

Decision Making for Human-in-the-loop Robotic Agents via Uncertainty-Aware Reinforcement Learning

Siddharth Singi^{*†}, Zhanpeng He^{*‡}, Alvin Pan[‡], Sandip Patel[†], Gunnar A. Sigurdsson[§],
Robinson Piramuthu[§], Shuran Song[‡] and Matei Ciocarlie[†]

<https://roamlab.github.io/hula/>

Abstract—In a Human-in-the-Loop paradigm, a robotic agent is able to act mostly autonomously in solving a task, but can request help from an external expert when needed. However, knowing when to request such assistance is critical: too few requests can lead to the robot making mistakes, but too many requests can overload the expert. In this paper, we present a Reinforcement Learning based approach to this problem, where a semi-autonomous agent asks for external assistance when it has low confidence in the eventual success of the task. The confidence level is computed by estimating the variance of the return from the current state. We show that this estimate can be iteratively improved during training using a Bellman-like recursion. On discrete navigation problems with both fully- and partially-observable state information, we show that our method makes effective use of a limited budget of expert calls at run-time, despite having no access to the expert at training time.

I. INTRODUCTION

Reinforcement Learning (RL) has shown great progress in learning decision-making for complex robotic skills [1–4] using experiences collected by a robotic agent exploring an environment and receiving reward signals. Traditional RL agents use a policy learned during training in order to act autonomously at deployment. However, even a well-trained agent can encounter situations when deployed that are hard to make decisions for, for reasons such as partial state observability, uncertain dynamics, changes in state distributions between training and testing, etc.

The Human-in-the-Loop (HitL) paradigm has been developed in robotics precisely for situations where an agent can act autonomously most of the time, but would still benefit from receiving assistance from an available (tele-)operator, usually assumed to be a human expert. This paradigm is particularly powerful if the agent itself makes the decision of when to request assistance, thus freeing the operator from having to monitor task progress. However, this approach gives rise to a critical decision-making problem on the agent’s part: when to request help? Too few such requests can lead to the robot making mistakes, but too many requests will overload the expert and lose the benefit of semi-autonomous operation [5].

^{*} denotes joint first authorship.

[†]Dept. of Mechanical Engineering, [‡]Dept. of Computer Science, Columbia University, New York, USA

[§]Amazon Alexa AI

This work was supported in part by the Columbia Center for Artificial Intelligence Technology in collaboration with Amazon.

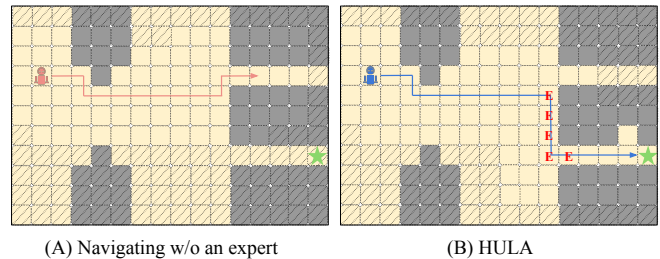


Fig. 1: An illustration of HULA, the method we propose in this paper. In a partially-observable environment, an agent without the help of an expert (A) cannot localize itself accurately due to partial observability, goes down the wrong passage and fails to reach the target. A HULA agent (B) decides to request assistance from an available external expert in the states marked with a red E and achieves the goal (denoted by a green star \star). In both cases, the agent can only observe a 5x5 grid around its current location; shadow areas represent unobserved regions throughout the navigation.

In this paper, we propose a DRL-based method for a HitL agent to make the critical determination of when to request expert assistance. We posit that the best moment to request such assistance is when the agent is highly uncertain in the successful outcome of the task. From an RL perspective, we relate this uncertainty to the variance of the return from the current state, as perceived by the agent. Numerous RL algorithms provide methods for the agent to estimate the expected return from a given state. We show that similar methods can be used to estimate the variance of return as well during the training process. At deployment time, the agent can then request expert assistance when its estimate of the return variance from the current state falls below a given threshold. We dub our method HULA, for Human-in-the-loop Uncertainty-aware Learning Agent, and illustrate its operation in Fig. 1.

Critically, HULA does not need to make any calls to the expert during training. This stands in contrast to a standard RL approach, where an agent could learn how to use an expert simply by making numerous such calls at train time. Nevertheless, our method is able to make effective use of a limited budget of expert calls in order to improve task performance. We summarize our key contributions as follows:

- To the best of our knowledge, we are the first to propose a method for an HitL RL agent to learn how to effectively budget its interactions with a human expert at deployment time, without needing any expert calls

during training.

- We show that the variance of return, which can be estimated during training using Bellman-like equations, is an effective measure for agent uncertainty and can be used at deployment to determine when to request assistance.
- Our experiments on discrete navigation problems with both fully- and partially-observable state information show that our method is as effective or better in managing its budget of expert calls compared to a standard learning approach that also makes thousands of expert calls during training.

II. RELATED WORK

Our work takes advantage of human assistance by learning a human-in-the-loop policy. Researchers have investigated how to incorporate human feedback to improve policy performance [6–8]. For instance, Knox and Stone [9] propose the TAMER framework, which learns a model to minimize the difference between the reinforcement function from a policy and a human. Celemin and Ruiz-del-Solar [10] propose COACH, an interactive framework that utilizes human corrective advice on the magnitude of continuous actions during training a policy. Compared to these works, which focus on using human assistance to derive high-performance policies in train time, HULA achieves higher task performance by learning a policy that can actively seek assistance from humans in test time.

Researchers have also investigated how to learn human-in-the-loop policies. For example, Kessler et al. suggest that robots can effectively ask for human assistance by modeling how an utterance affects human belief. Arakawa et al. propose DQN-TAMER [11], which is a method that incorporates a human observer model by real-time human feedback during training. Expected Local Improvement (ELI) [12] trains a state selector that suggests states requiring to query new actions from human experts. PAINT [13] learns a classifier to identify irreversible states and query the expert in case entering such states cannot be avoided. Hug-DRL [14] leverages a control transfer mechanism between humans and automation that corrects the agent’s unreasonable actions during training by human intervention, where it has demonstrated significant potential in autonomous driving applications. Barnett et al. [15] suggests learning a reward function of selecting from human teachers can lead to a querying policy that achieves good task performance. While these methods exploit human intervention during training, our work relies on the internal uncertainty of an agent and only uses an expert in test time.

HULA explicitly estimates the uncertainty of an RL agent and use it to make decisions about requesting an expert’s assistance. Prior work has explored representing the uncertainty of outcomes during RL training [16–18]. For instance, Kahn et al. use the variance of the dynamics model to represent the uncertainty of collision to avoid damage from high-speed collisions [19]. Ensemble Quantile Networks (EQN) [20] estimates both aleatoric and epistemic uncertainties via the combination of quantile networks (IQN) [21] and randomized

prior function (RPF) [22]. These works use the uncertainty of an agent to perform task-specific decision-making, while our work differs in using it to collaborate with a human. Moreover, our work represents an agent’s uncertainty using the variance of returns, which is a direct indicator of task performance.

The closest work to ours is RCMP, which learns a HitL policy that queries the expert when the epistemic uncertainty estimated by the variance of multiple value functions is high [23]. However, RCMP requires expert queries during training, whereas our work learns an uncertainty model based on the variance of return and does not query an expert in train time.

III. METHOD

Consider a problem for an autonomous agent formulated as a standard Markov decision process (MDP). A general MDP is defined as a tuple of (S, A, r, p) , where S represents the state space, A represents the action space, $r(s, a)$ is a reward function that evaluates immediate reward of action $a \in A$ in a state $s \in S$, and p represent a transition distribution $p(s_{t+1}|s_t, a_t)$. The goal of solving an MDP is to learn a policy $\pi(a|s)$ that maximizes its expected return $\mathbb{E}_\pi[R]$, where $R = \sum \gamma^t r(s_t, a_t)$ and γ is a discount factor.

Going beyond the standard MDP formulation, we now assume the availability of an expert who can give instructions to the robotic agent. When queried from a given state s_t , this expert can directly provide an action $a_{exp}(s_t)$ for the robot to execute. We assume that the expert always provides high-quality advice, i.e. a policy of always following expert-provided actions from all states will achieve a satisfactorily high return on the MDP problem above. However, such a policy would be impractical: we assume that the expert has limited bandwidth or availability, thus it is desirable for the agent to balance the goal of achieving high returns with the goal of not overloading the expert with requests.

Our method aims to determine when the agent should request expert assistance while making the most of a limited number of such calls available during deployment. Furthermore, we would like to limit (or eliminate) the number of calls made to the expert during the training phase.

To achieve this goal, our method leverages the uncertainty of the outcomes of the agent during exploration. Specifically, we use the variance of the return from a given state as a measure of uncertainty. Intuitively, a state with high variance of the return is one from which the agent, acting alone, could achieve a range of outcomes, ranging from very poor to very effective. We posit that these are the best states in which to request assistance. In contrast, states where the variance is low are those in which the agent is certain of the outcome of the task, and there are fewer ways in which external assistance can be of help.

Most RL algorithms work by estimating the expected return from a given state; this estimate is continuously refined during training time. We modify this training procedure also to maintain and refine an estimate of $\text{Var}_\pi(R)$, the variance of the returns under the learned policy, as an indicator of the agent’s uncertainty. At test time, our agent can use this

estimate of variance to decide when to request help. We detail these processes next.

A. Estimation for the variance of return

The expected return from a given state is encapsulated by the state value functions, which are explicitly relied on by most RL algorithms. However, while most RL algorithms are not concerned with the variance of the return, we would like to compute and maintain similar estimates for it during training. By definition, the variance of the return is defined by:

$$\text{Var}(R) = \mathbb{E}[R^2] - \mathbb{E}[R]^2$$

The expected return of taking action a_t in state s_t is given by the Q-function under policy π :

$$\mathbb{E}[R] = Q^\pi(s_t, a_t) \quad (1)$$

Since the transition function is not available, it is usually approximated via Bellman update:

$$Q^\pi(s_t, a_t) \leftarrow (1 - \alpha)Q^\pi(s_t, a_t) + \alpha(r(s_t, a_t) + \gamma \max_{a_{t+1} \in A} Q(a_{t+1}, s_{t+1})) \quad (2)$$

where α is a constant that is less than 1.

As shown in [24], the second moment of return $\mathbb{E}[R^2]$, which we call $M^\pi(s_t, a_t)$, can be approximated using a sampling-based method by:

$$M^\pi(s_t, a_t) = \mathbb{E}[R^2] = r^2(s_t, a_t) + 2\gamma \sum_{s_{t+1} \in S} p(s_{t+1}|s_t, a_t) Q^\pi(s_{t+1}, a_{t+1}) + \gamma^2 \sum_{s_{t+1} \in S} p(s_{t+1}|s_t, a_t) M^\pi(s_{t+1}, a_{t+1}) \quad (3)$$

Tamar et al. suggest that the solution to Eq.3 is unique under the assumption that the policy is proper. A policy is proper if a terminal state can be achieved from any initial state with a positive probability. Tamar et al. show that we can apply variants of TD(0) and LSTD(λ) with linear function approximation.

We observe that the formula of M^π satisfies a Bellman-like linear relation. Hence, we estimate M^π using a similar formula to Eq.2 with sampled data:

$$M^\pi(s_t, a_t) \leftarrow (1 - \alpha)M^\pi(s_t, a_t) + \alpha M'(s_t, a_t) \quad (4)$$

where

$$M'(s_t, a_t) = r^2(s_t, a_t) + 2\gamma Q^\pi(s_{t+1}, a_{t+1}) + \gamma^2 M(s_{t+1}, a_{t+1}) \quad (5)$$

Here, $a_{t+1} = \text{argmax}_{a_{t+1} \in A} Q(s_{t+1}, a_{t+1})$, since we assume a greedy agent that will take the actions that lead to max expected return during deployment.

Finally, combining M and Q , we can approximate the variance of returns in a state by:

$$\text{Var}(R) = M(s_t, a_t) - Q^2(s_t, a_t) \quad (6)$$

Algorithm 1 HULA Training

- 1: **while** returns have not converged **do**
 - 2: Collect H trajectories using the current control policy π with an MDP
 - 3: Update estimator of the Q-function Q_{θ_1} using Eq. (2)
 - 4: Update estimator of the M-function M_{θ_2} using Eq. (5)
 - 5: With updated M and Q , compute state-dependent variance using Eq. (6)
 - 6: **end while**
-

Hence, the variance of the return given a state and action pair only depends on the Q function as well as the M (second function), which in turn depends on Q and can be updated via a Bellman-like recursion.

B. HULA: Complete method

We are now ready to integrate the variance estimation into a complete learning algorithm. We use algorithms from the Q-learning family, as these already build an explicit estimate of the Q function at train time, and also use a greedy policy w.r.t. Q at run-time.

During training, in addition to the function estimator for the Q function (referred to as Q_{θ_1}), we build and update an estimator for M (referred to as Q_{θ_2}). This estimator can be of the same type as used for Q : for tabular Q-learning, we can use a table, while for Deep Q-Networks (DQN) [25] we use a deep neural network. At every iteration, we use Eqs. (4-5) to update the estimator for M . In the case of tabular Q-learning, Eq. (4) can be used directly, while in the case of DQN it can be transformed in a loss function analogous to the one used for Q . The complete procedure is shown in Alg. 1.

Finally, during deployment, the agent can use the trained estimators for the Q and M functions to compute the variance of the return for the current state. If this variance exceeds a threshold, the agent will request assistance and execute the action provided by the expert. Otherwise, the agent will execute the action prescribed by its own policy (in this case, a greedy selection based on the trained Q -function).

We note a key feature of our method: *it does not require the presence of an expert during train time*. The only change at train time compared to traditional, expert-less Q-learning is the addition of the function estimator for M , which will be used at run-time to help provide a variance estimator. In turn, the variance estimator is used to decide when to request assistance from the expert, which is only needed at deployment.

IV. EVALUATION

A. Environments

We evaluate HULA in a discrete navigation scenario, where the agent must reach a goal while avoiding obstacles and/or traps. Furthermore, we test our approach both on problems where the agent has exact knowledge of its current state (which is tractable via tabular Q-learning) and problems where the agent only has access to sensor data of limited

Algorithm 2 HULA Deployment

```
1: while not done do
2:   Compute variance  $v_t$  in current state using Eq. (6)
3:   if  $v_t \geq \epsilon$  then
4:     Execute action  $a_{exp}(s_t)$  provided by expert
5:   else
6:     Execute action from own policy  $\pi(a_t|s_t)$ 
7:   end if
8: end while
```

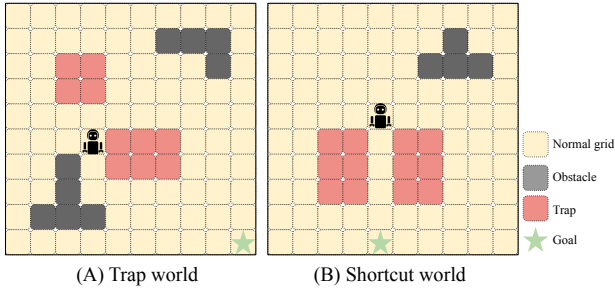


Fig. 2: **Fully-observable grid worlds.** The trap world (A) requires the agent to navigate around traps. In the shortcut world (B), the agent can leverage the expert to use the shortcut alley to achieve higher returns.

range, creating ambiguity (which requires more powerful function estimators, and the use of DQN).

1) *Fully-observable MDPs:* These are discrete navigation environments where the agent is provided its exact location as observation. In addition to the goal cell, the environment also contains obstacles and traps: colliding with a trap would terminate the episode, while colliding with an obstacle results in a failure of action, and the agent would stay in its original cell. The agent also receives a small penalty for each step it takes. Available actions for an agent are moving up, down, left, and right, and the observations are the coordinates (x, y) of the agent’s location. We use the maps shown in Fig. 2, and train and test individually on each map.

For this simple class of problems, an unsophisticated agent can easily learn to act optimally on its own. However, we introduce uncertainty in the form of a stochastic transition

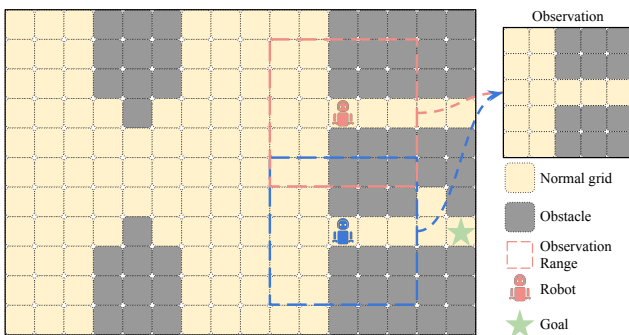


Fig. 3: **Partially-observable grid world.** In this environment, an agent may not localize itself by observing a surrounding 5x5 region. For example, the blue state and the red state result in identical observations.

function: at every step, after selecting an action, the agent moves in the desired direction with probability ψ , and moves in a random direction with probability $1 - \psi$ (in practice, we use $\psi = 0.45$). We think of this setting as a “slippery world”. In real-world scenarios, robots can experience similar action errors due to motor failures. Thus, even with an optimal policy, the robot may not reach its goal.

To help, we introduce an expert: when called, the expert provides the optimal action in the direction of the goal, and the action provided by the expert is not subject to transition function stochasticity. In this setting, the expert can thus be thought of as an “action corrector” with a better understanding of the environment dynamics, and whose actions always produce the expected results. Intuitively, we would expect an agent to call the expert in high-risk situations where it needs to avoid a wrong move, such as very close to traps. In this experiment, we test HULA on two maps: 1. trap world that requires an agent to navigate around traps with an expert; 2. shortcut grid where the agent can leverage the expert to take the shortcut for higher returns.

2) *Partially-observable MDPs:* To test how our algorithm performs in more complex environments, we extend to a case of partially observable MDPs where the agent does not have complete information about the system. Instead of observing state information as the (x, y) position in the grid, the observation now only includes a finite patch of grid cells around the agent’s current location (we use 5x5 grids in our implementation).

As shown in Figure 3, due to partial observability, it is not always possible for the agent to uniquely identify its own location in the map from observations. In such ambiguous regions, the same action taken based on identical observations can lead to utterly different results. However, other parts of the map are uniquely identifiable based on sensor observations, and thus an autonomous agent can always select the optimal action.

In this experiment, the expert observes the full state and always provides optimal actions to the agent. Intuitively, we would expect the agent to make use of the expert when traversing ambiguous areas of the map.

B. Evaluation Approach

The main evaluation metric for our approach focuses on its ability to make effective use of the expert: we would like to see how performance (measured as average episodic return) changes as a function of the number of expert calls made during deployment. We can measure this performance by varying the value of the variance threshold ϵ used in Alg. 2: for large values of ϵ , the agent will never make use of the expert; conversely, if ϵ is very small, the agent will call the expert in every state. By sweeping the value of ϵ between these extremes, we obtain more or less autonomous agents, and can plot performance as a function of the number of expert calls that result in each case.

As a baseline, we use a standard RL approach that simply integrates the expert into the training procedure. We refer to this baseline as the *penalty-based agent*. Specifically, the

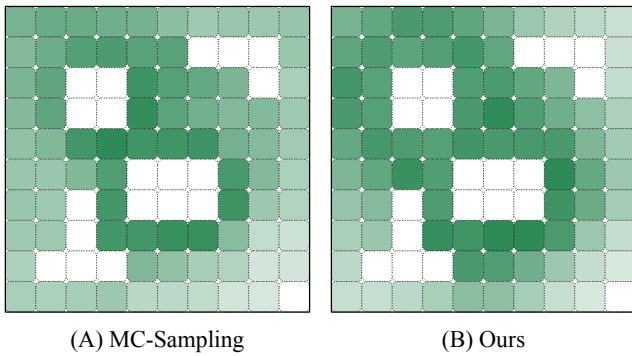


Fig. 4: **A learned variance map for the trap world environment.** (A) shows the return variance computed by Monte-Carlo Sampling with the same policy. (B) shows the estimated variance by our method (darker colors represent higher variances).

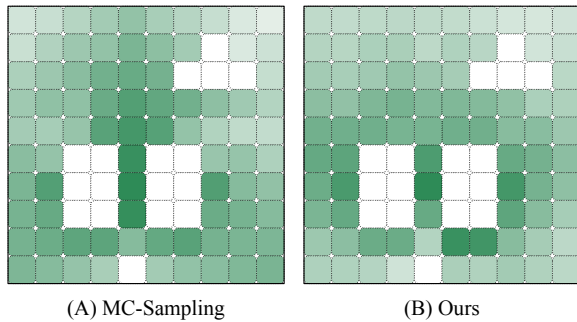


Fig. 5: Comparisons between a variance map learned by HULA and the ground truth for the shortcut world.

penalty-based agent treats calling expert as an extra action a_{call} that can be called alongside the other actions at both training and deployment time.

However, since the expert is optimal, the penalty-based agent will learn to always call for help. To learn a nontrivial policy that does not call the expert excessively, we employ a penalty of calling an expert in the reward function at train time: $r'(s, a) = r(s, a) + c$, where $c < 0$ is a penalty assessed only when calling an expert. By training this method with varying values of c , we again obtain agents that are more or less autonomous: the penalty-based agents trained with high c will call the expert less often, while those trained with low c will call more often. Again, we sweep the value of c between these extremes, and plot performance as a function of the number of expert calls that result in each case. However, for a fair comparison against HULA, we do not assess the expert penalty at deployment time.

V. RESULTS

A. Uncertainty Estimation

The first question we would like to answer is *whether HULA successfully captures the uncertainty of the agent and allocates the expert calls efficiently.* We compare the learned variances by our method with the ground truth variance of returns by Monte Carlo sampling the environment with the same policy without the assistance of an expert.

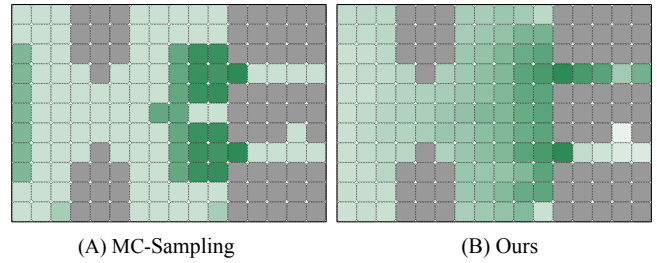


Fig. 6: Comparisons between a variance map learned by HULA and the ground truth for the partially-observable environment.

TABLE I: **Quantitative results of the expert call efficiency.** Accuracy of top- N variance states estimation. PO represents the partially-observable experiments.

Experiment	Trap world	Shortcut world	PO
Top-5	0.8	0.2	0.8
Top-10	0.8	0.6	0.7

As shown in Figure 4, in the fully-observable experiment, HULA successfully estimates a variance map that is similar to the ground truth. Specifically, it is uncertain about its outcome in states close to traps and confident when it is far from traps.

In the partially observable case, HULA identifies ambiguous states when the agent cannot localize itself in the map. In these states, the agent fails to produce an action faithfully in that the same action taken can lead to different results (e.g., hitting the boundary or reaching the goal). Our results align with the ground-truth variance map, which also has high uncertainty in a similar region.

To further investigate HULA’s efficiency in expert allocation, we compare the states with the highest N variances with those in the ground truth variance map. These states are usually where the agent calls the expert, especially when the agent has a budget of N expert calls.

Our results show that HULA can estimate the uncertainty of an RL agent accurately and allocate its expert call based on the return variance efficiently. As shown in Table I, in trap world, 80% of the top-5 and top-10 variance states estimated by HULA are also the top-5 and 10 accordingly in the ground-truth variance. In the shortcut world, although our method only correctly estimates 1 of 5 highest variance states, its performance is higher and reaches 60% in the top-10 case. This indicates that it allocates expert calls efficiently if given more budget. In the partially-observable experiment, our agent also recovers most of the high variance states in both the top-5 and top-10 evaluation.

B. Task Performance

In all experiments, our results show that HULA agents achieve higher performance when requesting help from experts compared to the ones that does not use an expert. In the trap world, as shown in Fig. 7, HULA achieves similar performance as the penalty-based agent while using the same number of expert calls. Interestingly, in the shortcut world, we find that the penalty-based method has a higher average return

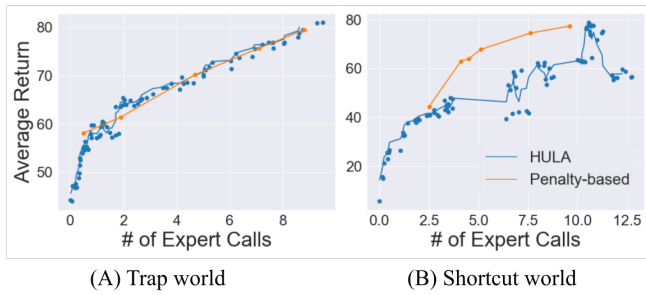


Fig. 7: # of expert calls vs Average returns for the fully-observable experiments. Curves are generated with the rolling mean method (window size of 4) for visualization purposes.

than HULA. This is because our method does not evaluate the effect of calling an expert during training. As shown in Fig. 8, the penalty-based agent learns to move towards states that ask for help from an expert since it knows it will call an expert in a future state, whereas HULA navigates around the uncertain states to avoid them. However, in the highest-variance states (e.g., states between traps), both agents are able to call the expert and complete the task.

In this experiment, we find that HULA is robust to different expert call budgets compared to the penalty-based agent. For example, in the shortcut world, the penalty-based agent cannot learn a policy that uses 2 expert calls to complete the task even if we perform hyper-parameter sweeping with the expert penalty c . This can be caused by the high expert penalty that modifies the original reward structure and hinders the learning of the task. In contrast, our method can flexibly incorporate different expert call budgets and maintain reasonable performance.

In the partially-observed grid world, the HULA agent outperforms the penalty-based agent when the allowed expert call is limited in the range of [4, 8]. When both agents are given enough expert assistance, they achieve similar performance. This indicates that HULA efficiently utilizes the expert to localize itself and improve the task performance. Surprisingly, our method does not have better performance than the baseline when only less than 2 expert calls are allowed. This is because HULA has higher variance in the crucial states (see Fig.6). With only 2 expert calls, it decide not to acquire experts' assistant in earlier steps. In contrast, the penalty-based agent learns to make some progress with the experts during training. However, limited by the number of expert calls, it also fails to achieve the goal.

An important feature of our method is that it does not introduce extra complexity in training an RL agent. Practically, we stop training when the Q function converges. Therefore, the training efficiency is similar to its underlying RL algorithm. In contrast, the penalty-based agent requires access to an expert during training time. Our experiments show that training each penalty-based agent for the fully-observable environment results in about 70000 expert calls. The requirement of an expert during train time limits its application on learning policies that interact with humans.

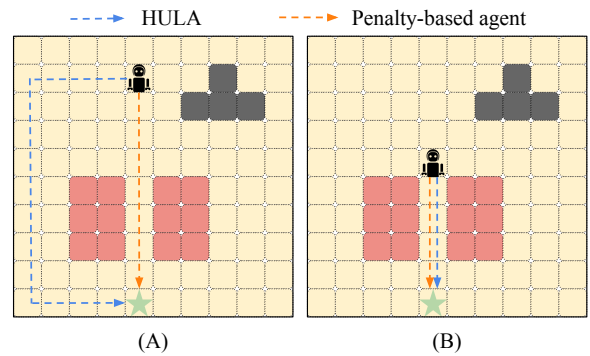


Fig. 8: Example rollouts of the HULA agent and the penalty-based agent in the shortcut world.

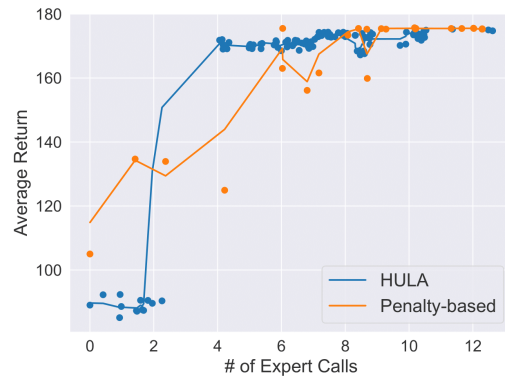


Fig. 9: # of expert calls vs Average returns for the partially-observable experiments. Curves are generated with the rolling mean method (window size of 4) for visualization purposes.

VI. CONCLUSION

We introduced HULA, a method to learn human-in-the-loop policies by estimating an RL agent's uncertainty. To capture the uncertainty, we directly estimate the variances of the returns using sampled data during the training of an RL agent. Our experimental results demonstrate that HULA can capture an RL agent's uncertainty and use it to request assistance from experts to achieve high task performance. An important feature of our method is that it does not require the presence of an expert during training.

The key limitation of our approach is that HULA requires the agent to have reasonably good explorations in the reward space. Such an agent can be hard to train in hard exploration problems, for example, in the sparse reward setting. However, this problem can be alleviated by applying advanced exploration strategies, for example, employing intrinsic motivation when learning a policy [26].

We envision that our approach can be applied to more complex problems. For example, we plan to extend HULA to continuous RL algorithms (e.g. DDPG [27]) and learn HitL policies to solve continuous control problems. We also would like to explore the direction of learning uncertain agents in other domains, such as language-guided navigation. Another extension to HULA is using it with offline reinforcement learning and utilizing the diversity of offline datasets to estimate an agent's variances.

REFERENCES

- [1] Oliver Kroemer, Scott Niekum, and George Dimitri Konidaris. “A Review of Robot Learning for Manipulation: Challenges, Representations, and Algorithms”. In: *CoRR* abs/1907.03146 (2019). arXiv: 1907.03146. URL: <http://arxiv.org/abs/1907.03146>.
- [2] Christian Daniel, Gerhard Neumann, Oliver Kroemer, and Jan Peters. “Hierarchical Relative Entropy Policy Search”. In: *Journal of Machine Learning Research* 17.93 (2016), pp. 1–50. URL: <http://jmlr.org/papers/v17/15-188.html>.
- [3] Jan Peters and Stefan Schaal. “Reinforcement learning of motor skills with policy gradients”. In: *Neural Networks* 21 (June 2008), pp. 682–. DOI: 10.1016/j.neunet.2008.02.003.
- [4] E. Theodorou, J. Buchli, and S. Schaal. “Reinforcement learning of motor skills in high dimensions: A path integral approach”. In: *2010 IEEE International Conference on Robotics and Automation* (2010), pp. 2397–2403.
- [5] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. “Power to the People: The Role of Humans in Interactive Machine Learning”. In: *AI Magazine* 35.4 (Dec. 2014), pp. 105–120. DOI: 10.1609/aimag.v35i4.2513. URL: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2513>.
- [6] Anis Najjar and Mohamed Chetouani. “Reinforcement learning with human advice. A survey”. In: *CoRR* abs/2005.11016 (2020). arXiv: 2005.11016. URL: <https://arxiv.org/abs/2005.11016>.
- [7] Brenna D. Argall, Brett Browning, and Manuela Veloso. “Learning robot motion control with demonstration and advice-operators”. In: *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2008, pp. 399–404. DOI: 10.1109/IROS.2008.4651020.
- [8] Jonathan Spencer, Sanjiban Choudhury, Matthew Barnes, Matthew Schmittle, Mung Chiang, Peter Ramadge, and Sidd Srinivasa. “Expert Intervention Learning”. In: *Autonomous Robots* 46.1 (Jan. 2022), pp. 99–113. ISSN: 1573-7527. DOI: 10.1007/s10514-021-10006-9. URL: <https://doi.org/10.1007/s10514-021-10006-9>.
- [9] W. Bradley Knox and Peter Stone. “Interactively Shaping Agents via Human Reinforcement: The TAMER Framework”. In: *The Fifth International Conference on Knowledge Capture*. Sept. 2009.
- [10] Carlos Celemin and Javier Ruiz-del-Solar. “An Interactive Framework for Learning Continuous Actions Policies Based on Corrective Feedback”. In: *Journal of Intelligent & Robotic Systems* 95.1 (July 2019), pp. 77–97. ISSN: 1573-0409. DOI: 10.1007/s10846-018-0839-z. URL: <https://doi.org/10.1007/s10846-018-0839-z>.
- [11] Riku Arakawa, Sosuke Kobayashi, Yuya Unno, Yuta Tsuboi, and Shin-ichi Maeda. “Dqn-tamer: Human-in-the-loop reinforcement learning with intractable feedback”. In: *arXiv preprint arXiv:1810.11748* (2018).
- [12] Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popović. “Where to add actions in human-in-the-loop reinforcement learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.
- [13] Annie Xie, Fahim Tajwar, Archit Sharma, and Chelsea Finn. “When to Ask for Help: Proactive Interventions in Autonomous Reinforcement Learning”. In: *arXiv preprint arXiv:2210.10765* (2022).
- [14] Jingda Wu, Zhiyu Huang, Chao Huang, Zhongxu Hu, Peng Hang, Yang Xing, and Chen Lv. “Human-in-the-loop deep reinforcement learning with application to autonomous driving”. In: *arXiv preprint arXiv:2104.07246* (2021).
- [15] Peter Barnett, Rachel Freedman, Justin Svegliato, and Stuart Russell. *Active Reward Learning from Multiple Teachers*. 2023. arXiv: 2303.00894 [cs.LG].
- [16] Max Jaderberg, Volodymyr Mnih, Wojciech M. Czarnecki, Tom Schaul, Joel Z. Leibo, David Silver, and Koray Kavukcuoglu. “Reinforcement Learning with Unsupervised Auxiliary Tasks”. In: *ArXiv* abs/1611.05397 (2016).
- [17] Shakir Mohamed and Danilo Jimenez Rezende. “Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning”. In: *NIPS*. 2015.
- [18] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. “Curiosity-Driven Exploration by Self-Supervised Prediction”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2017), pp. 488–489.
- [19] Gregory Kahn, Adam Villafior, Vitchyr Pong, Pieter Abbeel, and Sergey Levine. “Uncertainty-aware reinforcement learning for collision avoidance”. In: *arXiv preprint arXiv:1702.01182* (2017).
- [20] Carl-Johan Hoel, Krister Wolff, and Leo Laine. “Ensemble quantile networks: Uncertainty-aware reinforcement learning with applications in autonomous driving”. In: *arXiv preprint arXiv:2105.10266* (2021).
- [21] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. “Implicit quantile networks for distributional reinforcement learning”. In: *International conference on machine learning*. PMLR. 2018, pp. 1096–1105.
- [22] Ian Osband, John Aslanides, and Albin Cassirer. “Randomized prior functions for deep reinforcement learning”. In: *Advances in Neural Information Processing Systems* 31 (2018).
- [23] Felipe Leno Da Silva, Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. “Uncertainty-aware action advising for deep reinforcement learning agents”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 04. 2020, pp. 5792–5799.
- [24] Aviv Tamar, Dotan Di Castro, and Shie Mannor. “Learning the Variance of the Reward-To-Go”. In: *Journal of Machine Learning Research* 17.13 (2016), pp. 1–36. URL: <http://jmlr.org/papers/v17/14-335.html>.
- [25] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. “Playing Atari with Deep Reinforcement Learning”. In: *ArXiv* abs/1312.5602 (2013).
- [26] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. “Diversity is All You Need: Learning Skills without a Reward Function”. In: *CoRR* abs/1802.06070 (2018). arXiv: 1802.06070. URL: <http://arxiv.org/abs/1802.06070>.
- [27] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Manfred Otto Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. “Continuous control with deep reinforcement learning”. In: *CoRR* abs/1509.02971 (2015).