Integrating an Observer in Interactive Reinforcement Learning to Learn Legible Trajectories

Manuel Bied, Mohamed Chetouani

Integrating an Observer in Interactive Reinforcement Learning to Learn Legible Trajectories

Manuel Bied[1] and Mohamed Chetouani[1]

***Abstract*— An important aspect of Human-Robot-cooperation is that the robot is capable of clearly communicating its intentions to its human collaborator. This communication of intentions often requires the generation of legible motion trajectories. The concept of legible motion is usually not studied together with machine learning. Studying these fields together is an important step towards better Human-Robot cooperation. In this paper, we investigate interactive robot learning approaches with the aim of developing models that are able to generate legible motions by taking observer feedback into account. We explore how to integrate the observer feedback into a Reinforcement Learning (RL) framework. We do this by proposing three different observer algorithms as observer strategies in an interactive RL scheme and compare with one non-interactive RL algorithm as baseline. For the observer strategies we vary the method how the observer estimates how likely the agent is going for the target goal. We evaluate our approach on five environments and calculate the legibility of the learned trajectories. The results show that the legibility of the learned trajectories is significantly higher while integrating the feedback from the observer compared with a standard Q-Learning algorithm not using the observer feedback.**

I. INTRODUCTION

Humans and robots working together - so called Human-Robot cooperation - has recently become a popular area of research. This cooperation can allow robots and humans to accomplish more sophisticated tasks. When it comes to cooperation, one crucial difference between humans and other species is the capability to share goals and intentions [1]. In order to mimic these capabilities in Human-Robot collaboration, it is necessary to equip the robot with the capability of shared goals and intentions. One important aspect to achieve this intention sharing is that the robot understands what the human is doing, for example to predict human motions [2]. Another important aspect is to enrich the robot with behavior that can be well understood by humans. In general this problem requires the robot to behave more transparently or with explainability [3]–[5]. Depending on the task, this constraint on the robot's behavior requires the robot's motion trajectories to be either predictable or legible. While predictable and legible motion trajectories can correlate, they are *"fundamentally different and often contradictory properties of motion"* [6]. While legibility requires the knowledge of possible goals: *"Plan legibility reduces ambiguity over possible goals that might be achieved"* [5], predictability requires the knowledge of a goal/planning problem: *Plan predictability reduces ambiguity over possible plans, given a goal/planning problem* [5]. Both concepts have in common that some kind of observer of the actions of the robot exists. This observer reasons about the possible intentions, either possible goals or plans, of the robot. Aforementioned concepts are rarely studied in combination with machine learning. A suitable candidate to use machine learning in this context is reinforcement learning (RL) [7]. The basic idea of RL is that the robot, or more generally an agent, interacts with the environment and receives a reward from the environment. The goal of the agent is to chose its actions in a way that maximizes the received total reward. RL offers a computational framework to learn how choose these actions in an optimal way. The classical RL approach does not offer the possibility to add a human to the loop, depriving the framework of integrating valuable task knowledge from the human. This gap has already been addressed in research on interactive RL investigating different models of human-feedback [8]–[12]. Usually these approaches have the goal to speed up the learning process of the agent or enable it to

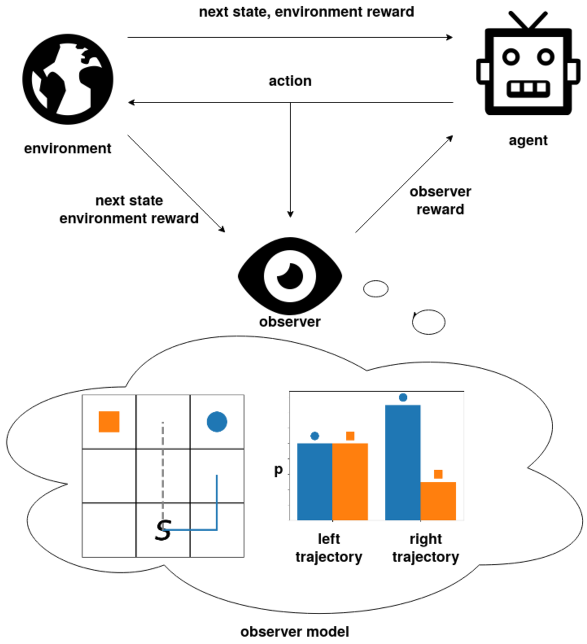

Fig. 1. Setup of the observer RL framework. As in a regular RL setting the agent interacts with the environment and receives a reward after each taken action. The observer gives an additional reward as feedback to the agent based on how well the observer can infer which of multiple possible goals the agent is targeting. This additional feedback results in the agent learning more legible trajectories.

[1] The authors are with Institut des Systèmes Intelligents et de Robotique, CNRS UMR 7222, Sorbonne Université, 75005 Paris, France {bied, chetouani}@isir.upmc.fr

find better solutions. To the best of our knowledge, none of the work on interactive RL explores the role of an observer that reasons about the intent of the agent in an interactive RL framework. However, integrating such an observer into interactive RL is an important step towards using RL for human-robot collaboration. In this paper we explore how to integrate observer feedback into RL algorithms to learn legible (motion) trajectories. We add an observer that gives feedback to the agent to improve the legibility of the learned policy. The interaction scheme of our proposed system is illustrated in Fig. 1.

II. Related Work

Terms like explicability [13], [14], legibility [6], transparency [15] and predictability [16] have become popular in recent research on artificial agents. These terms describe, depending on their definition, similar or contradicting concepts. A comprehensive overview of the different concepts is presented in [5]. All concepts have in common that they assume some kind of observer that tries to infer the intentions of the agent. This idea goes along with the concept of Theory of Mind [17], [18]. Theory of Mind is the capacity of attributing a mental state to other people i.e. to infer unobservable beliefs, desires and intentions and interpret actions in relations to these mind state. Research on Theory of Mind [19]–[22] suggests that humans interpret observed behaviors as goal-directed actions. Csibra and Gergeley [21] identify two types of inference: ”action-to-goal” and ”goal-to-action”. Based on this idea Dragan et al. [6] present a framework to quantify predictability (”goal-to-action” inference) and legibility (”action-to-goal” inference) of trajectories. In [23] they extend the work and use it to generate legible (motion) trajectories by introducing constrained legibility optimization. This framework is used in [24] to create legible pointing trajectories. As we use this legibility metric in our model, we will present it in III-B. Similar to the concept of legibility is the work on sensorimotor communication (SMC) [25], [26]. SMC uses the same channel as the to-be-executed action for communication. [25] proposes to use a signaling distribution of a trajectory in order to facilitate its recognition by another person. However, no other work uses an interactive RL scheme to learn more legible behavior. Some similar ideas can be found in [27], where the agents in a multi-agent RL setting integrate the intent of the other agents when calculating the optimal action, but do not express their intent themselves. [28] employs different approximate-inference Inverse Reinforcement Learning (IRL) variations to model how humans infer an agent's objective function and use an algorithmic teaching approach [29]–[31] to generate a set of environments to increase the probability of inferring the correct objective function. For each environment the optimal trajectory according to the objective function is shown, therefore it's not about comparing different trajectories in one environment like in our approach. The work of Ho et al. [32], [33] combine the idea of IRL and communication via the means of pedagogical reasoning [34]. [35] shows similar to this idea that there is a difference between people solving and teaching a sensorimotor task, furthermore that people perceive a significant higher portion of negative than positive demonstrations as informative. [36] uses the data of Ho et al. [33] and find that it's safer to assume a literal human even when people try to be pedagogic.

While no other work investigates the integration of legibility with classical RL algorithms, there exists a substantial amount of research on how to put a human in the loop of RL. Integrating evaluative feedback from humans into RL is sometimes called human-centered RL. A survey on human-centered RL topic is provided by [37]. One important model to mention is TAMER [38], [39]. TAMER directly models the human rewards and myopically learns from this model. TAMER itself is not a RL technique, however further work TAMER+RL [9], [40], [41] integrates TAMER with RL. Furthermore, other models that integrate human feedback into RL use concepts borrowed from TAMER (e.g. COACH [12], [42]) or extend it (e.g. ACTAMER [43]).

III. Integrating Observer Feedback on Legibility into Interactive RL

In this work we are interested in the combination of a RL system with an observer that reasons about the goals of the learner to increase the legibility of the learned trajectories. In order to achieve this we use a Markov Decision Process (MDP) in combination with Reward Shaping to model the learning problem. We add the observer to the equation by modeling the observer with different strategies to estimate how likely the agent is going for the target goal.

A. Interactive RL

The standard way of formalizing reinforcement learning problems is the use of a MDP. A MDP is defined as tuple $(S, A, \mathcal{T}, R, \gamma)$. S is the set of states (state-space), A is the set of actions (action-space), $\mathcal{T} : S \times A \times S \to P(s' \mid s, a)$ defines the state-transition probability function, with $P(s'|s, a)$ representing the probability that the agent transitions to state s' when taking the action a, $R : S \times A \times S \to \mathbb{R}$ defines the reward $r(s, a, s')$ that the agent receives when transitioning from state s to the new state s' while taking action a. $\gamma \to [0, 1]$ is the discount factor describing how much rewards for the recent decision are taking into account. The agent's objective is to maximize the cumulative received reward. A policy is a mapping from state to action $\pi : S \to A$. The value of taking an action a while being in state s and following the policy π can be described with the action-value function for policy π denoted as Q^{π}. A standard approach to solve problems formulated like this is Q-Learning. Q-Learning will also serve us as baseline to compare to. For Q-Learning we use a simple one-step Q-learning defined by:

$$\begin{aligned} Q(s_t, a_t) \leftarrow & \; Q(s_t, a_t) \\ & + \alpha[r_{t+1} + \gamma max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \end{aligned} \tag{1}$$

The step size α with $0 < \alpha < 1$ defines how strongly to move towards the new estimate at each iteration, the larger α, the larger the step towards the new estimate. The discount factor γ with $0 \leq \gamma \leq 1$ determines how strongly to take future rewards into account. When γ is 0 the agent will only consider current rewards and with increasing γ the agent takes future rewards more strongly into account. For action taking we use the exploration rate ϵ, i.e. with a probability of ϵ the agent takes a random agent and the reward maximizing action of the current policy otherwise.

We add the observer to the system by using Reward Shaping [44]. The original MDP reward is replaced by $R'(s,a)$ by adding the weighted reward from the observer $\hat{O}$ to it.

$$R'(s,a) = R(s,a) + \beta \cdot \hat{O}(s,a) \tag{2}$$

The reward from the environment and the reward from the observer are of different nature and can, in general, differ in scale. The weighting factor β can be used to accommodate for this fact. We will compare different algorithms for $\hat{O}$ to model different observer strategies. These algorithms will be presented in III-C. While other (more sophisticated) methods like Policy Shaping and Value Shaping [9], [10], [40], [45] to integrate Human feedback into the classical RL formulation exist, Reward Shaping will suffice as proof of concept for the feasibility of our approach. Ng et al. [44] describe the necessary requirements for Reward Shaping to preserve the optimal policy, if these requirements are not met positive-reward cycles can occur. Note that the way we are employing reward shaping does not meet these requirements.

B. Legibility

In order to formally evaluate the legibility $\lambda(\xi)$ of a trajectory ξ, we use the legibility metric proposed by [6]. Following this line of work the observer needs to be able to confidently infer the correct (=the target) goal g^* after only observing a part of the whole trajectory to the goal $\xi_{s_0 \to s_t}$ starting at s_0 and ending at the intermediate point s_t. The trajectory is more legible the faster this confident inference happens. Imagine an observer watching an agent acting in the environment shown in Fig. 1 in the observer model part. The observer tries now to infer as fast as possible for which goal the agent is going for. The right trajectory (solid blue line) is more legible than the left trajectory (dashed grey line), because for the right trajectory it seems more likely that the agent is going for the target goal on the right. For the left trajectory it is still not clear for which goal the agent is going for, the next step could either be to the left or to the right. Fig. 2 illustrates the concept of legibility in a discrete environment. The agent is aiming for the goal to the right (blue circle). There is an alternative goal on the left side (orange square). The more the trajectories go to the right side the higher is the resulting legibility. We will use this environment in our first experiment and refer to it as environment 1. The described properties of legibility are captured by the following equation [6]:

$$\lambda(\xi) = \frac{\int P(g^*|\xi_{s_0 \to s_t}) f(t) dt}{\int f(t) dt} \tag{3}$$

We integrate over the probability to infer the target goal given the current trajectory $P(g^*|\xi_{s_0 \to s_t})$. Therefore, higher inference probability of the right goal will result in a higher legibility. The second requirement is that this inference should happen as fast as possible. $f(t)$ provides a simple function to give higher weights to earlier parts of the trajectory. We use $f(t) = T - t$ with T as duration of the trajectory as suggested in [6]. Now we need to calculate the probability $P(g|\xi_{s_0 \to q})$ starting of with Bayes's Rule:

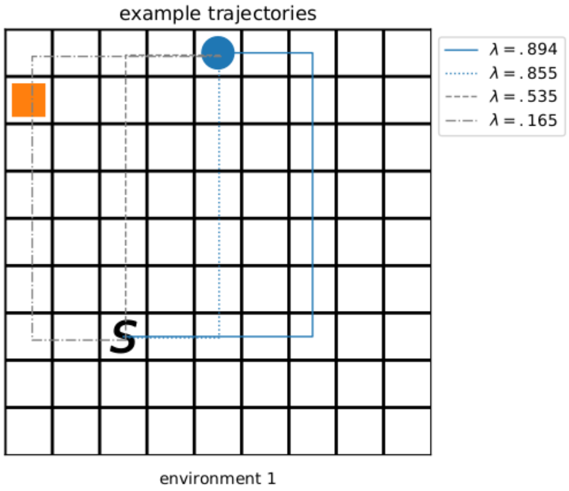

Fig. 2. Different example trajectories with corresponding legibility (λ) in environment 1. The start position is marked with 'S', the target goal is marked with a blue circle and the alternative goal is marked with an orange square.

$$P(g|\xi_{s_0 \to q}) \propto P(\xi_{s_0 \to q}|g) P(g) \tag{4}$$

$P(g|\xi_{s_0 \to q})$ is the probability that the agent follows $\xi_{s_0 \to q}$ when the agent targets a possible goal $g \in \mathcal{G}$. q can be any intermediate point. The prior probability of a goal $P(g)$ is assumed to be known, otherwise a uniform prior can be used. $P(\xi_{s_0 \to q}|g)$ can be computed as the ratio of all trajectories from s_0 to g that pass through $\xi_{s_0 \to q}$ to all trajectories from s_0 to g [6]:

$$P(\xi_{s_0 \to q}|g) = \frac{\int_{\xi_{q \to g}} P(\xi_{s_0 \to q \to g})}{\int_{\xi_{s_0 \to g}} P(\xi_{s_0 \to g})} \tag{5}$$

Following the assumption that trajectories are separable [46], i.e. $P(\xi_{s_0 \to q \to g}) = P(\xi_{s_0 \to q}) P(\xi_{q \to g})$, leads to:

$$P(\xi_{s_0 \to q}|g) = \frac{P(\xi_{s_0 \to q}) \int_{\xi_{q \to g}} P(\xi_{q \to g})}{\int_{\xi_{s_0 \to g}} P(\xi_{s_0 \to g})} \tag{6}$$

At this point, a model is required to express the probability of a trajectory in the eyes of an observer $P(g|\xi_{s_0 \to q})$. The principle of maximum entropy as suggested by [46] is adopted to model this probability as $P(\xi) \propto \exp(-C(\xi))$. $C(\xi)$ is the cost associated with trajectory ξ, therefore the probability of a trajectory decreases exponentially with increasing costs, leading to [6]:

$$P(\xi_{s_0 \to q}|g) \propto \frac{\exp(-C(\xi_{s_0 \to q}) \int_{\xi_{q \to g}} \exp(-C(\xi^*_{q \to g}))}{\int_{\xi_{s_0 \to g}} \exp(-C(\xi^*_{S \to G}))} \tag{7}$$

These integrals are computationally challenging and [47] derive an approximation with the assumptions that C is quadratic and its Hessian is constant. Under these assumptions according to Laplace's method we have $\int \exp(-C(\xi_{s_0 \to q})) \approx k \exp(-C(\xi^*_{s_0 \to q}))$, with the constant k and $\xi^*_{s_0 \to q}$ as the optimal trajectory from s to q w.r.t. C. Plugging this expression into (7) and using $z = \sum_G P(g|\xi_{s_0 \to q})$ to normalize the probability leads to [6]:

$$P(g|\xi_{s_0 \to q}) = \frac{1}{z} \frac{\exp(-C(\xi_{s_0 \to q}) - C(\xi^*_{q \to g}))}{\exp(-C(\xi^*_{s_0 \to g}))} P(g) \quad (8)$$

Approximating the cost C with the quadratic trajectory length in workspace punishes the agent from unnecessarily long paths [6]: $C = \sum_t \|\xi_{s_0 \to s_{t+1}} - \xi_{s_0 \to s_t}\|^2$. For the discrete case with step size of one, C is equivalent to the Manhattan distance. In situations with multiple goals, an agent can make trajectories more and more legible and never reaching a score of one while increasing the cost w.r.t to C more and more. In order to prevent the agent to go to far away from the observer's expectation, [6] propose to use a regularizer: $L(\xi) = \lambda(\xi) - \mu C(\xi)$. We did not apply this regularizer, because the agent is already being punished by the environment for longer trajectories. [6] show furthermore in an experiment with real humans that for legible trajectories the participants were faster able to infer the target goal with higher probability correctly.

C. Modeling the Observer

We compare four algorithms: Q-Learning (Q-L), Q-OBS-D, Q-OBS-P and Q-OBS-L. These algorithms differ in the strategy the observer $\hat{O}$ implements. An overview of the ideas for the used functions for $\hat{O}$ is shown in Table I. The algorithms only differ in the choice of $\hat{O}$ inserted in (2). The main difference is between Q-Learning as baseline algorithm which we consider non-interactive and the other three algorithms which we consider interactive. The main purpose of implementing different versions of the interactive methods is to explore how to integrate an observer that reasons about the possible goals of the agent in interactive RL. In the following we explain the proposed algorithms.

1) Q-L: Using the trivial equation for $\hat{O}$:

$$\hat{O} = 0 \quad (9)$$

is equivalent to plain Q-Learning. Q-Learning does not use any information from the observer and is therefore not interactive. Since Q-Learning only takes the rewards from the environment into account, it has no information on legibility. However, this does not mean that the learned trajectories can not be legible, we can expect that some trajectories more legible than others. Therefore, Q-Learning will serve us as comparison to have a baseline how legible the trajectories are just by chance.

2) Q-OBS-D:

$$\hat{O}(s, a, s') = \frac{1}{z} \exp(-\sigma d(s', g^*)) \quad (10)$$

d is the distance from s' to the goal using the Manhattan distance. z is partition function of the softmax distribution in order to normalize the probability to one. σ is the temperature parameter to adjust how 'sharp' the distribution peaks around the maximum. (10) only depends on the current state s' and not on the observed trajectory snippet. We consider this approach as a naive approach to estimate goal probability and expect it to work in some cases, as it gives an incentive to reduce the distance to the target goal early on, however in more complex configurations, e.g. when the target goal is behind another goal, this approach might not work. Therefor we expect it to work at least as good as Q-L, and in some cases even better.

TABLE I

THE DIFFERENT USED OBSERVER FUNCTIONS.

alg.	observer function idea
Q-L	Non-interactive baseline algorithm using no observer function
Q-OBS-D	Interactive algorithm using softmax of goal distance as observer function
Q-OBS-P	Interactive algorithm using the cost of the observed trajectory in comparison with the cost of the optimal trajectory as observer function
Q-OBS-L	Interactive algorithm using the legibility of the observed trajectory as observer function

3) Q-OBS-P: For Q-OBS-P we use the probability to reach a goal given a snippet of trajectory given with (8):

$$\hat{O}(\xi_{s_0 \to q}) = P(g^*|\xi_{s_0 \to q}) \quad (11)$$

Since this method uses a goal probability that has successfully been employed in previous research [6], [23], [24] it seems like a more suitable candidate to estimate the goal probability than Q-OBS-D and we expect it to perform better. For Q-OBS-L we directly use the legibility as feedback from the observer. For the discrete case with K as the number of steps for reaching q and s_k as the state after k steps (3) becomes:

$$\hat{O}(\xi_{s_0 \to q}) = \frac{\Sigma_{k=0}^{K} P(g|\xi_{s_0 \to s_k}) f(k)}{\Sigma_{k}^{K} f(k)} \quad (12)$$

Using directly the legibility is not a goal probability, since it does not sum up to one for all goals, nevertheless it contains by definition information on how confident the observer is that the agent is going for the target goal. Therefore we also expect this method to also perform better than Q-OBS-D.

IV. Experiments

The goal of the experiments is to evaluate the ability of the algorithms presented in Section III-C to increase the legibility of the learned trajectories. Q-Learning will serve as non-interactive baseline to compare to. Q-OBS-D, Q-OBS-P and Q-OBS-L integrate information information on the goal probability into the model and are expected to perform better. We evaluated the approach on five different environments. For the first environment there are only two possible goals, and we use it to illustrate the approach. For the environments 2 – 5 we use three goals and changed the configuration of these goals relative to each other.

The parameters were set intuitively. First we set the parameters that Q-Learning performed reasonable well and kept these parameters for the interactive algorithms. The parameters specific to the interactive algorithms were then set to perform reasonable well, but not tweaked to achieve the best possible performance. The parameters were kept for all environments. For the rewards from the environment we used: reaching the target goal $r_g = 0$, penalty for unvisited state different from the target goal $r_p = -0.1$, penalty for already visited state different from the target goal $r_{p2} = -0.2$. For the Q-Learning relevant parameters we used: $\alpha = 0.9$, $\gamma = 0.9$ and $\epsilon = 0.1$. The q-table was initialized with random values from 0 to 2. For Q-OBS-D we set $\sigma = 0.3$ For implementation reasons, to address the problem of positive loops we use $\beta = \beta_1 \beta_2$, with $\beta_1 = -r_p$ and $\beta_2 = 2$. By setting the parameters like this, we assure that agent does not achieve a net gain larger than 0 by cycling back and forth. However, the possible looping behavior drastically limits the choice to set β. Each algorithm was trained in 100 sessions for 120 episodes on each environment.

A. *Environment 1*

1) Description: The first environment was used to check the feasibility of the approach and includes only two goals: the target goal and one alternative goal. The size of the grid of the first environment is 9x9 and is visualized in Fig. 2 alongside with four example trajectories and the corresponding legibility. The first trajectory (from left to right) is sup-optimal in terms of steps towards the target goal and the legibility is low, the second and the third trajectory are both optimal, however the third trajectory yields a higher legibility because one can infer earlier for which goal the agent is aiming. The fourth trajectory is sup-optimal but the legibility is the highest of the shown trajectories. There are multiple optimal trajectories, when using Q-Learning, there is no reason for the agent to prefer one optimal trajectory over another optimal trajectory. Since the learning is stochastic, we expect the agent to sometimes learn an optimal trajectory with a higher legibility and other times with a lower legibility. We do not expect to learn with Q-Learning trajectories with a even higher legibility. When integrating the observer feedback, we expect the learned trajectories to be more legible and sometimes to even learn trajectories that are sub-optimal, but more legible than the most legible optimal trajectory. While we show only two possible optimal trajectories to the goal, there are more possible optimal trajectories to the goal than these two. These trajectories only differ in the legibility. Since the Q-table is randomly initialized and an ϵ-greedy exploration strategy is used the learning process is stochastic. Technically, we are not learning trajectories, but policies - once a policy is learned the trajectory generated by that policy are deterministic. When speaking about the learned trajectories, we are strictly speaking about the trajectories that are generated by the learned policies.

2) Results: As aforementioned even when only considering only the optimal trajectories there is a large number of possible trajectories. During the training processes of the different algorithms a large number of different trajectories have been learned. It is not possible to visualize the differences of the different algorithms in only one graph. Therefore we will use different methods to illustrate the occurred differences. First, we will have a look into the five best and five worst trajectories w.r.t. the legibility as illustrated in Fig. 3. If we now have a look at the legibility of the best and worst trajectories (w.r.t. λ) learned by Q-Learning, we see that these values are lower than legibility of the best and worst trajectories learned by the interactive algorithms. From Fig. 2 we know that the more legible trajectories tend to go to the right earlier on and are more on the right side of the grid in general. We can see that the trajectories of the interactive algorithms also tend to lie more on the right side of the grid world. The legibility of the different algorithms averaged over 100 runs for environment 1 is reported in Fig. 4. All algorithms that integrate a non-zero observer reward perform significantly better than plain Q-Learning. Q-OBS-P performs best regarding the legibility of the learned trajectories.

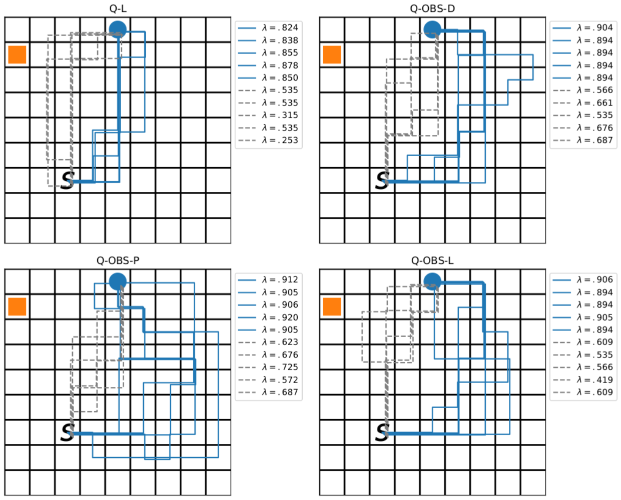

Fig. 3. The five best (solid blue lines) and five worst (dashed grey lines) learned trajectories w.r.t. λ for the different algorithms of environment 1.

B. *Environments 2 – 5*

While we showed in environment 1 that all interactive algorithms perform better than Q-Learning, we tested the approach on four additional environments to test the limits of our approach. This time we included an additional alternative goal. The grid size for tasks 2 – 5 is 9x9 as in environment 1. All tasks have three goals, the target goal and two alternative goals. The different environments can be seen in Fig. 5. We varied the relative configuration of the goals to evaluate the influence on the performance of the algorithms. In environment 2 – 4 the position of the alternative goals

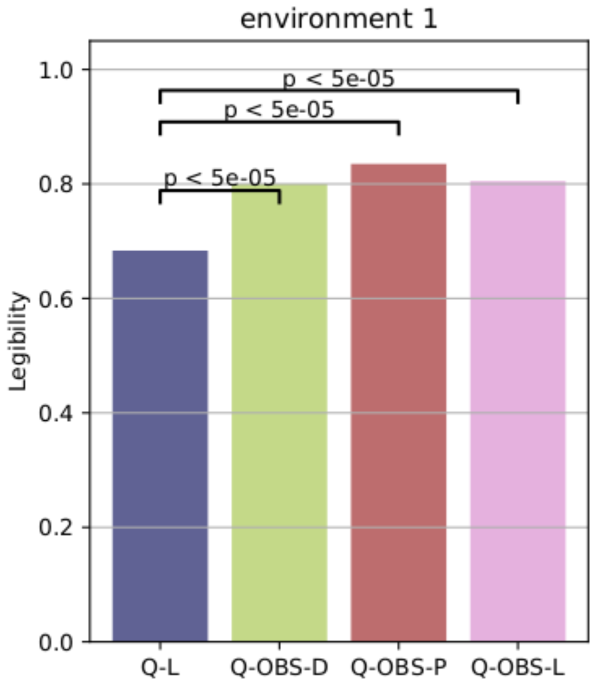

Fig. 4. Mean of the legibility for the different algorithms in environment 1. The significance level was calculated using the Mann-Whitney U test.

stays the same, we only vary the position of the target goal. In environment 5 there is no obvious more legible trajectory, so we do not expect including the observer feedback to perform better.

1) Results: The legibility of the different algorithms averaged over 100 runs for environment 2 – 5 is reported in Fig. 6. While Q-OBS-D significantly improved the legibility of the learned trajectory for environment 1, there is no significant difference for environment 2 and 4. As for environment 1, Q-OBS-P is the best performing algorithm for all environments with a significant higher legibility in comparison to Q-Learning. Q-OBS-P performs significantly better than Q-Learning for environments 2 – 4, but not for environment 5. For environment 5, from a human perspective, there is no more legible trajectory than the (only) optimal trajectory i.e. just going straight from start to the target goal. The only optimal trajectory has a legibility of $\lambda = 0.388$. In Fig. 7 we see the five best and five worst trajectories for environment 5. We see that in terms of the metric that all algorithms generated some trajectories with $\lambda > 0.388$.

V. Discussion

The results show that the interactive algorithms perform better than Q-Learning. Our main focus is on showing that the interactive approaches are useful in comparison to the non-interactive approaches. In order to support our main message it is not really important which of the interactive algorithm performs best. A major limitation is the use of Reward Shaping and a next step will be to replace it with a better suited method like Policy Shaping. Therefore it is not useful to put effort into analyzing differences in the approaches based on Policy Shaping, especially since the parameters were not tuned for every algorithm to perform to its best. In our approach we simulated the observer giving additional rewards to the agent. One possible idea is to employ a real human in the loop giving the observer feedback. Research on human feedback in RL (e.g. [8], [48]–[50]) suggests that probably real humans will behave differently than our models, therefore the framework might not work. However, our model might be useful even when no observer giving feedback is present. Since the agent has all the information the observer has, we could integrate the observer model internally into the agent. The agent could improve it's behavior by expecting to being watched. An approach like this would go into the direction of theory of mind. One downside of our approach is that the observe needs to know all the present goals in the setting to infer the goal probability. In robotics this is a strong assumption. An interesting problematic arises in this context: the robustness of the legibility if the observer has only partial knowledge of the goals. Also interesting is that in environment 5 there are more legible trajectories than the trajectory that goes directly from start to the target goal. From a human point of view there are arguably no more legible trajectories. For the legibility metric this happens, because for example going to the right after passing the level of the two alternative goals, drastically decreases the probability of the left goal and increases the probability of the right goal, therefore the target goal probability also increases. Simultaneously the length of the trajectory increases leading to a change of the weights for

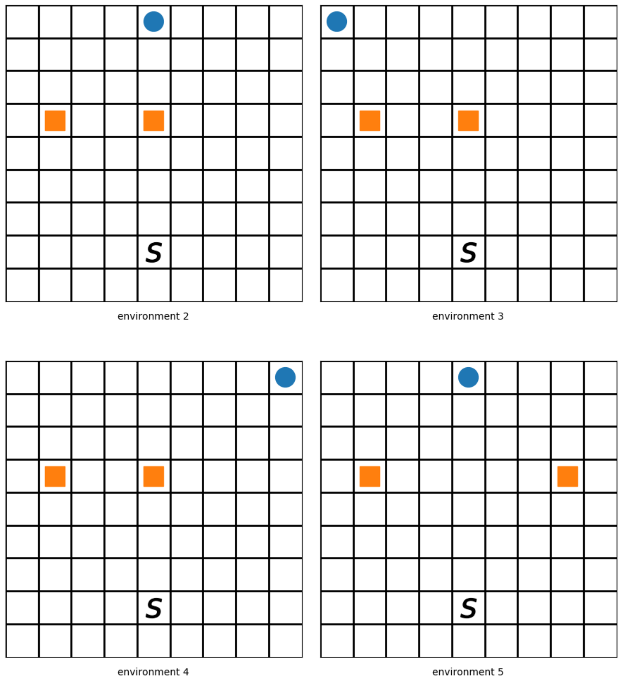

Fig. 5. Environment 2 – 5, for the environments 2 – 4 only the position of the target goal was varied, for environment 5 there is no legible path from a human point of view.The target goal is marked as a blue circle, the alternative goals as orange squares and the start with 'S'.

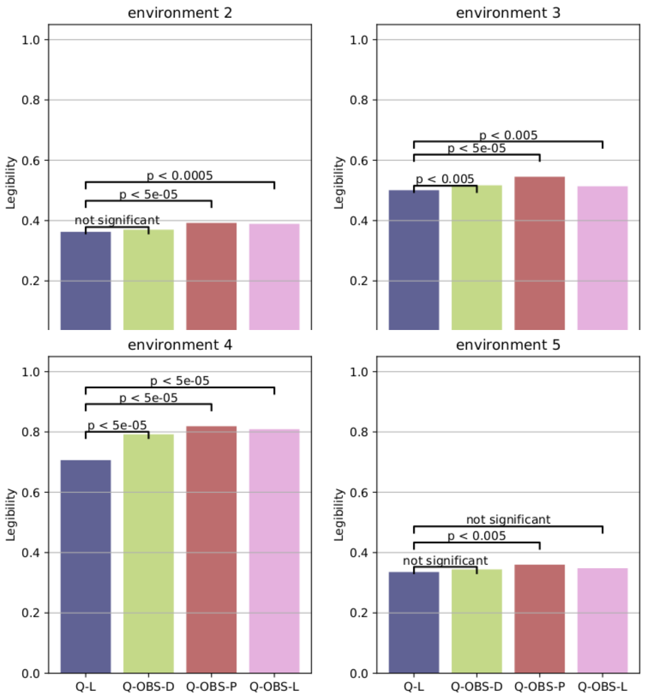

Fig. 6. Mean of the legibilities for Task 2 – 5 for the different algorithms. The significance level was calculated using the Mann-Whitney U test.

each part of the distribution. These two changes together can lead to an increase in legibility. It is not clear, if this will be a relevant problem, for longer continuous trajectories. That's another limitation we did not address in this work - scaling the approach up to a more complex task than just a grid world, possibly also using a robot with multiple degrees of freedom instead of just a point robot.

VI. Conclusions and Future Work

In this work, we were interested in integrating observer feedback into RL to increase the legibility of the learned trajectories. We proposed three interactive RL algorithms by integrating observer feedback and compared them to the non-interactive Q-Learning. We showed that the interactive RL approaches learn trajectories with a significantly higher legibility and that even a simple approach can perform at least as good as Q-Learning. From that, we conclude that when it comes to Human-Robot cooperation it is useful to integrate reasoning about the goal probabilities in order to increase the legibility of the trajectories. While we used Reward Shaping as a simple mechanism to integrate the feedback, the problem of positive-reward cycle is limiting the power of the approach. In future work we will consider other shaping mechanisms as Policy Shaping, as it will probably work better in experiments with real humans, which is another direction we are aiming for. Furthermore, we plan to extend the experiment to a more complex environment.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 765955.

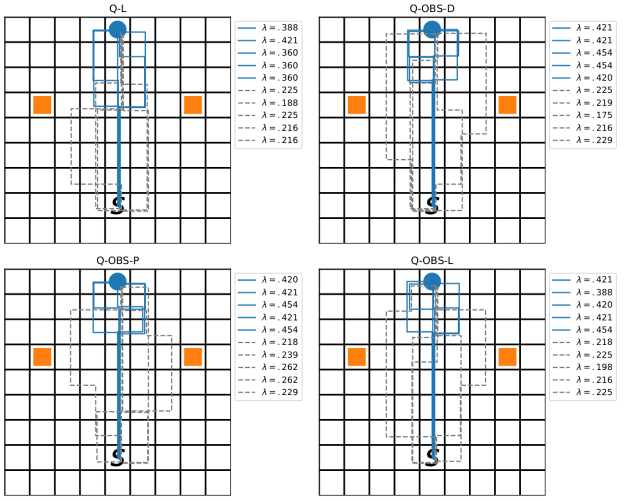

Fig. 7. The five best (solid blue lines) and five worst (dashed grey lines) learned trajectories w.r.t. λ for the different algorithms of environment 5.

References

[1] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll, "Understanding and sharing intentions: The origins of cultural cognition," *Behavioral and Brain Sciences*, vol. 28, no. 5, p. 675–691, 2005.

[2] J. Mainprice, R. Hayne, and D. Berenson, "Predicting human reaching motion in collaborative tasks using inverse optimal control and iterative re-planning," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 885–892.

[3] S. Wallkotter, S. Tulli, G. Castellano, A. Paiva, and M. Chetouani, "Explainable agents through social cues: A review," 2020.

[4] J. Broekens and M. Chetouani, "Towards transparent robot learning through tdrl-based emotional expressions," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[5] T. Chakraborti, A. Kulkarni, S. Sreedharan, D. E. Smith, and S. Kambhampati, "Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior," *CoRR*, vol. abs/1811.09722, 2018. [Online]. Available: http://arxiv.org/abs/1811.09722

[6] A. D. Dragan, K. C. T. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 03 2013, pp. 301–308.

[7] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 1998.

[8] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artificial Intelligence*, 2008.

[9] W. B. Knox and P. Stone, "Reinforcement learning from simultaneous human and mdp reward," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, ser. AAMAS '12. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2012, p. 475–482.

[10] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, and A. L. Thomaz, "Policy shaping: Integrating human feedback with reinforcement learning," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 2625–2633. [Online]. Available: http://papers.nips.cc/paper/5187-policy-shaping-integrating-human-feedback-with-reinforcement-learning.pdf

[11] J. MacGlashan, M. K. Ho, R. T. Loftin, B. Peng, D. L. Roberts, M. E. Taylor, and M. L. Littman, "Interactive learning from policy-dependent human feedback," *CoRR*, vol. abs/1701.06049, 2017. [Online]. Available: http://arxiv.org/abs/1701.06049

[12] C. Celemin, J. R. del Solar, and J. Kober, "A fast hybrid reinforcement learning framework with human corrective feedback," *Autonomous Robots*, vol. 43, pp. 1173–1186, 2019.

[13] A. Kulkarni, Y. Zha, T. Chakraborti, S. G. Vadlamudi, Y. Zhang, and S. Kambhampati, "Explicable planning as minimizing distance from expected behavior," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, ser. AAMAS '19. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2019, p. 2075–2077.

[14] Y. Zhang, S. Sreedharan, A. Kulkarni, T. Chakraborti, H. Zhuo, and S. Kambhampati, "Plan explicability and predictability for robot task planning," in *ICRA 2017 - IEEE International Conference on Robotics and Automation*. United States: Institute of Electrical and Electronics Engineers Inc., 7 2017, pp. 1313–1320.

[15] A. M. MacNally, N. Lipovetzky, M. Ramirez, and A. R. Pearce, "Action selection for transparent planning," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, ser. AAMAS '18. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2018, p. 1327–1335.

[16] J. F. Fisac, C. Liu, J. B. Hamrick, S. S. Sastry, J. K. Hedrick, T. L. Griffiths, and A. D. Dragan, "Generating plans that predict themselves," *CoRR*, vol. abs/1802.05250, 2018. [Online]. Available: http://arxiv.org/abs/1802.05250

[17] D. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences*, vol. 4, no. 4, pp. 515–629, 1978.

[18] C. L. Baker, R. Saxe, and J. B. Tenenbaum, "Action understanding as inverse planning," *Cognition*, vol. 113, pp. 329–349, 2009.

[19] A. N. Meltzoff, "Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children." US, pp. 838–850, 1995.

[20] S. C. Johnson, "The recognition of mentalistic agents in infancy," *Trends in Cognitive Sciences*, vol. 4, no. 1, pp. 22–28, 2000.

[21] G. Csibra and G. Gergely, "'obsessed with goals': functions and mechanisms of teleological interpretation of actions in humans." *Acta psychologica*, vol. 124 1, pp. 60–78, 2007.

[22] E. J. Carter, J. K. Hodgins, and D. H. Rakison, "Exploring the neural correlates of goal-directed action and intention understanding," *NeuroImage*, vol. 54, no. 2, pp. 1634 – 1642, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811910011808

[23] A. D. Dragan and S. S. Srinivasa, "Generating legible motion," in *Robotics: Science and Systems*, 2013.

[24] R. Holladay, A. Dragan, and S. Srinivasa, "Legible robot pointing," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, vol. 2014, 08 2014.

[25] G. Pezzulo, F. Donnarumma, and H. Dindo, "Human sensorimotor communication: A theory of signaling in online social interactions," *PLoS ONE*, 2013.

[26] G. Pezzulo, F. Donnarumma, H. Dindo, A. D'Ausilio, I. Konvalinka, and C. Castelfranchi, "The body talks: Sensorimotor communication and its brain and kinematic signatures," *Physics of Life Reviews*, vol. 28, pp. 1–21, 2019. [Online]. Available: https://doi.org/10.1016/j.plrev.2018.06.014

[27] S. Qi and S. Zhu, "Intent-aware multi-agent reinforcement learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 7533–7540.

[28] S. H. Huang, D. Held, P. Abbeel, and A. D. Dragan, "Enabling robots to communicate their objectives," *ArXiv*, vol. abs/1702.03465, 2017.

[29] S. Goldman and M. Kearns, "On the complexity of teaching," *J. Comput. Syst. Sci.*, vol. 50, no. 1, pp. 20–31, Feb. 1995. [Online]. Available: http://dx.doi.org/10.1006/jcss.1995.1003

[30] M. Cakmak and M. Lopes, "Algorithmic and human teaching of sequential decision tasks," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, ser. AAAI'12. AAAI Press, 2012, p. 1536–1542.

[31] X. Zhu, A. Singla, S. Zilles, and A. N. Rafferty, "An overview of machine teaching," *CoRR*, vol. abs/1801.05927, 2018. [Online]. Available: http://arxiv.org/abs/1801.05927

[32] M. K. Ho, M. Littman, J. MacGlashan, F. Cushman, and J. L. Austerweil, "Showing versus doing: Teaching by demonstration," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 3027–3035. [Online]. Available: http://papers.nips.cc/paper/6413-showing-versus-doing-teaching-by-demonstration.pdf

[33] M. K. Ho, M. L. Littman, F. Cushman, and J. L. Austerweil, "Effectively learning from pedagogical demonstrations," in *Annual Conference of the Cognitive Science Society (CogSci)*, 2018.

[34] P. Shafto, N. D. Goodman, and T. L. Griffiths, "A rational account of pedagogical reasoning: Teaching by, and learning from, examples," *Cognitive Psychology*, vol. 71, pp. 55–89, 2014.

[35] M. Bied and M. Chetouani, "Exploring the difference between solving and teaching in sensorimotor tasks," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 139–141. [Online]. Available: https://doi.org/10.1145/3371382.3378284

[36] S. Milli and A. D. Dragan, "Literal or pedagogic human? analyzing human model misspecification in objective learning," *CoRR*, vol. abs/1903.03877, 2019. [Online]. Available: http://arxiv.org/abs/1903.03877

[37] G. Li, R. Gomez, K. Nakamura, and B. He, "Human-centered reinforcement learning: A survey," *IEEE Transactions on Human-Machine Systems*, 05 2019.

[38] W. Bradley Knox and P. Stone, "Tamer: Training an agent manually via evaluative reinforcement," in *2008 7th IEEE International Conference on Development and Learning*, Aug 2008, pp. 292–297.

[39] W. B. Knox and P. Stone, "Interactively shaping agents via human reinforcement: The tamer framework," in *Proceedings of the Fifth International Conference on Knowledge Capture*, ser. K-CAP '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 9–16. [Online]. Available: https://doi.org/10.1145/1597735.1597738

[40] ——, "Combining manual feedback with subsequent mdp reward signals for reinforcement learning," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, ser. AAMAS '10. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2010, p. 5–12.

[41] ——, "Augmenting reinforcement learning with human feedback," in *ICML 2011 Workshop on New Developments in Imitation Learning*, July 2011.

[42] C. Celemin and J. Ruiz-del-Solar, "Coach: Learning continuous actions from corrective advice communicated by humans," in *2015 International Conference on Advanced Robotics (ICAR)*, July 2015, pp. 581–586.

[43] N. A. Vien, W. Ertel, and T. C. Chung, "Learning via human feedback in continuous state and action spaces," *Applied Intelligence*, vol. 39, no. 2, pp. 267–278, 2013. [Online]. Available: https://doi.org/10.1007/s10489-012-0412-6

[44] A. Y. Ng, D. Harada, and S. J. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Proceedings of the Sixteenth International Conference on Machine Learning*, ser. ICML '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, p. 278–287.

[45] A. Najar, O. Sigaud, and M. Chetouani, "Interactively shaping robot behaviour with unlabeled human instructions," *CoRR*, vol. abs/1902.01670, 2019. [Online]. Available: http://arxiv.org/abs/1902.01670

[46] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *Proc. AAAI*, 01 2008, pp. 1433–1438.

[47] A. Dragan and S. Srinivasa, "Formalizing assistive teleoperation," in *Proceedings of Robotics: Science and Systems*, July 2012.

[48] A. L. Thomaz and C. Breazeal, "Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance," in *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, ser. AAAI'06. AAAI Press, 2006, p. 1000–1005.

[49] M. K. Ho, M. L. Littman, F. Cushman, and J. L. Austerweil, "Teaching with rewards and punishments: Reinforcement or communication?" in *CogSci*, 2015.

[50] M. K. Ho, F. Cushman, M. L. Littman, and J. L. Austerweil, "People teach with rewards and punishments as communication, not reinforcements." *Journal of Experimental Psychology: General*, vol. 148, no. 3, pp. 520–549, 2019.