Human in the Lifelong Reinforcement Learning Loop

Abstract—**Deploying embodied agents in real-world settings brings safety, adaptability, and costs challenges. As Reinforcement Learning (RL) in robotics remains expensive to build in such scenarios, generic decision models are often trained at a high scale of the simulation. Then, transferring to a real-world environment becomes the starting point of a lifelong adaptation learning procedure. For a realistic scenario of service robotics, adaptation with the help of human local experts is one major challenge for the large adoption of such a paradigm. In this paper, we introduce a novel learning paradigm that improves adaptation efficiency to real-environment thanks to an original integration of human expert feedback that iteratively improves the agent behavior toward an expected use. Our approach involves a computational mechanism of influence for experts to alter the behavior of the agent in a defined space that does not need to be aligned to the action space of the targeted environment. This approach can be distinguished from the state-of-the-art approaches where the human is solely defined as Oracle over the targeted action space of the considered environment. In this proposition, the human expert, also called a coach, can sequentially assess the capabilities of the current agent and the constraints of the environment on the go while defining influence signal altering the agent actions. We illustrate our approach in the task of autonomous localization and mapping and show initial results using the AI-Habitat simulator.**

I. Introduction

Embodied agent deployments in public areas emphasize the issue of robots adaptation to ever-changing environments. As agent, we refer to the sequential decision making model that defines, at each step, the action of the robot. Autonomous robots raise the question of the cost of deployment, the risk of sharing the physical environment with humans and the limited guarantee that robots behavior can provide. Another major concern is the ability to adapt the robots either to a new place or within a same environment to evolving practices and goals. This situation brings the need for lifelong learning for the robot and this concern is shared by researchers from diverse research communities studying the role of humans in the large variety of machine learning paradigms and associated protocols [6]. This general concern is also recently labeled under the term human-centered AI [11]. In this paper, we introduce a novel approach to adapt sequential decision model giving humans the ability to adapt agents for deployment in real-world settings. Interacting with human in the loop during training has shown significant advantages in numerous real world situations. First, it allows to improve the data efficiency of the learning. Second, it makes possible the introduction of robots in new environments in a safe manner [15]. In the following of this paper, we describe a new formalization of human in the loop where human experts influence the behavior of a sequential decision agent in order to adapt it to fit their needs. This approach can be distinguished from the state of the art approaches of human-in-the-loop where the human is solely defined as Oracle over the targeted action space of the considered environment. We assume that one or several human experts can coach an agent introduced to real or facing a real environment change. As human experts, we refer here to humans that are knowledgeable regarding the task to be done, the environment where the agent is evolving and associated risks. No requirements of expertise in algorithms, nor models of sequential decision making are assumed. We evidence the technological feasibility of this new learning paradigm with an experiment. Finally, we illustrate our approach in the task of autonomous localization and mapping and show initial results over state-of-the-art methods, using the AI-Habitat simulator.

II. Human in the loop in Reinforcement Learning

Besides the performances of modern sequential decision learning approaches since Deep-Q learning [8], [10], recent research works addressing real-world situations where humans interact during such a learning process can be mentioned.

First, DQN-TAMER [2] uses face recognition of a supervising human for helping inducing a policy for maze navigation. As a possible extension, this study shows a good example of the necessity for humans to help the learning agent in a natural language. Nonetheless, while face recognition approach is an interesting reward signal to use, it is limited to very simple environments and an action frequency which is low. As an alternative, we propose to introduce a task-specific signal of supervision that will directly bias the behavioral policy.

In more complex environments and tasks, it is often needed to identify where to interact in Human-in-the-Loop learning process. As an example, [9] emphasis the importance of requiring user feedback when the expected impact of this information is high. Indeed, It is desirable to directly interact with the learning agent if the situation requires it. In this work, we get inspired by this approach but we propose to improve it: letting the user decide where and how in the agent environment he considers his advice to get the more impact.

In the specific context of autonomous driving, this question has been addressed by the live supervision of a human driver [14]. It seems natural to supervise autonomous driving with a driver live corrections as reinforcement signal. In this context, a shared environment forms a convenient situation to align human local preferences with agent decision model as the situation involves a common usage of the steering wheel which correspond to a necessary alignment of the action space of the

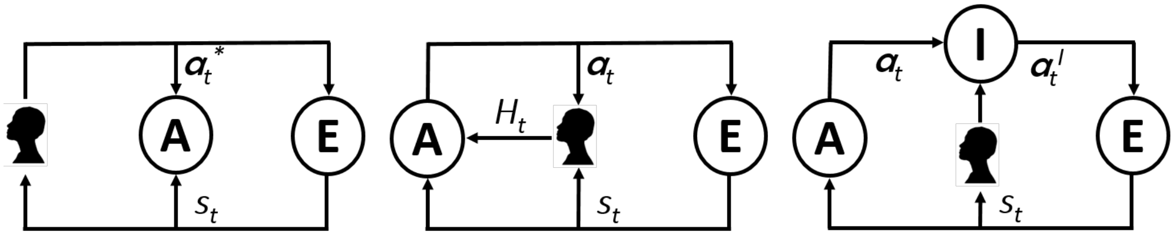

Fig. 1. From (left) classic imitation learning scheme to (middle) Evaluative feedback to (right) our proposed coaching approach of human-in-the-loop sequential learning. The black head is the human expert, also named coach, E is the environment, A is the agent. In this last approach, I is the influence module that transforms the influence signal produced by the coach and the agent action into the influenced action

agent and the human expert. In this work, we reproduce this ideal scheme when the task cannot be supervised in real-time by a driver, but by regular external observations of a task by experts and without such action space alignment constraint. Agent-Agnostic Human-in-the-Loop Reinforcement Learning [1] is an interesting overview of the various algorithmic possibilities offered by the two possible interventions on this learning procedure: altering the action, or change the reward. Action alteration aims at reducing the exploration/exploitation necessity of learning by leveraging human knowledge about the task at hand.

Finally, reward shaping [5] is another approach of learning efficiency improvement that is available at learning time. On the other hand, action alteration is also available at exploitation time.

III. Influence for human in the loop of sequential decision

Figure 1 describes our proposed method using the notation introduced in [15] and details how we differentiate from it. In comparison to the state of the art methods which mainly consist in either directly fine-tuning the actions of the agent or to alter its reward function from a feedback provided by the human trainer, our method introduces a conversion step that input both the current action of the agent and an influence signal provided by the trainer and output a influenced action to execute into the environment.

So, given a trained sequential decision policy, our approach is decomposed in two steps. First, the outputs of the policy are altered with local constraints, called influence, specified by the human experts and conditioned by the current action and observed state of the environment. This first step involves is performed iteratively by a human expert improving the adaptation of the system by continuously monitoring the impact of the influence on the robot behavior in relation with the expected behavior. In this first adaptation step, where the agent's policy remains unchanged, can degrades the performance of the model. The second step integrates the established influence into the actual model in two possible manners, model distillation or reward shaping.

As an example, in the context of autonomous navigation, influence can be defined with a direction and an associated strength. Influence can be defined as relative or absolute. On the first hand, an influence is defined as absolute as a preferential distribution that is computed to favor the action going in the desired direction, given the current pose of the agent, e.g. west orientation. On the other hand, an influence is called relative when the computed distribution favors the influence direction, e.g. right orientation. The strength of an influence amplifies or smooths the action distribution. Then, local amplification, depending on the distance between the agent and the influence location, can be applied too. When computed for each state, this preference is applied to the output of the policy, e.g. using dot-product before selecting the decision.

Formally, a two sequential steps process leverages this influence signal.

A. Influence for action alteration

First, the agent's behavior is constrained using $\mathbf{B}(a, s, r, \psi) = \pi(a, s, r) * p + \epsilon(p)$, where p is the computed preferential distribution of action, knowing the influence and the current state : $p = \mathbf{P}(\psi, s)$, and ψ is the influence provided by humans.

Finding an adaptation scheme requires human expertise since human criteria and constraints can extend the perception capacities of the agent. The proposed influence mechanism is external to the model as it post-processes the decision logits to alter and improve the decisions. The alteration of actions is first meant for securing the behavior of the policy with respect to the task at hand before adapting it.

B. Learning from influence

The preferences designed by experts are used to fine-tuned the original policy. The reward shaping function is defined as follows: $\mathbf{R}(a, s, r, \psi) = r + \phi(a, p)$ with is the reward granted for having chosen the a action under the preference distribution computed for the state s under the influence Ẇe choose to define the reward shaping function : $\phi\ (a, p) = \epsilon(max(p) - p[a])$

In this method, we propose to loosely uncouple the exploration of an adaptation executing in a given environment from its learning as fine-tuning of the current policy. Indeed, human feedback can be built from the observation of the exploration they suggest to run. Once the explored influence is safe and confirms to adapt to the new context, it can reinforce the

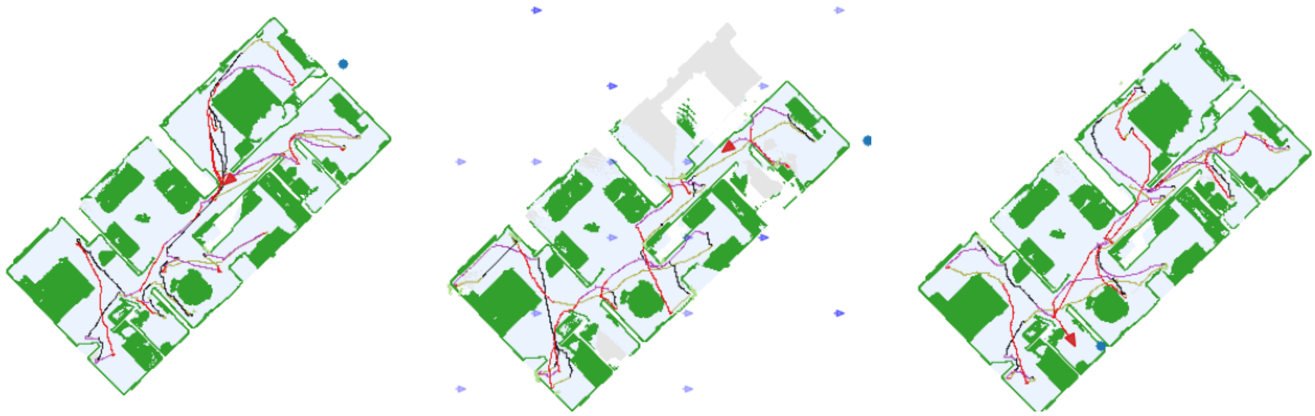

Fig. 2. After defining ops to measure the collision risk. $nb(overlap\_pixels)/exploration\_rate$. Original mapping and navigation (**Left** 88 ops) under right-side influence (**Middle** 24 ops). **Right:** Resulting influenced policy

task using state of the art approaches of behavior cloning and model distillation, defined using the constraints given by the human observer during the first step of the proposed adaptation approach.

IV. Experiments

We illustrate our approach using a scenario of autonomous mapping in an indoor environment. We use the 40 first scenes of Gibson validation dataset of autonomous navigation task over the AI-Habitat framework [12]. We assume a neural network trained for Simultaneous Localization and Mapper sharing the space with a fleet of delivery robots to update them with map modifications. When blocked in the traffic, the mapper maps false new obstacles. To avoid this, it would desirable that both the mapper and robot adopt a right-side navigation behavior. Our study focuses on the mapper navigation task, the Active SLAM mapper [4] under a right-side walk influence within the AI-habitat validation dataset. As for now, the computation of preference from influence is deterministic.

Task compliance: During this first step of the method, by adding a set of local influences to go right, humans can first secure the SLAM model from unexpected collisions. While the safe mapper maps less completely, it navigates more on the right of the corridors, lowering the risk of frontal collisions. Figure 2 illustrates robot trajectories where red is south, black is north, pale green is east and purple is west oriented. Overlapping of red and black is rare, as for purple and green. In the ideal appliance of the right-side influence, purple is upper than green and black is to the right of red. The resulting mapping also illustrates where grey areas correspond to unmap ones. Figure 3 details the collisions observed for the originals, influenced and fine-tuned policies.

Reinforced under influence: after the second step, the SLAM model acquires a safer behavior as illustrated in Figure 3. We verified that ops(native) ¿ ¿ ops(influenced) ¿˜ ops(reinforced) on most of the considered indoor scenes. As we reinforced on various scenes under the same generic scheme designed for our example scene, points of influence were unchanged so not ideally placed for all the scenes for some of them.

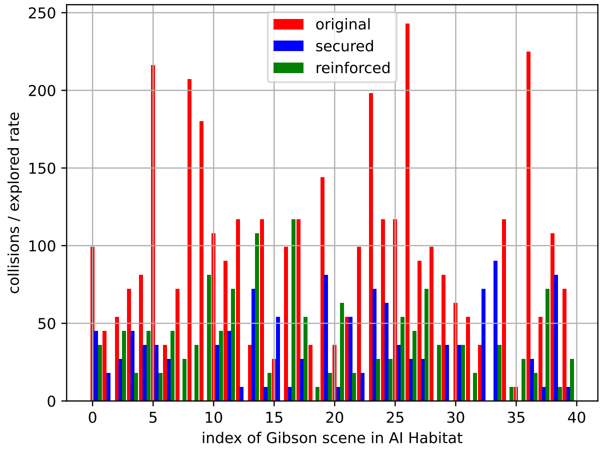

Fig. 3. Collision rates over 40 indoor environments for the task of mapping in original, secured and reinforced models, lower the better.

V. Conclusion and future work

In this paper, we introduce a new learning paradigm that considers a signal defined as influence, provided by human experts, as a reasonable proposition to answer several challenges regarding long-term human robot interaction. In particular, it improves the learning performances, which is critical to foster autonomous robot deployment in everyday environments. So far, in human in the loop for decision sequential making, the first attempts were to overcome technological limitations of algorithms and training on data by supplementing the learning protocol with human inputs in order to increase the performance of the algorithm. The shift initiated with our proposition is to evolve from a *targeted task*, that is known to be optimized, to a preferred behavior, that does not need to be fully anticipated by the human experts to be developed. Indeed, the posture of the human in the loop shifts from being an omniscient oracle to an actor involved in an iterative loop. This shift brings two valuable outcomes. First, it creates the context for human experts and embodied agents to jointly fine-tune an appropriate behavior. Maintaining the humans in the decision loop is a way to keep them in control which is known to be critical in achieving safety, trust and reliability [13]. Second, it provides a solution where human experts do not need knowledge and development skills about algorithms, or models of sequential decision making to act on improving how this technology can support them. This *blurring the line between developers and end-users* is identified as a critical point to foster wide adoption of AI based technology [3]. The initial framework presented in this contribution open new research questions that will guide future work. One relates to how human expert feedback will be collected. The feedback can rely on various modalities such as haptic, graphical user interface or natural conversations [2]. We foresee a diversity of human experts ranging from the workers involved in the supervision and maintenance of embodied agents to physically closed individuals to a robot executing its task. The influence results in the aggregation of the feedback of all human experts, as proposed in [7]. Another element to consider is the information that we will need to share with the human expert in order to provide efficient coaching . In fact, all the different coaches will have a representation of the behavior of the robot. In addition, sharing details on the task of the embodied agent, its perception of the environment, its current state might impact the feedback.

References

[1] David Abel, John Salvatier, Andreas Stuhlmüller, and Owain Evans. Agent-agnostic human-in-the-loop reinforcement learning. *arXiv preprint arXiv:1701.04079*, 2017.

[2] Riku Arakawa, Sosuke Kobayashi, Yuya Unno, Yuta Tsuboi, and Shin-ichi Maeda. Dqn-tamer: Human-in-the-loop reinforcement learning with intractable feedback. *arXiv preprint arXiv:1810.11748*, 2018.

[3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[4] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. *arXiv preprint arXiv:2004.05155*, 2020.

[5] Luíza C. Garaffa, Maik Basso, Andréa Aparecida Konzen, and Edison Pignaton de Freitas. Reinforcement learning for mobile robotics exploration: A survey. *IEEE transactions on neural networks and learning systems*, PP, 2021.

[6] Bahar Irfan, Aditi Ramachandran, Samuel Spaulding, Sinan Kalkan, German I Parisi, and Hatice Gunes. Lifelong learning and personalization in long-term human-robot interaction (leap-hri). In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 724–727, 2021.

[7] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. Webuildai: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–35, 2019.

[8] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[9] Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popović. Where to add actions in human-in-the-loop reinforcement learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[10] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[11] Michael Muller, Plamen Agelov, Shion Guha, Marina Kogan, Gina Neff, Nuria Oliver, Manuel Gomez Rodriguez, and Adrian Weller. Neurips 2021 workshop proposal: Human centered ai.

[12] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019.

[13] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[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. *arXiv preprint arXiv:2104.07246*, 2021.

[15] Ruohan Zhang, Faraz Torabi, L. Guan, Dana H. Ballard, and Peter Stone. Leveraging human guidance for deep reinforcement learning tasks. In *IJCAI*, 2019.