continuous classic mountain-car OpenAI-Gym environment.

Continuous cart-pole and mountain-car are used in this study as continuous classical OpenAI-Gym environment. The objective of the cart-pole system is to adjustably control the cart by taking continuous and unlimited actions. The cart has two degrees of freedom (DoF) to balance with the horizontal axes. The system state is parameterized by orientation, position and velocity of both pole and card. The stability of this system is defined by orientation from -12° to 12° , and position deviation between -2.4 to 2.4 (see Figure 3a). Whenever the system gets unbalanced, a negative signal as a punishment acknowledgement is sent back to the system and it will be reset.

The next OpenAI-Gym considered in this study is continuous mountain-car illustrated in Figure 3b. This environment presents a car on a sinuous curve track, located between two mountains. The goal is to drive up the right mountain; however, the car's engine is not strong enough to scale the mountain in a single pass. Therefore, the only way to succeed is to drive back and forth to build up momentum. Here, the reward is greater if you spend less energy to reach the goal.

The results of applying different algorithms (see Algorithms 1–4) on a continuous classic cart-pole in an OpenAI-Gym environment are presented in Figure 4a. At each step, the agent is rewarded for balancing the horizontal axes. The results for both “Hybrid A3C/IL” and “Hybrid SARSA/IL” show that the proposed algorithms based on integration-imitation learning and reinforcement learning can overcome the stand-alone reinforcement learning, A3C and SARSA. Figure 4a presents Hybrid A3C/IL converging faster (in Episode # 70) than Hybrid SARSA/IL. The reason for the acceleration of the convergence by Hybrid A3C/IL generally is based on the accuracy of policy-based reinforcement learning in continuous environments. This is proven by comparing stand-alone SARSA and A3C presented by blue and pink dots in the figure. It shows that the value-based reinforcement learning (SARSA here) in a continuous environment like a cart-pole is not satisfying regarding data efficiency. Finally, Hybrid A3C/IL increases the data efficiency of the cart-pole environment by 85.7%, 53.8% and 14.2% compared to SARSA, A3C and Hybrid SARSA/IL, respectively.

Algorithm 3 Hybrid State-action reward-state-action imitation learning (SARSA/IL)**Require:**

Exact requirements of Algorithm 1,

Initialization:

Same as Algorithm 1,

while Q is not converged **do** select $[O, A] \in O_s$ **while** O is not terminated **do** $R \leftarrow R(O, A)$ $O' \leftarrow T(O, A)$ $A' \leftarrow \phi(O')$ $e(O, A) \leftarrow e(O, A) + 1$ $\sigma \leftarrow \lambda Q(O_{n+1}, A_{n+1}) - Q(O_n, A_n)$ **for** $[O', A'] \in O_s$ **do** **if** f is exist: **then**

Consider Equation (5)

Consider Equation (1)

Consider Equation (3)

Consider Equation (4)

end if $Q(O', A') \leftarrow Q(O', A') + \alpha * \gamma * \sigma * e(O', A')$ $e(O', A') \leftarrow \gamma * \lambda * e(O', A')$ $O' \leftarrow O$ $A' \leftarrow A$ **end for** **end while****end while****return** Q **Algorithm 4** Hybrid Asynchronous Advantage Actor-Critic Agents Imitation Learning (Hybrid A3C/IL)**Require:**

requirements of Algorithm 1,

Initialization:

Meta parameters: θ, ω **while** $T < T_{max}$ **do** $\theta \leftarrow 0$ and $\omega \leftarrow 0$ **while** O is not terminated **do** $A' \leftarrow \phi(O')$ **if** f is exist: **then**

Consider Equation (5)

Consider Equation (1)

Consider Equation (3)

Consider Equation (4)

end if $R \leftarrow R(O, A)$ $O' \leftarrow T(O, A)$ **for** $i = t - 1 : t_{start}$ **do** update $R, d\theta$ and $d\omega$ **end for** **end while****end while****return** Q

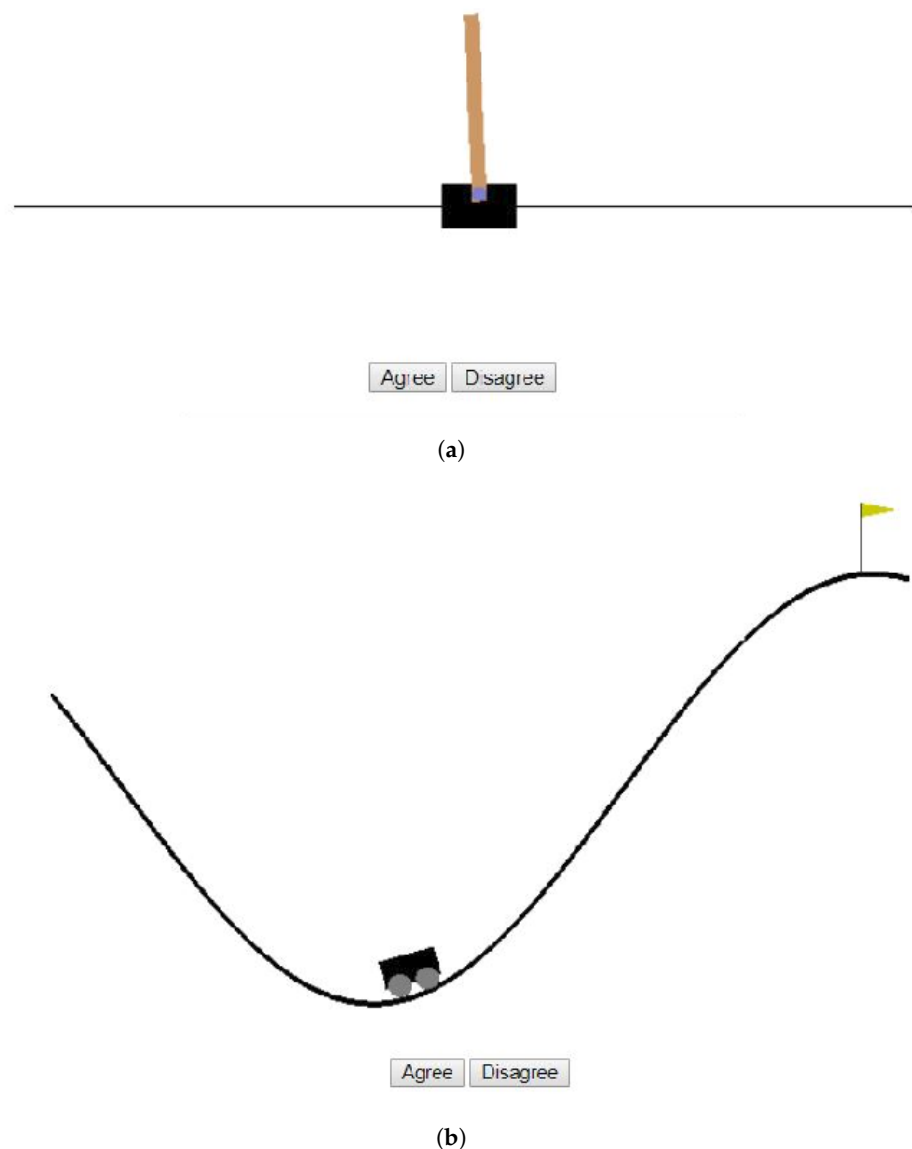
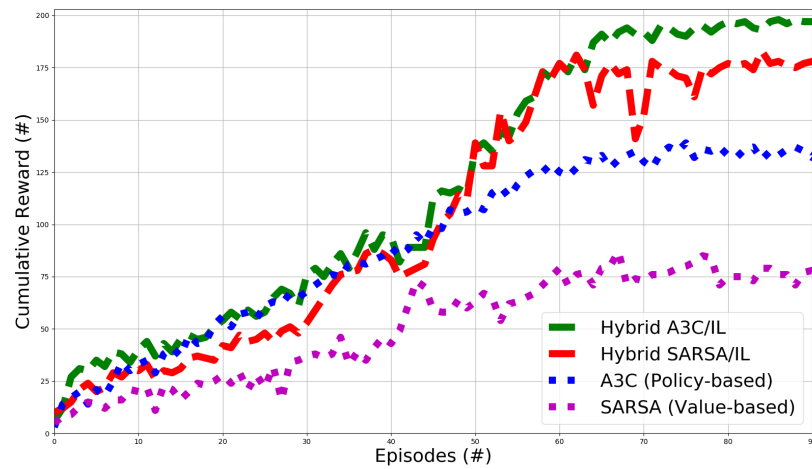
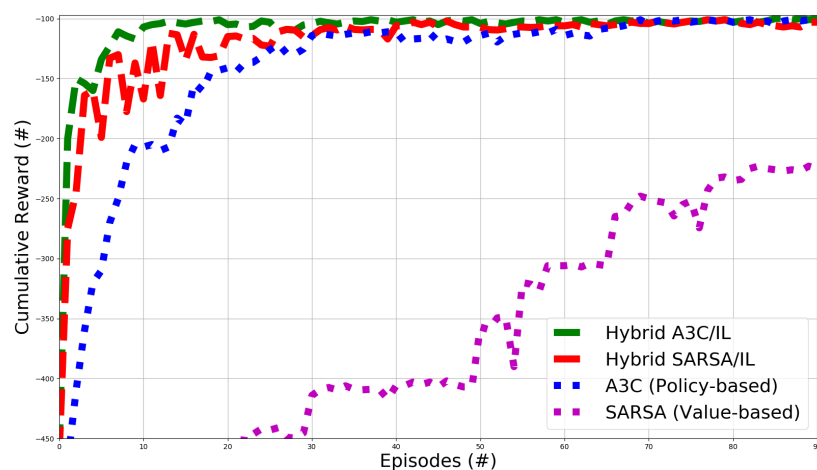


Figure 3. Continuous classic OpenAI-Gym environment: (a) Continuous cart-pole; (b) continuous mountain-car in the proposed framework.

Moreover, the achievements of utilizing Algorithms 1–4 on continuous classic mountain-car in an OpenAI-Gym environment are shown in Figure 4b. In this environment, the agent receives a punishment (negative reward) for each step of an episode. The maximum performance of this environment is -100 cumulative rewards and it shows that the minimum number of steps in an epoch to gain the flag on top of the hill in the continuous environment is 100. Like the cart-pole environment, the results for both “Hybrid A3C/IL” and “Hybrid SARSA/IL” show that the proposed algorithms based on the integration–imitation learning and reinforcement learning outperform the A3C and SARSA, as examples of policy-based and value-based RL. Figure 4b presents Hybrid A3C/IL and Hybrid SARSA/IL converging faster concerning the two RL algorithms. However, Hybrid SARSA/IL shows lots of oscillations before stabilising at episode # 20. The reason for these fluctuations is that the value-iteration-based RL cannot act well in complicated continuous environments regarding exploration and evaluation of the tuple of state and action. Hybrid A3C/IL and Hybrid SARSA/IL both increase data efficiency about 60% and 33.4% compared to SARSA and A3C, respectively.



(a)



(b)

Figure 4. Experimental results: (a) Continuous classic cart-pole OpenAI-Gym environment, (b) continuous classic mountain-car OpenAI-Gym environment.

5. Conclusions

In this paper, a novel approach is proposed which combines IL with different types of RL methods, namely, state-action-reward-state-action (SARSA) and Asynchronous Advantage Actor–Critic Agents (A3C), to take advantage of both the IL and RL methods. Moreover, we address how to effectively leverage the teacher’s feedback for the agent learner to learn sequential decision-making policies. The results of this study on a simple OpenAI-Gym environment show that Hybrid A3C/IL increases the data efficiency of the cart-pole environment by 85.7%, 53.8% and 14.2% compared to SARSA, A3C and Hybrid SARSA/IL, respectively. Moreover, the results on a complicated OpenAI-Gym environment show that Hybrid A3C/IL and Hybrid SARSA/IL both increase data efficiency by about 60% and 33.4% compared to SARSA and A3C, respectively.

Author Contributions: Also, all authors have contributed equally. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Mnih, V.; Badia, A.P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 1928–1937.
- Meriçli, C.; Veloso, M.; Akin, H.L. Complementary humanoid behavior shaping using corrective demonstration. In Proceedings of the 2010 10th IEEE-RAS International Conference on Humanoid Robots, Nashville, TN, USA, 6–8 December 2010; pp. 334–339.
- Christiano, P.F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; Amodei, D. Deep reinforcement learning from human preferences. *arXiv* **2017**, arXiv:1706.03741.
- Ross, S.; Gordon, G.; Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 627–635.
- Moore, R.; Caines, A.; Rice, A.; Buttery, P. Behavioural Cloning of Teachers for Automatic Homework Selection. In Proceedings of the International Conference on Artificial Intelligence in Education, Chicago, IL, USA, 25–29 June 2019; pp. 333–344.
- Englert, P.; Paraschos, A.; Peters, J.; Deisenroth, M.P. Model-based imitation learning by probabilistic trajectory matching. In Proceedings of the 2013 IEEE International Conference on Robotics and Automation, Karlsruhe, Germany, 6–10 May 2013; pp. 1922–1927.
- Ross, S.; Bagnell, D. Efficient reductions for imitation learning. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; pp. 661–668.
- Daumé, H.; Langford, J.; Marcu, D. Search-based structured prediction. *Mach. Learn.* **2009**, *75*, 297–325. [[CrossRef](#)]
- Ho, J.; Ermon, S. Generative adversarial imitation learning. *arXiv* **2016**, arXiv:1606.03476.
- Attia, A.; Dayan, S. Global overview of imitation learning. *arXiv* **2018**, arXiv:1801.06503.
- Sasaki, F.; Yohira, T.; Kawaguchi, A. Sample efficient imitation learning for continuous control. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
- Judah, K.; Fern, A.P.; Dietterich, T.G. Active imitation learning via reduction to iid active learning. *arXiv* **2012**, arXiv:1210.4876.
- Hester, T.; Vecerik, M.; Pietquin, O.; Lanctot, M.; Schaul, T.; Piot, B.; Horgan, D.; Quan, J.; Sendonaris, A.; Osband, I.; et al. Deep q-learning from demonstrations. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
- Brys, T.; Harutyunyan, A.; Suay, H.B.; Chernova, S.; Taylor, M.E.; Nowé, A. Reinforcement learning from demonstration through shaping. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015.
- Griffith, S.; Subramanian, K.; Scholz, J.; Isbell, C.L.; Thomaz, A.L. Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in Neural Information Processing Systems*; Georgia Institute of Technology: Atlanta, GA, USA, 2013; pp. 2625–2633.
- Wang, G.F.; Fang, Z.; Li, P.; Li, B. Transferring knowledge from human-demonstration trajectories to reinforcement learning. *Trans. Inst. Meas. Control* **2018**, *40*, 94–101. [[CrossRef](#)]
- Ng, A.Y.; Harada, D.; Russell, S. *Policy Invariance under Reward Transformations: Theory and Application to Reward Shaping*; ICML: Long Beach, CA, USA, 1999; Volume 99, pp. 278–287.
- Pilarski, P.M.; Dawson, M.R.; Degris, T.; Fahimi, F.; Carey, J.P.; Sutton, R.S. Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning. In Proceedings of the 2011 IEEE International Conference on Rehabilitation Robotics, Zurich, Switzerland, 29 June–1 July 2011; pp. 1–7.
- Isbell, C.L.; Kearns, M.; Singh, S.; Shelton, C.R.; Stone, P.; Kormann, D. Cobot in LambdaMOO: An adaptive social statistics agent. *Auton. Agents -Multi-Agent Syst.* **2006**, *13*, 327–354. [[CrossRef](#)]
- Thomaz, A.L.; Breazeal, C. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artif. Intell.* **2008**, *172*, 716–737. [[CrossRef](#)]
- Taylor, M.E.; Suay, H.B.; Chernova, S. Integrating reinforcement learning with human demonstrations of varying ability. In Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2, International Foundation for Autonomous Agents and Multiagent Systems, Taipei, Taiwan, 2–6 May 2011; pp. 617–624.
- Ziebart, B.D.; Maas, A.L.; Bagnell, J.A.; Dey, A.K. *Maximum Entropy Inverse Reinforcement Learning*; AAAI: Chicago, IL, USA, 2008; Volume 8, pp. 1433–1438.
- Chevalier-Boisvert, M.; Bahdanau, D.; Lahlou, S.; Willems, L.; Saharia, C.; Nguyen, T.H.; Bengio, Y. BabyAI: First steps towards grounded language learning with a human in the loop. *arXiv* **2018**, arXiv:1810.08272.
- Cauchard, J.R.; Zhai, K.Y.; Spadafora, M.; Landay, J.A. Emotion encoding in human-drone interaction. In Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, New Zealand, 7–10 March 2016; pp. 263–270.

-
25. Jaderberg, M.; Czarnecki, W.M.; Dunning, I.; Marris, L.; Lever, G.; Castaneda, A.G.; Beattie, C.; Rabinowitz, N.C.; Morcos, A.S.; Ruderman, A.; et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science* **2019**, *364*, 859–865. [[CrossRef](#)] [[PubMed](#)]
 26. Peng, B.; MacGlashan, J.; Loftin, R.; Littman, M.L.; Roberts, D.L.; Taylor, M.E. A need for speed: Adapting agent action speed to improve task learning from non-expert humans. In Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, Singapore, 9–13 May 2016.