Learning from Explanations With Maximum Likelihood Inverse Reinforcement Learning

Silvia Tulli (silvia.tulli@gaips.inesc-id.pt)
Universidade de Lisboa Instituto Superior Tecnico https://orcid.org/0000-0002-6826-370X
Francisco S. Melo
INESC-ID: Instituto de Engenharia de Sistemas e Computadores Investigacao e Desenvolvimento em Lisboa
Ana Paiva
INESC-ID: Instituto de Engenharia de Sistemas e Computadores Investigacao e Desenvolvimento em Lisboa
Mohamed Chetouani
Sorbonne Universite

Research Article

Keywords: Explainable Inverse Reinforcement Learning, Learning from Demonstration, Maximum Likelihood Inverse Reinforcement Learning, Explainable Agents, Interactive Machine Learning

Posted Date: March 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1439366/v1

Learning from Explanations with Maximum Likelihood Inverse Reinforcement Learning

Silvia Tulli[1,2*], Francisco S. Melo[1], Ana Paiva[1] and Mohamed Chetouani[2]

[1]Department of Computer Science and Engineering, INESC-ID, Instituto Superior Técnico.
[2]Department of Mechanical, Acoustic, Electronics and Robotics Sciences, Institute for Intelligent Systems and Robotics, CNRS UMR 7222, Sorbonne Université.

*Corresponding author(s). E-mail(s): silvia.tulli@gaips.inesc-id.pt;
Contributing authors: fmelo@inesc-id.pt; ana.paiva@inesc-id.pt;
mohamed.chetouani@sorbonne-universite.fr;

Abstract

Our research effort takes inspiration from human social learning mechanisms to focus on situations in which an expert guides a learner through explanations. The proposed approach incorporates explanations into maximum likelihood inverse reinforcement learning. We computationally evaluate explanations against other teaching signals (reward, demonstration and explanation) in three navigational scenarios. The generated explanations are also evaluated in a user study with 150 participants. The user study investigates participants' preferences between the different types of teaching signals and the impact of contextual situations, i.e., distance from the task's goal, on their preferences. Our simulations' results show that explanations lead to better performance compared to reward and demonstration signals, and that explanations are preferred by human teachers in situations where the goal is far from the learner.

Keywords: Explainable Inverse Reinforcement Learning, Learning from Demonstration, Maximum Likelihood Inverse Reinforcement Learning, Explainable Agents, Interactive Machine Learning

1 Introduction

As robots and other autonomous agents enter our homes, hospitals, schools, and workplaces, it is critical to find ways to adapt their behaviour to tasks that occur unexpectedly, thus *learning* through natural, real-time interactions with the environment and its inhabitants [1]. There exists several learning strategies used by both humans and agents to learn a new task.

An approach to learning from other agents is imitation. Enabling machines to learn a desired behavior by imitating an expert's behavior has been proven to be a powerful tool to speed up the learning process [2]. This approach is inspired by human *imitation learning* (IL) processes, and is also known as *learning from demonstration* (LfD) [3, 4], *programming by demonstration* (PbD) [5], and *teaching by showing* [6]. Imagine having to define the reward function for an aerobatic helicopter flight [7]. The reward function would consist of many features describing relevant aspects for controlling the helicopter, e.g., desired velocity, centripetal acceleration, pitch and so on, that are difficult to detail using rewards. By recording a pilot's flight and using *imitation learning*, we can obtain the reward weights that result in policies that bring us closer to the expert, and drastically reduce the online computational cost of learning how to perform aerobatic helicopter moves. Due to its advantages, IL finds its application in many sequential decision problems including continuous and discrete optimization problems, and it is particularly useful in control problems where writing down the reward function that specifies how different desiderata should be traded off is challenging [8], e.g., driving a car [9], modeling human intents to navigate in a crowd environment [10], or synchronizing the lips of a cartoon character [11].

To exemplify the intents behind the experts' demonstrations and direct the learner towards crucial aspects of the task, humans often substitute or complement rewards and demonstrations with other teaching signals, such as explanations. The act of explaining can be thought of as a mean to transfer knowledge between an explainer, i.e., someone who is in possession of explanatory information, and an explainee, i.e., someone or a group of people who is thought not to possess it already [12, 13]. This process has been identified as the *social process* of the explanation [14]. In a continuous interaction between the explainer and her counterpart, the main goal of the explainer is to provide enough information to the explainee so that they can understand the causes of some fact or event. This process contemplates the active role of the explainee, which can ask for explanations by querying the explainer. In addition to the *social process*, explanation has been described also as a *cognitive process* and a *product* [15, 16]. The *cognitive process* concerns with abductive inference, a form of logical inference that starting from the observation or set of observations seeks for the simplest and more likely conclusion, i.e., explanatory hypotheses [17, 18].

In summary, explanations describe how and why something works the way it does, allowing humans to solve ambiguities in their current knowledge state [19, 20] and evaluate observed actions.

In the context of intelligent agents, explanations could be a valuable way to concisely describe a task or extrapolate useful information from a set of demonstrations, thus decrease the number of examples needed to replicate a behavior and generalize a certain knowledge to unseen situations.

There exists two compelling lines of research concerned respectively with how to incorporate the knowledge of an expert, and how to summarize the behavior of an expert.

So far, the work that has been done shows that inverse reinforcement learning (IRL) algorithms are helpful to integrate various types of previous knowledge [21–24]. The expert's knowledge is encoded in a set of demonstrations, each demonstration comprises samples exemplifying the behavior of the expert, e.g., the action selected in a specific state. To reduce the number of samples needed to learn the observed task, inverse reinforcement learners can query the demonstrator about specific states [25], rank demonstrations to extrapolate the underlying intent of the best demonstration [26], or learn progressively more challenging source tasks [27].

Moreover, the possibility to integrate statistical methods to estimate the parameters of an assumed probability distribution, given some demonstrations, allows further improvements in the performance of IRL agents [28].

Solutions for explaining the decisions of sequential decision making agents include: techniques to answer questions such as *"Why has this recommendation been made?"* by populating generic templates with domain-specific information from the task [29, 30], approaches to map action queries such as *"When do/will you action?"* into policy explanations by inspecting the states in which the input action is the most likely one [31–33], and methods to generate counterfactual explanations of behaviour based on the causal relationships between variables of interest [34, 35].

Human evaluations of agents' explanations generally take the form of a user study and examine the knowledge gained through task prediction performance, i.e., the explainee, recipient of the explanations, would be able to provide a better prediction if explanations successfully made the model intelligible. Comparatively, to the best of our knowledge, the evaluation of agents' explanations as teaching signals in learning and teaching scenarios is hardly explored. This work aims at understanding whether reasoning upon explanations of an expert would make a learning agent more efficient, and validating how this learning approach would be valuable in teaching scenarios involving humans. First, we introduce and formalize a method to integrate explanations into maximum likelihood inverse reinforcement learning. Second, we computationally evaluate our method on three navigational scenarios using three different types of teaching signals, i.e., reward, demonstration, explanation. Results indicate that explanations lead to better performance in all scenarios. Finally, we conduct a user study using the implemented teaching signals and evaluate participants' preferences in four different situations each constituted by a set of eight positions of the learner with respect to the goal. Results show that explanations are preferred when the learner is far from the goal.

2 Background

We formalize the problem of *learning from explanations* (LfE) as an inverse reinforcement learning problem and use ideas from optimal control such as the incompletely-known *Markov Decision Process*, and *value function*. A *Markov Decision Process* (MDP) encodes the task of an agent by describing its sequential decision making as a tuple $(\mathcal{S}, \mathcal{A}, \{P_a\}, r)$ where $\mathcal{S}$ represents the state-space, $\mathcal{A}$ the action-space, and $\{P_a, a \in \mathcal{A}\}$ denotes the set of transition probabilities defining the dynamics of the MDP, i.e.,

$$P_a(s' \mid s) \triangleq \Pr(S_{t+1} = s' \mid S_t = s, A_t = a).$$

$r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function. An MDP describes the interaction between a decision-maker and its environment. The goal of the decision-maker is to determine the series of actions that maximize the agent's total discounted reward - i.e. $TDR = E\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}\right]$ where R_{t+1} is the random reward received by the agent at time step $t+1$ as a consequence of performing some action A_t in state S_t. The principle used to select a series of actions, i.e., the policy, is a mapping $\pi : S \times \mathcal{A} \to [0, 1]$. A policy π can be described as the probability of choosing a certain action in a given state $P_a[A_t = a \mid S_t = s] = \pi(s, a)$. The value $v^\pi(s)$ of a policy π can be denoted as:

$$v^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) \mid S_0 = s \right].$$

where $v^\pi(s)$ is the sum of the discounted rewards for all time steps, and γ is the discount factor assigning the importance to the rewards obtained over time. Whereas the reward signal r specifies what is good in an immediate sense, the *value function* $v^\pi(s)$ specifies what is good in the long run [36].
$v^\pi(s)$ is the value that the decision-maker expects to collect by starting in state s and selecting its actions according to π such that the value for all states is the higher $v^{\pi^*}(s) \geq v^\pi(s)$, with v^{π^*} denoting the optimal value following optimal policy π^*. The optimal value function verifies the recursive relation:

$$v^{\pi^*}(s) = \max_{a \in \mathcal{A}} \left[r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P_a(s' \mid s) v^{\pi^*}(s)(s') \right]$$

Conversely, the optimal q-function, or action-value function, is the value that the decision-maker expects to collect starting from state s, taking action a, following policy π. The optimal q-function is defined as:

$$q^*(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P_a(s' \mid s) v^{\pi^*}(s')$$

The q-function summarizes all the relevant information the agent has to know by providing a ranking for the actions according to how useful they are

for a particular goal that the agent has. By using this function, at each state the agent can search for the action that has the maximum q^* value. Both the optimal policy π^* and v^{π^*} can be computed from q^* as:

$$\pi^*(s) = \underset{a \in A}{arg\,max}\ q^*(s,a)$$

$$v^*(s) = \max_{a \in \mathcal{A}} q^*(s,a)$$

Following the notation in the literature [37–40], we will assume that the reward function to be learned, henceforth denoted as r^*, can be represented as a linear combination of features, i.e., given a set of $\mathcal{K}$ features $\phi_1, \ldots, \phi_K$ with $\phi_k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for $k = 1, \ldots, K$ and for some weight vector w^*,

$$r^*(s,a) = \sum_{k=1}^{K} \phi_k(s,a) w_k^* = \phi^\top(s,a) w^*$$

Thus, we can rewrite the *value function* of a policy as:

$$V_R^\pi = \mathbb{E}_{s_0 \sim S_0}\left[\sum_{t=0}^{\infty} \gamma^t w^T \phi\left(s_t\right) \mid \pi\right] = w^T \phi(\pi)$$

We can then call the remaining term $\phi(\pi)$ the feature expectations of the policy π. To estimate the value of the linear combination of features describing the expert's reward function we use *maximum likelihood estimation* (MLE).

Given that the reward function f and the dynamics of the system M are known linear mappings and assuming the data is large enough, finding θ involves solving a system of linear equations. Maximum likelihood estimation has been used in previous work in inverse reinforcement learning, more details on this can be found in section 3.

3 Related Work

The literature addressing the topic of sequential decision making agent's explanations fall into two main categories: work on learning from explanations, and work on generating explanations. Our work lies in the overlap of these two.

3.1 Learning Rewards from Explanations

In the context of reinforcement learning and inverse reinforcement learning particular emphasis has been placed on incorporating verbal instructions by either combining natural language with demonstrations [41] and by using sentiment analysis to filter natural language input into advice [42, 43].

The work of Macglashan et al. [41] introduces an architecture for sentence–trajectory pairs, where the learner has access to natural language input,

and demonstrations of appropriate behavior. Their architecture include maximum likelihood inverse reinforcement learning to estimate the expert's reward function from linguistic feedback available at different stages in the learning process.

Accounting for the fact that humans often do not specify state information, e.g., *Mario should jump on enemies*, the work of Krening et al. [42] propose a method named *object-focused advice* in which the human advice is tied to objects instead of specific states and is generalized over the objects' state space. Their experiment collects information on the nature of explanations, the accuracy of their sentiment analysis to filter explanations, and the performance of agents trained with *object-focused advice*. Their results show that their method is able to capture human explanations without state information and increases the performance of reinforcement learning agents. They observe that free-form explanations, i.e., human explanations not constrained to a template, vary in many ways. The level of detail and abstraction used to describe desired actions seem to reflect the amount of prior knowledge the learning agent is assumed to have. However, while a sentiment filter can process free-form explanations, the majority of the sentences are not actionable and cannot be directly utilized as advice.

In line with the use of sentiment analysis to filter linguistic feedback, in the work of Sumers et al. [43] the learner grounds linguistic feedback to elements of the task, e.g., *"Good job"* refers to prior behavior, whereas *"You should have gone to the living room"* refers to an action, and assigns a positive or negative sentiment to that behavior. The positive or negative valence of the behavior implies a positive or negative rewards on its features. Linguistic feedback are processed as *evaluative*, *imperative*, and *descriptive* feedback. An evaluative feedback corresponds to a scalar value in response to the agent's actions and have positive or negative valence (+1/-1). An imperative feedback gives information about the correct action in a given state by mapping the language input into a set of state-action pairs. A descriptive feedback provides information about the state transition function, i.e., how the teacher's preference changes in response to an action. Their results show that the learner is able to learn from all types of linguistic feedback, obtaining the best scores when trained with descriptive feedback.

The use of linguistic feedback as teaching signals to transfer knowledge has been studied both in the context of human social learning and machine learning. Explanations help establish a connection between what has been observed and its causes, and serve as a principled basis for generalization [14]. Consequently, explanations scaffold causal learning and have a crucial role in inference [44]. Following this idea, our work also generate explanations in the form of sentence-trajectories and uses maximum likelihood inverse reinforcement learning to find a weighting of the state features that (locally) maximizes the probability of these trajectories. We provide a framework to learn from explanations and allow a fair comparison with other types of teaching signals, i.e., reward, demonstration. In addition, we evaluate the generated teaching

signals in a user study, accounting for different situations and positions of the learner with respect to the goal.

3.2 Explainable Reinforcement Learning

Research efforts in explainable reinforcement learning introduce the concepts of *minimal sufficient explanation* and *explanation* for compactly explaining action preferences via decomposed reward [30]. Alternative approaches use causal models to generate explanations for *why* and *why not* questions [35].

Juozapaitis et al. [30] study reward decomposition[1] for the purpose of explanation and focus on pairwise action explanations where the goal is to explain why one action is preferred to another in a particular state. They define *reward difference* explanations (RDX) as the difference of the decomposed q-vectors $\Delta(s, a_1, a_2) = \vec{Q}(s, a_1) - \vec{Q}(s, a_2)$. Each component of the RDX is a positive or negative reason $\Delta_c(s, a_1, a_2)$ for the preference depending on whether a_1 has an advantage (disadvantage) over a_2 with respect to reward type c. The *minimal sufficient explanation* is a more compact version of RDX comprises of a small set of the most important reasons about why an action is preferred to another.

In extension, Madumal et al. [35] formalize an *action influence model* to learn the quantitative influences that actions have on variables of interest. Their computational evaluations are accompanied with a user study to measure the participants' performance in task prediction, explanation satisfaction, and trust.

Similarly to Juozapaitis et al. [30], our work provides information about the positive or negative valence of an action in a certain state by comparing the q-values of two possible actions. Thanks to this comparison our system builds explanations similar to the ones generated by causal approaches, thus replying to *why* and *why not* questions. In addition, our approach accounts for the goodness of the state and includes a parameter to indicate how trustworthy the explanations are.

Differently from previous works [35, 43], we evaluate explanation against other types of teaching signals, i.e., reward and demonstration, controlling for the situation and the position of the learner with respect to the goal.

4 Learning from Explanations

We present a maximum likelihood inverse reinforcement learning approach to allow agents to learn from explanation. Within this framework we define three different teaching signals: reward, demonstration, explanation. A reward signal emulates the reinforcement learning approach and includes information about the state, the action and the reward associated with the state-action pair. A demonstration signal mimics the learning from demonstration approach and

[1] Reward decomposition consists in decomposing a reward function by specifying a set of reward components/types

constitutes of information about the state and the action. Finally, a explanation signal gives information about the state, the action, the contrastive action, the next state and the goodness of that state.

4.1 Learning a task

Throughout this document, we consider a learner who knows a rewardless MDP, i.e., $(\mathcal{S}, \mathcal{A}, \{P_a\}, \gamma)$, and must learn a task description in the form of a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. We assume that the reward to be learned, henceforth denoted as r^*, can be represented as a linear combination of features as described in section 2.

We denote by v^*_w, π^*_w, and q^*_w the optimal value function, policy, and q-function for the MDP $\left(\mathcal{S}, \mathcal{A}, \{P_a\}, \phi^\top w, \gamma\right)$. The goal of the learner is, therefore, to recover some weight vector w^*.

4.2 Learning a task from rewards

The first approach to learning the task is given samples of the reward. A sample consists of a triplet (s, a, r), where r is a reward observed upon performing a in state s. Specifically, we assume that the sample rewards r correspond to independent observations of $r^*(s, a)$ corrupted by zero-mean Gaussian noise with known precision η, so that:

$$\Pr\left(s, a, r \mid r^* = \phi^\top w\right) = \text{Normal}\left(r - \phi^\top(s, a)w; 0, \eta\right) \tag{1}$$

The maximum likelihood estimate for w^* thus comes:

$$\begin{aligned}
\hat{w}^* &= \text{argmax}_{w \in \mathbb{R}^K} \prod_{n=1}^{N} \text{Normal}\left(r_n - \phi^\top\left(s_n, a_n\right) w; 0, \eta\right) \\
&= \text{argmax}_{w \in \mathbb{R}^K} \sum_{n=1}^{N} \log \text{Normal}\left(r_n - \phi^\top\left(s_n, a_n\right) w; 0, \eta\right) \\
&= \text{argmin}_{w \in \mathbb{R}^K} \sum_{n=1}^{N} \left(r_n - \phi^\top\left(s_n, a_n\right) w\right)^2
\end{aligned} \tag{2}$$

Then, given a set of samples $\{(s_n, a_n, r_n), n = 1, \ldots, N\}$, we want to compute a weight vector w to minimize the loss:

$$L(w) = \sum_{n=1}^{N} \left(r_n - \phi^\top\left(s_n, a_n\right) w\right)^2 \tag{3}$$

Using standard stochastic gradient descent we get the online update:

$$w_{n+1} = w_n + \alpha_n \phi\left(s_n, a_n\right)\left(r_n - \phi^\top\left(s_n, a_n\right) w_n\right) \tag{4}$$

4.3 Learning a task from demonstrations

The second approach we consider is to recover the task from sample demonstrations. We consider a demonstration as a pair (s, a), indicating that the optimal action in state s is a. We assume that the demonstrations are independent and subject to noise, such that

$$\Pr\left(s, a \mid r^* = \phi^\top w\right) = \sigma_w(s, a; \eta) \triangleq \frac{\exp\left\{\eta Q_w^*(s, a)\right\}}{\sum_{a' \in \mathcal{A}} \exp\left\{\eta Q_w^*\left(s, a'\right)\right\}}, \tag{5}$$

where η is a *confidence parameter*, indicating how trustworthy the demonstrations are. Smaller η allows for more imprecise demonstrations, while larger η requires more precise demonstrations. The expert draws actions from a Boltzmann distribution (softmax) over the learned q-values. The maximum likelihood estimate for w^* thus comes

$$\begin{aligned} \hat{w}^* &= \sum_{n=1}^{N} \log \sigma_w\left(x_n, a_n; \eta\right) \\ &= \operatorname{argmin} \sum_{n=1}^{N} \left(\log \sum_{a' \in \mathcal{A}} \exp\left\{\eta Q_w^*\left(s_n, a'\right)\right\} - \eta Q_w^*\left(s_n, a_n\right) \right). \end{aligned} \tag{6}$$

Then, given a set of demonstrations $\{(s_n, a_n), n = 1, \ldots, N\}$, we want to compute w to minimize the loss

$$L(w) = \sum_{n=1}^{N} \left(\log \sum_{a' \in \mathcal{A}} \exp\left\{\eta Q_w^*\left(s_n, a'\right)\right\} - \eta Q_w^*\left(s_n, a_n\right) \right). \tag{7}$$

For a single sample (s_n, a_n),

$$\frac{\partial L(w)}{\partial Q_w^*\left(s_n, a\right)} = \eta\left(\sigma_w\left(s_n, a; \eta\right) - \delta_{a, a_n}\right) \tag{8}$$

where δ_{a,a_n} is the Kronecker delta function. On the other hand, let

$$P_{\pi_w^*}\left(s' \mid s\right) = \sum_{a \in \mathcal{A}} \pi_w^*(a \mid s) P_a\left(s' \mid s\right) \tag{9}$$

and let $\Phi_{\pi_w^*}$ denote $|\mathcal{S}| \times K$ with s, k element given by

$$\left[\Phi_{\pi_w^*}\right]_{s,k} = \phi_{\pi_w^*, k}(s) \triangleq \sum_{a \in \mathcal{A}} \pi_w^*(a \mid s) \phi_k(s, a). \tag{10}$$

We have that

$$\frac{\partial Q_w^*\left(s_n, a\right)}{\partial w_k} = \phi_k\left(s_n, a\right) + \gamma P_a\left(s_n\right)\left(I - \gamma P_{\pi_w^*}\right)^{-1} \Phi_{\pi_w^*, k}, \tag{11}$$

where $P_a(s)$ is (row) vector corresponding to sth row of P_a. We ignored the dependence of π_w^* on w. This finally yields

$$\nabla_w L(w) = \sum_{a \in \mathcal{A}} \eta \left(\phi\left(s_n, a\right) + \gamma \left(P_a\left(s_n\right)\left(I - \gamma P_{\pi_w^*}\right)^{-1} \Phi_{\pi_w^*} \right)^\top \right) \left(\sigma_w\left(s_n, a\right) - \delta_{a, a_n}\right) \tag{12}$$

and we get the update

$$w_{n+1} = w_n + \alpha_n \sum_{a \in \mathcal{A}} \eta \left(\phi(s_n, a) + \gamma \left(P_a(s_n) \left(I - \gamma P_{\pi^*_w} \right)^{-1} \Phi_{\pi^*_w} \right)^\top \right) \left(\sigma_w(s_n, a) - \delta_{a,a_n} \right) \tag{13}$$

4.4 Learning a task from explanations

We finally consider learning a task from explanations. We consider an explanation a tuple (s, a, b, s', v) with $s, s' \in \mathcal{S}$, $a, b \in \mathcal{A}$, and $v \in \{-1, +1\}$ determines the positive or negative valence of the explanation (good/bad). The natural language explanation can be written with following semantics:

- In state s action a is *better* than action b because it will eventually lead you through state s' and that is *good*.
- In state s action a is *worse* than action b because it will eventually lead you through state s' and that is *bad*.

We consider that, given the target reward r,

$$r_{\pi^*}(s) = \sum_{a \in \mathcal{A}} \pi^*(a \mid s) r(s, a) \tag{14}$$

- A state s is *good* in the above sense, if $r_{\pi^*}(s) > 0$ for the optimal policy π^*_r given that reward.
- A state s is *bad* in the above sense, if $r_{\pi^*}(s) < 0$ for the optimal policy π^*_r given that reward.
- An action a is *better* than b in state s and leading to state s' if the following two conditions are cumulatively met:
  - $Q^*(s, a) > Q^*(s, b)$ (a is better than b)
    - The transition probabilities by first taking action a and then following π^* lead to larger transition probability from s to s' than those same probabilities taking action b.
- An action a is *worse* than an action b in state s and leading to state s' if the following two conditions are cumulatively met:
  - $Q^*(s, a) < Q^*(s, a')$ (a is worse than b)
    - The transition probabilities by taking action a in s and following π^* elsewhere lead to larger transition probability from s to s' than those same probabilities taking action b.
- Finally, we consider that an action a in state s leads to state b if the policy $\hat{\pi}$ defined as

$$\hat{\pi}(s') = \begin{cases} a & \text{if } s' = s \\ \pi^*(s') & \text{otherwise} \end{cases} \tag{15}$$

is such that

$$P_a^\infty(s' \mid s) \triangleq (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P_{\hat{\pi}}^t(s' \mid s) > 0. \tag{16}$$

Let

$$\sigma(z;\eta) = \frac{1}{1+\exp\{-\eta z\}} \tag{17}$$

where η is *confidence parameter*, indicating how trustworthy the explanations are. We consider

$$\Pr\left(s,a,s',b,v \mid r^* = \phi^\top w\right) = \sigma\left(v r_{\pi_w^*}(s');\eta\right)\sigma\left(v\left(Q_w^*(s,a) - Q_w^*(s,b)\right);\eta\right)P_a^\infty(s' \mid s) \tag{18}$$

assuming, as before, that the explanations are independent and potentially *noisy*. Then,

$$\begin{aligned}-\log\Pr\left(s,a,s',b,v \mid r^* = \phi^\top w\right) = &\log\left(1+\exp\left(-\eta v r_{\pi_w^*}(s')\right)\right)\\ &+\log\left(1+\exp\left(-\eta v\left(Q_w^*(s,a) - Q_w^*(s,b)\right)\right)\right)\\ &-\log\left(P_a^\infty(s' \mid s)\right)\end{aligned} \tag{19}$$

Then, given a set of explanations, $\{(s_n,a_n,b_n,s_n',v_n), n=1,\ldots,N\}$, we get the loss function

$$\begin{aligned}L(w) = &\sum_{n=1}^{N}\log\left(1+\exp\left(-\eta v r_{\pi_w^*}\left(s_n'\right)\right)\right)\\ &+\log\left(1+\exp\left(-\eta v\left(Q_w^*(s,a) - Q_w^*(s,b)\right)\right)\right)\\ &-\log\left(P_a^\infty(s' \mid s)\right)\end{aligned} \tag{20}$$

which can be optimized online using standard stochastic gradient descent. We have

$$\nabla_w \log\left(1+\exp\left(-\eta v r_{\pi_w^*}(s')\right)\right) = \eta v \phi_{\pi_w^*}(s')\left(\sigma\left(v r_{\pi_w^*}(s');\eta\right) - 1\right) \tag{21}$$

where we again disregard the dependence of π_w^* on w. Similarly,

$$\begin{aligned}\nabla_w \log(1+\exp(-\eta v(Q_w^*(s,a) - Q_w^*(s,b)))) \quad &= \eta v(\phi(s,a) - \phi(s,b)\\ +\gamma((P_a(s) - P_b(s))(I - \gamma P_{\pi_w^*})^{-1}\Phi_{\pi_w^*})^\top)&(\sigma(v(Q_w^*(s,a) - Q_w^*(s,b));\eta) - 1)w\end{aligned} \tag{22}$$

Finally, we have that,

$$P_a^\infty = (I - \gamma P_{\hat{\pi}})^{-1} \tag{23}$$

and the computation of $\frac{\partial P_a^\infty}{\partial w_k}$ is far from trivial, since the dependence of $\hat{\pi}$ on w is highly nonlinear and, in general, non-differentiable. To make the computation feasible, we instead consider a smooth approximation to $\hat{\pi}$, whereby

$$\hat{\pi}\left(a' \mid s'\right) \approx \begin{cases} 1.0 & \text{if } s = s' \text{ and } a = a' \\ \sigma_w\left(a' \mid s'\right) & \text{otherwise} \end{cases} \tag{24}$$

with σ_w defined in 4.2 Then,

$$\begin{aligned} \frac{\partial P_a^{\infty}}{\partial w_k} &= \gamma \left(I - \gamma P_{\hat{\pi}}\right)^{-1} \frac{\partial P_{\hat{\pi}}}{\partial w_k} \left(I - \gamma P_{\hat{\pi}}\right)^{-1} \\ &= \gamma P_a^{\infty} \frac{\partial P_{\hat{\pi}}}{\partial w_k} P_a^{\infty} \end{aligned} \tag{25}$$

with

$$P_{\hat{\pi}}(s' \mid s) = \sum_{a \in \mathcal{A}} \hat{\pi}(a \mid s) P_a(s' \mid s). \tag{26}$$

This yields

$$\begin{aligned} \nabla_w P_{\hat{\pi}}(s' \mid s) &= \sum_{a \in \mathcal{A}} \nabla_w \hat{\pi}(a \mid s) P_a(s' \mid s) \\ &= \sum_{a \in \mathcal{A}} \frac{\hat{\pi}(a \mid s)}{\hat{\pi}(a \mid s)} \nabla_w \hat{\pi}(a \mid s) P_a(s' \mid s) \\ &= \sum_{a \in \mathcal{A}} \hat{\pi}(a \mid s) \nabla_w \log \hat{\pi}(a \mid s) P_a(s' \mid s). \end{aligned} \tag{27}$$

Using the results from learning from demonstrations method,

$$\begin{aligned} \frac{\partial P_{\hat{\pi}}(s' \mid s)}{\partial w_k} = \sum_{a \in \mathcal{A}} \hat{\pi}(a \mid s) \sum_{a' \in \mathcal{A}} \eta(\phi_k(s, a') + \gamma(P_{a'}(s) P_a^{\infty} \phi_{\pi_w^*, k})^{\top})(\sigma_w(s, a') \\ - \delta_{a,a'}) P_a(s' \mid s) \end{aligned} \tag{28}$$

Finally, putting everything together,

$$\frac{\partial \log P_a^{\infty}(s' \mid s)}{\partial w_k} = \frac{\gamma}{P_a^{\infty}(s' \mid s)} \left(P_a^{\infty} \frac{\partial P_{\hat{\pi}}}{\partial w_k} P_a^{\infty} \right)_{s,s'} \tag{29}$$

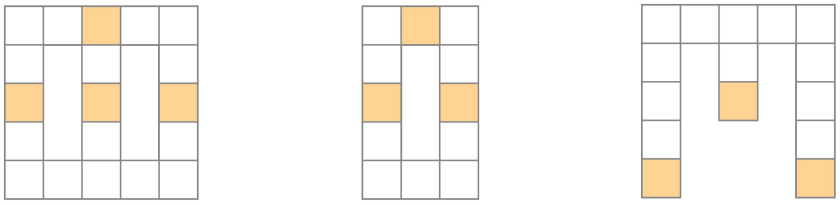
5 Experiments

5.1 Simulation Experiment

We evaluate the learning from explanations (LfE) framework to determine if leads to better performance compared to other types of learning approaches, i.e., learning from rewards and demonstrations. The effectiveness of the learned policy is evaluated based on a agent's capability of using the learned reward in the original problem. We refer to this capability as the performance of the agent. Our hypothesis **(H1)** is that agents will learn more efficiently from explanations than from both rewards and demonstrations.

Methodology

We consider the environments depicted in Figure 1. The agent operates in a grid, and is able to move in four directions, i.e., *up*, *down*, *left*, *right*, or stay in place *stay*. Each action moves the agent deterministically to an adjacent cell, factoring in obstacles. The agent is rewarded for navigating from an initial random state A to the goal B, which is one of the colored cells, in the most efficient way. In the simulation, the movement actions succeed with probability 0.8 and fail with probability 0.2. The reward is a linear combination of features, each corresponding to the indicator for one of the states in orange. Each environment has a different configuration. Environment 1a is a 5×5 grid with 19 possible states and four objects. Environment 1b is a 3×5 grid with 12 states and three objects. Environment 1c is a 6×5 grid with 15 states and three objects. The combination of corridors and objects simulates ambiguous situations in which explanations might help in choosing the best action depending on the state of the agent and the reward associated with each of the objects.



(a) Env 1 (b) Env 2 (c) Env 3

Fig. 1: Navigational environments used for the computational evaluation of the learning from explanation (LfE) framework.

We designed an experiment with three conditions: (1) learning from rewards (LfR), (2) learning from demonstrations (LfD), (3) learning from explanations (LfE). At the beginning, we generate a random reward as a convex combination of features, and compare the performance of an expert, a LfR updater, a LfD updater and a LfE updater in randomly selected sample rewards, demos and explanations.

We start by defining the samplers. The reward sampler selects a random state s and a random action a from the given MDP M and provides the corresponding noisy reward r as a linear combination of features from a standard normal distribution. The demonstration sampler selects a random state s and chooses an action a from a Boltzmann distribution (softmax) over the learned q-values. The explanation sampler samples a random initial state s, a random initial action a, a random second action (different from the first) b, a random via state with non-zero reward (to fit the good or bad description) based on the discounted state visitation frequencies given a policy π. After ensuring

that there is a next state to sample, the explanation sampler finally checks if the next state is good or bad comparing the q-values. The q-values provide information on whether, on the long run, a certain state or action will lead to better rewards. The reward, demonstration or explanation samples generate natural language strings that take the forms of: "The reward in state 7 when performing action Left is -0.05.", "In state 7 you should perform action Up.", "In *state*, *action* is *better/worse* than *action* because it may eventually lead you through *state* and that is *better/worse*", respectively.

Procedure

We run a comparative study. Every 10 steps, for each of the three approaches, we: estimate the reward, compute the associated optimal policy π^* from the reward parameters w^*, and evaluate that policy in the correct MDP. We test the proposed approaches by providing the learner with 15 samples, using each of the three methods. Each sample is selected randomly according to the corresponding distributions. The results are depicted in Figure 2, and correspond to the performance of the policy of the learner using the learned reward in the original problem. Each plot corresponds to the average of 30 independent runs where, in each run, the parameters w^* are sampled randomly from a Gaussian distribution with mean 0 and unit standard deviation. For reproducibility, we set 40 random seeds.

Results

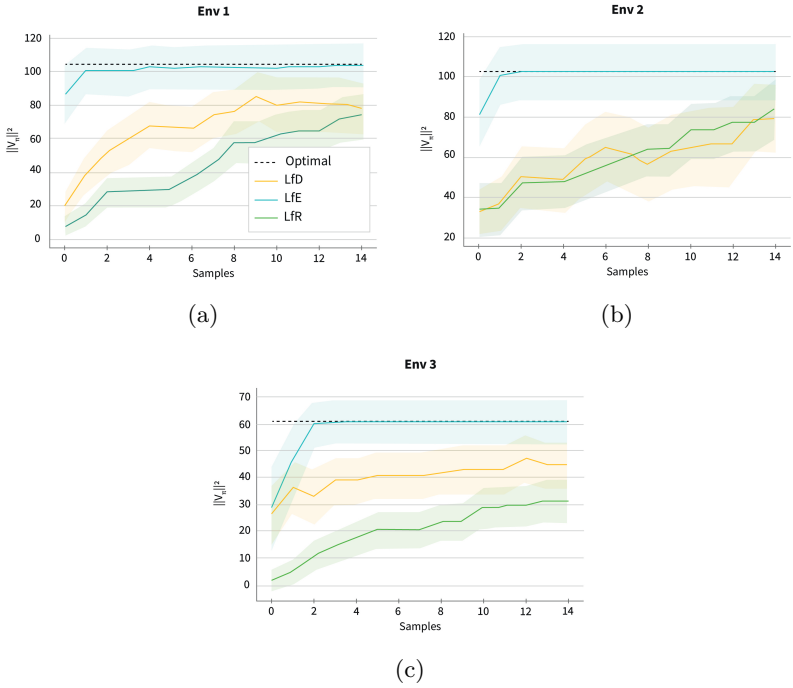


Fig. 2: Average return against the number of samples grouped by condition. Mean and confidence intervals for 40 seeds.

We plot the average return against the number of samples grouped by condition (Figure 2). The black dotted line corresponds to the expert's performance. As expected, and similar to results found by [43] when agents' were learning from descriptive feedback, the performance of the agent increases when learning from explanation samples. LfE agent outperforms both the LfD and LfR agents reaching near-optimal performance in all the environments. Looking closer to type of samples that hindered better performance, the analysis of the average returns after 1 learning update revealed a statistically significant difference between the performance of the LfE and LfD agents (Env 1: $t = -3.709$, $p = 0.000$, Env 2: $t = -3.495$, $p = 0.001$, Env 3: $t = -2.891$, $p = 0.007$). Moreover, the LfD agent performs better than the LfR agent in 1a and 1b, and achieves similar performance in 1b. All the agents eventually learn how to perform the task. The results of the simulations validate **H1**, showing trends that LfE approach leads to more efficient learning. That is, the LfE algorithm enables the learner to imitate the expert behavior achieving better performance than LfR and LfD do all over the tasks. This result confirms that LfE is

more sample efficient than LfR and LfD approaches in terms of the expert demonstration.

5.2 User Study

To validate our approach with humans, we evaluated the different teaching signals in a user study. The teaching signals generated with our system, take the form of sentences including the information detailed in section 4. We seek to investigate how humans select teaching signals and if their choice changes depending on the position of the learner with respect to the goal.

We hypothesize that when choosing among teaching signals **(H2)** humans will generally prefer explanations. Moreover, **(H3)** humans would prefer explanations over both rewards and demonstrations depending on the contextual situation, i.e., how close is the goal with respect to the learner position.

Methodology

The experiment consist of four phases: (1) familiarization, (2) structured teaching signals, (3) free-form explanations, and (4) subjective evaluation around the informativeness of the structure of the teaching signal. Participants had the possibility to navigate the environment until they were comfortable with it. Within the study, we propose four situations based on the goal (Figure 4), each situation including eight relevant states depicting the learner distance from the goal: far, ambiguous and adjacent. Far goals are equal or more than 3 steps away from the player, ambiguous goals are two steps away from the player and had a negative colored cell at equal distance, adjacent goals are next to the player. Situations were accompanied with information about the goal as well as a set of structured teaching signals. Examples of structured teaching signals include: (1) demonstration, i.e., *"In cell 17 you should move right."*, (2) reward, i.e., *"The amount of points you get in cell 17 when moving left is around 0.05."*, (3) explanation, i.e., *"In cell 17, moving Right is better than Up because it may eventually lead you through cell 11 and that is good."* Finally, we ask three exploratory questions to rate the informativeness of the general structure of the explanations we employ throughout the study.

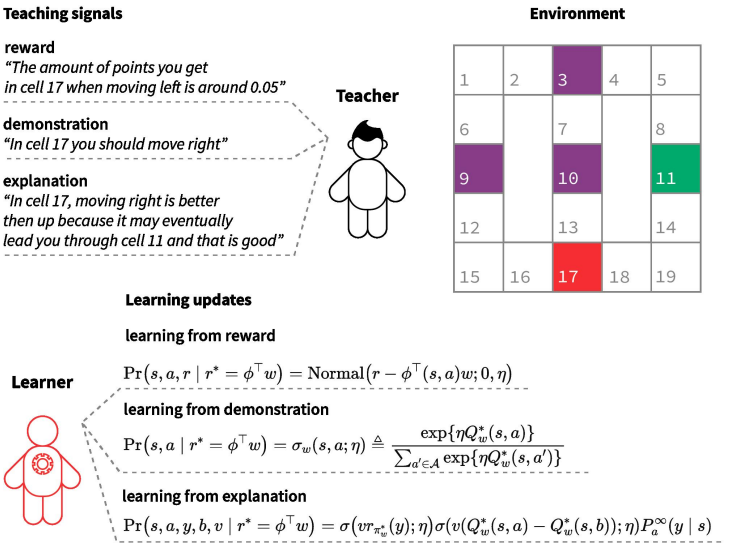


Fig. 3: An example of three teaching signals for the depicted situation. The learner is in cell 17 and has to go to cell 11. Participants are asked to choose among three types of teaching signals. The teaching signals used for the user study were the ones we generated and used for training the agents.

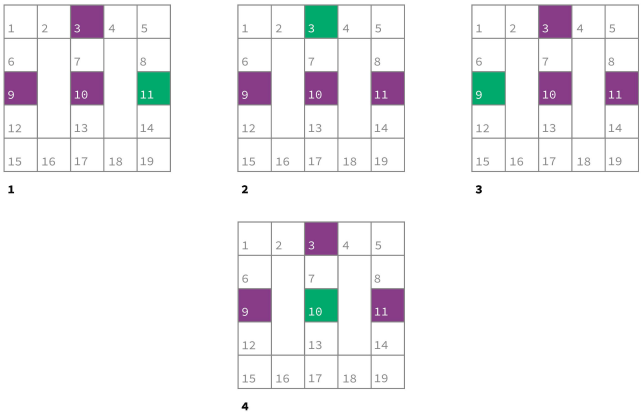


Fig. 4: The environment consists of 19 states disposed across three corridors. Each reward feature is represented by the colored states, i.e., states 9, 10, 11 and 3. Each configuration of the reward features represents a situation: Situation 1 - reaching state 11, Situation 2 - reaching state 3, Situation 3 - reaching state 9, Situation 4 - reaching state 10.

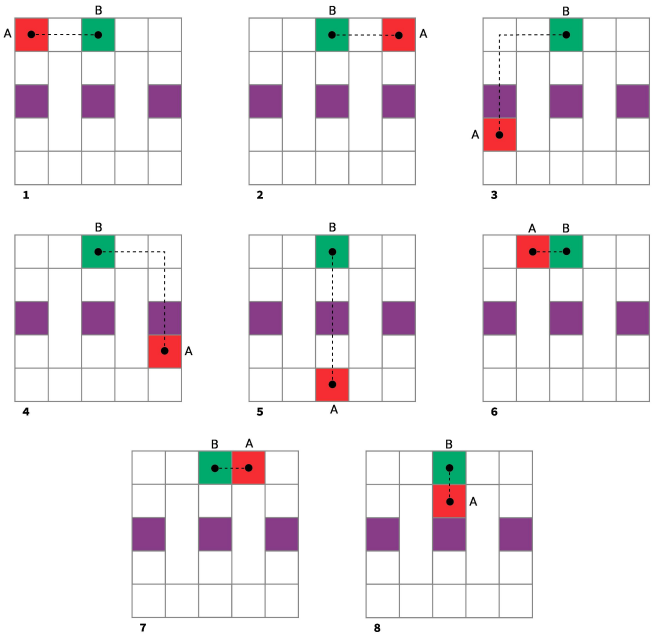


Fig. 5: An example of the eight positions of the learner, represented by the red checker, with respect to the goal, i.e., reaching the green cell. The learner can be in an ambiguous position (case 1 and 2), far (case 3, 4, and 5) or adjacent to the goal (case 6, 7, 8). An ambiguous position is a position where the learner is equally distant to a cell leading to a negative reward and the goal cell leading to a positive reward, a far position is a position where the player is at minimum 4 cells from the goal, an adjacent position is a position where the player is next to the goal.

To measure humans' explanatory preferences **(H2)**, we ask participants to select the best teaching signal among demonstration, reward, and explanation, and to provide an open answer to the question *"If you were asked to provide an explanation in this situation, what would that be?"*. To measure whether the contextual situation affects human's explanatory preferences **(H3)** we cluster the situations based on the learner's distance from the goal.

Participants

We recruited 150 participants using Prolific[2]. All participants were English speakers and gave informed consent to participate (age M = 40.24, SD = 13.43, gender Female = 91, Male = 59). We introduced some attention and verification questions in order to ensure the quality of the data. We asked four multi-answer questions related to the scenario (e.g., *"What is the goal of the player? (Check all that apply)"*) and computed an attention score based on the number of correct answers. The criteria to exclude participants were: not

[2] https://prolific.co/

having completed the entire experiment; and having an attention score of less than 70%. Consequently, we ran the data analysis on both the entire sample and on the reduced sample of 60 more reliable participants (age M = 39.83, SD = 12.85, Female = 33).

Materials

The self-assessed questionnaire included some demographic questions (age, gender, higher level of education), two items regarding participants self-perceived familiarity with navigational games and with reinforcement learning, three items regarding the perceived informativeness of the proposed teaching signals and four validation questions randomly dispersed in the questionnaire to evaluate their understanding of the rules of the game.

Procedure

After replying to the self-assessed questionnaire, participants were asked to consider a single-player video game based on a two-dimensional navigational environment in which a player, represented by a red checker, has to reach a hidden goal. Their role was to guide an hypothetical learner by providing the most informative teaching signals by either choosing among the structured signals and/or writing their own feedback.

Results

General Preference To obtain an overview of the participants' explanatory preferences, we summed up the number of teaching signals in all situations. The one-way analysis of variance of the teaching signals shows that there is a significant difference on how participants selected demonstrations, rewards and explanations (Kruskal-Wallis H test: $H = 16.810, p = .000$). The Mann-Whitney U test between explanations ($M = 13.375, SE = 5.909, SD = 24.763$) and demonstrations ($M = 4.82, SE = 6.120, SD = 33.858$) revealed a significant difference ($U = 10.0; p = .02$). A significant different was also found between demonstrations and rewards ($M = 1.625, SE = .374, SD = 5.402$)($U = 64, p = .00$), and explanations and rewards ($U = 61, p = .002$). These results suggest that participants' generally prefer demonstrations.

Influence of Contextual Situation The Kruskal-Wallis H test of the teaching signals, i.e., explanations, demonstrations, rewards, revealed that the main effect of the situation was significant across situation Situation 3 ($H(1, 150) = 13.022, p = .001$)), and Situation 4 ($H(1, 150) = 16.810, p = .000$). The specific values per each teaching signal were: Situation 3 [explanations ($M = 12.75, SE = 3.648, SD = 9.653$), demonstrations ($M = 37.875, SE = 7.024, SD = 18.583$), rewards ($M = 1.875, SE = 1.315, SD = 3.479$)] Situation 4 [explanations ($M = 13.375, SE = 5.909, SD = 9.653$), demonstrations ($M = 45.0, SE = 6.120, SD = 18.583$), rewards ($M = 1.625, SE = .374, SD = 3.479$)].

The analysis of variance in teaching signals revealed a main effect of the player distance from the goal for adjacent goals in Situation 2 ($H(3, 150) = 4.705, p = .049$) and in all situations for far goals: Situation 1 ($H(3, 150) = 4.705, p = .006$), Situation 2 ($H(3, 150) = 4.705, p = .025$), Situation 3 ($H(3, 150) = 2.0, p = .014$), Situation 4 ($H(3, 150) = 4.705, p = .02$). These results indicate that the teaching signals are selected differently based on the position of the player with respect to the goal.

Overall, participants reduce the number of explanations when the player is adjacent to the goal, while consistently choose to give more explanations when the player is far from the goal.

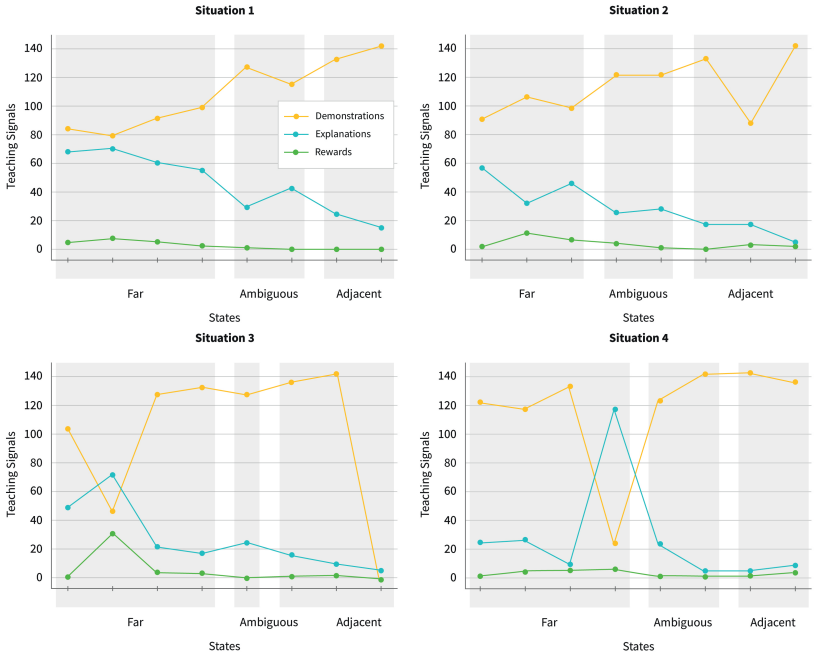


Fig. 6: Teaching signals per situations against position of the learner.

Our results show that participants prefer to teach through demonstrations significantly more than rewards and explanations, therefore **H2** is not confirmed. Moreover, in all situations participants consistently tend to decrease the number of explanations as the learner get closer to the goal, confirming **H3**.

5.3 Discussion

In this work we explored the role of explanation in learning and teaching tasks. First, we investigated whether or not explanations can lead to better

performance and introduced the *learning from explanations* (LfE) approach for recovering a reward function from explanations of another agent. Second, we evaluated the generated explanations in a user study. Hence, we compared the performance of an agent learning from explanations against demonstrations and rewards.

Regarding **H1** we predicted that agents would learn more efficiently from explanations than other types of teaching signals. The simulations confirmed our hypothesis by showing that in the explanation condition the agent was capable of achieving better performance. With our approach rewards were chosen to maximize the likelihood of the data given as a set of traces of optimal behavior, allowing us to combine a supervised-learning component with a flexible hypothesis class given as input. The LfE algorithm is simple, with relatively low computational cost per iteration. However, the maximum likelihood algorithm generally behaves well when the expert's demonstration are representative of the task [28]. Further work should focus on testing this approach in other learning scenarios adopting more robust and scalable implementations.

According to **H2**, we expected humans to generally prefer explanations over rewards and explanations to guide a player towards a hidden goal. However, this was not supported. Instead, we found that participants preferred to teach through demonstrations significantly more than rewards and explanations. Our results are in line with the work of [45] which compares *simplicity* and *probability* in causal explanation, and states that simpler explanations are preferred and judged more likely. Thus, teaching signals which invoke a limited number of causes whilst still conveying relevant aspects of a task are favored, i.e., demonstrations.

In **H3**, we hypothesized that humans would choose explanations depending on the contextual situation. We observed that in all situations participants consistently tend to decrease the number of explanations as the player get closer to the goal. This might be due the fact that explanations conveys complex concepts better than demonstrations while relying on shared context to permit high bandwidth. In contrast, demonstrations are lower-bandwidth but more robust [46], therefore seen as more useful in situations in which there is little ambiguity, i.e., when the learner is not far from the goal.

6 Conclusion

Throughout this work we explore the problem of learning from explanations and provide a framework to compare learning from explanations with learning from other types of teaching signals. We present an application of maximum likelihood inverse reinforcement learning to the problem of training an agent to follow different teaching signals representing high-level tasks. We evaluate our approach in three navigational scenarios. We then undertake a user study with 150 participants to investigate humans' preferences between the different types of teaching signals and the impact of contextual situations on their choice. The first takeaway from this work is that we can improve agents performance by

integrating explanations into IRL. The second takeaway, derived from the user study, is that in the context of interactive task learning, humans might prefer different types of teaching signals depending on the contextual situations.

As a consequence of our approach we show that explanations can be a more succinct, robust, and transferable way to represent tasks. This approach is presumably robust enough to be applicable to a large range of sequential planning tasks in both human-agent and agent-agent settings.

In our work, we introduced explanations as a teaching signal by comparing the valence of actions and states. Yet explanations can take several other forms. The explanations can vary between teachers and with respect to their assumptions about the learner's knowledge and learning capabilities [47]. For example, the type of information that an explanation presents varies with the teachers' familiarity with the task or their model of the agent's functioning etc. Explanations can also result from a dialogue or interaction structure [25, 48]. Such diverseness in explanation are omnipresent in the real world and it would be interesting to incorporate them in future studies since many complex dynamics can emerge that we might be overlooking when assuming predefined templates and fixed level of abstraction of the task. Furthermore, future works shall study different methods to learn from explanation, for example, by testing different combinations of distributions and statistical inference (i.e., Bayesian inference) [49]. Lastly, we advocate studying the effect of generated agents' explanations with human learners.

Acknowledgement

We would like to thank Rebecca Stower and Ramona Merhej for proofreading the work.

Declarations

Funding. This work received funding from the EU Horizon 2020 research and innovation program for grant agreement No 765955 ANIMATAS project, and was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UIDB/50021/2020.

Conflict of Interest. The authors have no relevant financial or non-financial interests to disclose.

Ethical Approval. All participants were recruited using Prolific and gave informed consent to participate.

References

[1] Laird, J.E., Gluck, K.A., Anderson, J.R., Forbus, K.D., Jenkins, O.C., Lebiere, C., Salvucci, D.D., Scheutz, M., Thomaz, A.L., Trafton, J.G., Wray, R.E., Mohan, S., Kirk, J.R.: Interactive task learning. IEEE Intelligent Systems **32**, 6–21 (2017)

[2] Osa, T., Pajarinen, J., Neumann, G., Bagnell, J.A., Abbeel, P., Peters, J.: An algorithmic perspective on imitation learning. Found. Trends Robotics **7**, 1–179 (2018)

[3] Argall, B., Chernova, S., Veloso, M., Browning, B.: A survey of robot learning from demonstration. Robotics Auton. Syst. **57**, 469–483 (2009)

[4] Billard, A., Grollman, D.H.: Robot learning by demonstration. Scholarpedia **8**, 3824 (2013)

[5] Billard, A., Calinon, S., Dillmann, R., Schaal, S.: Robot programming by demonstration. In: Springer Handbook of Robotics (2008)

[6] Montgomery, J.F., Bekey, G.A.: Learning helicopter control through "teaching by showing". Proceedings of the 37th IEEE Conference on Decision and Control (Cat. No.98CH36171) **4**, 3647–36524 (1998)

[7] Abbeel, P., Coates, A., Quigley, M., Ng, A.: An application of reinforcement learning to aerobatic helicopter flight. In: NIPS (2006)

[8] Abbeel, P., Ng, A.Y.: Inverse Reinforcement Learning, pp. 554–558. Springer, Boston, MA (2010). https://doi.org/10.1007/978-0-387-30164-8_417

[9] Pomerleau, D.A.: Alvinn: An autonomous land vehicle in a neural network. In: NIPS (1988)

[10] Ziebart, B.D., Ratliff, N.D., Gallagher, G., Mertz, C., Peterson, K.M., Bagnell, J.A., Hebert, M., Dey, A.K., Srinivasa, S.S.: Planning-based prediction for pedestrians. 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, 3931–3936 (2009)

[11] Taylor, M.E., Stone, P.: Behavior transfer for value-function-based reinforcement learning. In: AAMAS '05 (2005)

[12] Lewis, D.: Causal explanation. In: Lewis, D. (ed.) Philosophical Papers Vol. Ii, pp. 214–240. Oxford University Press, ??? (1986)

[13] de Graaf, M.M.A., Malle, B.: How people explain action (and autonomous intelligent systems should too). In: AAAI Fall Symposia (2017)

[14] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. CoRR **abs/1706.07269** (2017) arXiv:1706.07269

[15] Lombrozo, T.: The structure and function of explanations. Trends in Cognitive Sciences **10**(10), 464–470 (2006). https://doi.org/10.1016/j.tics.2006.08.004

[16] Chin-Parker, S., Bradner, A.: Background shifts affect explanatory style: how a pragmatic theory of explanation accounts for background effects in the generation of explanations. Cognitive Processing **11**, 227–249 (2009)

[17] Aliseda, A.: Abductive Reasoning vol. 330. Springer, ??? (2006)

[18] Campos, D.G.: On the distinction between peirce's abduction and lipton's inference to the best explanation. Synthese **180**(3), 419–442 (2011)

[19] Ny Vasila and Azzurra Ruggeri and Tania Lombrozo: When and how children use explanations to guide generalizations. Cognitive Development **61** (2021)

[20] Michael M Chouinard: Children's questions: a mechanism for cognitive development. Monographs of the Society for Research in Child Development **72**, 126 (2007)

[21] Choi, J., Kim, K.: MAP inference for bayesian inverse reinforcement learning. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F.C.N., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a Meeting Held 12-14 December 2011, Granada, Spain, pp. 1989–1997 (2011). https://proceedings.neurips.cc/paper/2011/hash/3a15c7d0bbe60300a39f76f8a5ba6896-Abstract.html

[22] Chan, A.J., van der Schaar, M.: Scalable bayesian inverse reinforcement learning. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, ??? (2021). https://openreview.net/forum?id=4qR3coiNaIv

[23] Ramachandran, D., Amir, E.: Bayesian inverse reinforcement learning. In: Veloso, M.M. (ed.) IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007, pp. 2586–2591 (2007). http://ijcai.org/Proceedings/07/Papers/416.pdf

[24] Hadfield-Menell, D., Dragan, A.D., Abbeel, P., Russell, S.J.: Cooperative inverse reinforcement learning. CoRR **abs/1606.03137** (2016) 1606.03137

[25] Lopes, M., Melo, F.S., Montesano, L.: Active learning for reward estimation in inverse reinforcement learning. In: Buntine, W.L., Grobelnik, M., Mladenic, D., Shawe-Taylor, J. (eds.) Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II. Lecture Notes in Computer Science, vol. 5782, pp. 31–46. Springer, ??? (2009). https://

doi.org/10.1007/978-3-642-04174-7_3. https://doi.org/10.1007/978-3-642-04174-7_3

[26] Brown, D.S., Goo, W., Nagarajan, P., Niekum, S.: Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 783–792. PMLR, ??? (2019). http://proceedings.mlr.press/v97/brown19a.html

[27] Theodorou, A., Wortham, R., Bryson, J.: Why is my robot behaving like that? designing transparency for real time inspection of autonomous robots. AISB Workshop on Principles of Robotics (2016)

[28] Vroman, M.C.: Maximum likelihood inverse reinforcement learning. (2014)

[29] Khan, O.Z., Poupart, P., Black, J.P.: Minimal sufficient explanations for factored markov decision processes. In: Proceedings of the Nineteenth International Conference on International Conference on Automated Planning and Scheduling. ICAPS'09, pp. 194–200. AAAI Press, ??? (2009)

[30] Juozapaitis, Z., Koul, A., Fern, A., Erwig, M., Doshi-Velez, F.: Explainable reinforcement learning via reward decomposition. (2019)

[31] Hayes, B., Shah, J.A.: Improving robot controller transparency through autonomous policy explanation. 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI, 303–312 (2017)

[32] Amir, D., Amir, O.: Highlights: Summarizing agent behavior to people. AAMAS / (2018)

[33] Tsirtsis, S., De, A., Gomez-Rodriguez, M.: Counterfactual explanations in sequential decision making under uncertainty. ArXiv **abs/2107.02776** (2021)

[34] Zhang, J., Bareinboim, E.: Fairness in decision-making — the causal explanation formula. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (2018)

[35] Madumal, P., Miller, T., Sonenberg, L., Vetere, F.: Explainable reinforcement learning through a causal lens. ArXiv **abs/1905.10958** (2020)

[36] Barto, A.G., Sutton, R.S., Anderson, C.W.: Looking back on the actor-critic architecture. IEEE Trans. Syst. Man Cybern. Syst. **51**(1), 40–50

(2021). https://doi.org/10.1109/TSMC.2020.3041775

[37] Abbeel, P., Ng, A.: Apprenticeship learning via inverse reinforcement learning. Proceedings of the twenty-first international conference on Machine learning (2004)

[38] Ziebart, B.D., Maas, A.L., Bagnell, J.A., Dey, A.K.: Maximum entropy inverse reinforcement learning. In: AAAI (2008)

[39] Sadigh, D., Dragan, A.D., Sastry, S.S., Seshia, S.A.: Active preference-based learning of reward functions. In: Robotics: Science and Systems (2017)

[40] Pirotta, M., Restelli, M.: Inverse reinforcement learning through policy gradient minimization. AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (2016)

[41] Babes-Vroman, M., MacGlashan, J., Gao, R., Winner, K., Adjogah, R., desJardins, M., Littman, M.S., Muresan, S.: Learning to interpret natural language instructions. (2012)

[42] Krening, S., Harrison, B., Feigh, K.M., Isbell, C.L., Riedl, M., Thomaz, A.: Learning from explanations using sentiment and advice in rl. IEEE Transactions on Cognitive and Developmental Systems **9**(1), 44–55 (2017)

[43] Sumers, T.R., Ho, M.K., Griffiths, T.L.: Show or tell? demonstration is more robust to changes in shared perception than explanation. CoRR **abs/2012.09035** (2020) 2012.09035

[44] Lee, M.S., Admoni, H., Simmons, R.: Machine teaching for human inverse reinforcement learning. Frontiers in Robotics and AI **8** (2021)

[45] Lombrozo, T.: Simplicity and probability in causal explanation. Cognitive Psychology **55**(3), 232–257 (2007). https://doi.org/10.1016/j.cogpsych.2006.09.006

[46] Sumers, T., Ho, M.K., Hawkins, R.D., Griffiths, T.: Show or tell? teaching with language outperforms demonstration but only when context is shared. PsyArXiv (2021)

[47] Jara-Ettinger, J.: Theory of mind as inverse reinforcement learning. Current Opinion in Behavioral Sciences **29**, 105–110 (2019)

[48] Madumal, P., Miller, T., Vetere, F., Sonenberg, L.: Towards a grounded dialog model for explainable artificial intelligence. Workshop on Socio-cognitive Systems IJCAI **abs/1806.08055** (2018) arXiv:1806.08055

[49] Vélez, N., Gweon, H.: Learning from other minds: an optimistic critique

of reinforcement learning models of social learning. Current Opinion in Behavioral Sciences **38**, 110–115 (2021)