ZKG INTERNATIONAL

ISSN: 2366-1313

# A Critical Evaluation of Interactive Reinforcement Learning Through Social Reactions

*Vijaya Kumar Elpula, Research Scholar and Dr. Suribabu Potnuri, Professor*
*Department of Computer Science and Engineering, J.S. University, Shikohabad, U.P., India email: vijayakumarvarma@gmail.com*

## ABSTRACT

The process of a reinforcement learning agent acquiring new abilities involves the agent seeing and reacting to its environment. There has been a limited amount of implementation of reinforcement learning in real applications because of the problem of sample efficiency. In addition to accelerating the learning process as a whole, the objective of developing interactive reinforcement learning was to make it simpler for normal people to instruct and assess bots. There is the potential for a broad range of hardware-delivered, naturally occurring interactions (such as facial expressions, speech, or gestures) to serve as input for agent learning, taking inspiration from actual biological learning circumstances that occur in the real world. In addition to this, the agent is able to acquire knowledge from both multimodal and unimodal sensory input. The purpose of this study is to examine various approaches for delivering feedback and methods for learning from human social interaction. These techniques are used to interactive reinforcement learning robots. Last but not least, we discuss a few issues that have not been addressed and possible directions for additional research.

Keywords--Human agent/robot interaction, interactive reinforcement learning, interactive shaping,social interaction.

## I INTRODUCTION

Many real-world issues have been solved with surprising success using reinforcement learning (RL) [1, 2]. A growing number of researchers are turning to RL—now known as deep RL—as a solution to end-to-end learning in sequential choice problems, thanks to developments in deep learning [3]. The issue of sample efficiency, however, has

ISSN: 2366-1313

severely restricted the practical use of RL and deep RL. To acquire an effective strategy for playing a video game, for instance, an RL agent may need millions of training samples [3]. Most real-world applications of RL and deep RL will include agents or robots that interact with humans in their homes. The importance of, and need for, contact between agent and human will only grow. Thereby, the agent's learning might be guided by the vast amounts of user-generated information.

Human trainers may guide agents' learning in several methods, including by demonstrating concepts, giving instructions and guidance, and offering evaluation feedback [4]-[12]. A human user may demonstrate something to an agent either physically or via remote control [10], [13]. Inverse reinforcement learning [14] is one method of learning from demonstrations; it allows agents to maximize policies by learning a reward function from given demonstrations. The majority of demos involve solving RL tasks using one-time interactions utilizing inverse RL or initializing the agent's policy. Nevertheless, it may be very challenging for human trainers without expertise to provide top-notch demonstrations in sectors involving complicated tasks.

Using natural language is another method that human trainers guide agents to learn [15]. Improving reinforcement agent learning often requires encoding the advice into a computer language or mapping it from a formal language to natural language [16], [17]. By associating intermediate shaping incentives with free-form natural language instructions, the agent may also learn from instruction.

One may pick up a reward function from them and use it to learn how to obey linguistic instructions [18, 19]. Optimum action advice, optimum gain-risk advice, and other forms of guidance are studied for their impact on the agent's learning performance [20]. The human user may also teach the bots by providing evaluation comments. The phrase "human-centered reinforcement learning" is an approach to agent learning that relies on human evaluates for feedback [21]. Algorithms for learning may vary depending on how evaluative feedback—whether it's numerical reward, discrete categorization, or policy—is understood. Buttons and mouse clicks provide the majority of human input in these experiments. They may educate the agent

ISSN: 2366-1313

more organically by delivering feedback via emotions, gestures, or even natural languages, drawing inspiration from real-life biological learning settings. Both of these forms of social feedback—demonstration, guidance, and instruction—and the agent's own internal evaluations may help it grow.

The purpose of this paper is to review the current literature on training agents to solve reinforcement learning tasks using various forms of human social feedback, such as evaluation feedback and advice/instruction, since there are already some survey papers on learning from demonstration and observation [10], [22]. Both model-based and model-free approaches, as in conventional RL, are viable options for learning from human input. Humans may utilize either a unimodal or multimodal approach to provide feedback. Furthermore, the agent has the option to learn from a variety of human social input sources, or from both environmental incentives and human feedback. In several difficult reinforcement learning tasks, including RL benchmarking domains [23], [24], Atari games [25], simulation robotic control [8], [26], and actual robot navigation [27], these evaluated techniques have shown encouraging outcomes.

## II BACKGROUND

All of the algorithms presented in this work are based on reinforcement learning, which is initially described in this section. The next step is to provide interactive reinforcement learning, in which agents get input from human trainers and use it to improve their performance.

### A. REINFORCEMENT LEARNING

Agents may learn to tackle sequential decision-making issues using the Reinforcement Learning framework [2]. The tuple , where S is a collection of states and A is an action set, may be used to express a sequential decision problem as an MDP. The likelihood of a transition is denoted by T. The TV show! In the interval [0, 1], the reward function R is defined as R V S AS!. <. The discount factor, which controls the present value of future benefits, is 2 [0, 1]. A policy, denoted as V S A, represents the learnt behaviour of the agent. In this policy, the probability of choosing an action, denoted as ~(s; a) D Pr(at D ajst D s), is defined for each state *s*. The agent's goal is to maximize the predicted cumulative

ISSN: 2366-1313

reward in order to learn the best policy. There are three common ways to classify RL algorithms: policy search, value function, and actor-critic. Policy search algorithms absorb policies at their own pace. In order to generate the policy, value function techniques estimate the state value function V~ (s) and the action value function Q~ (s; a). Both the policy and value functions are simultaneously learned via actor-critical approaches. There is room for optimization and approximation in the policy and value functions. Function approximation is a common use case for deep neural networks in deep RL. Figure 1 displays the conventional RL framework.

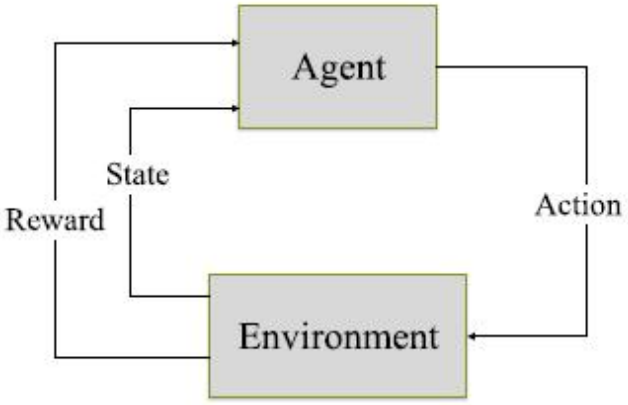
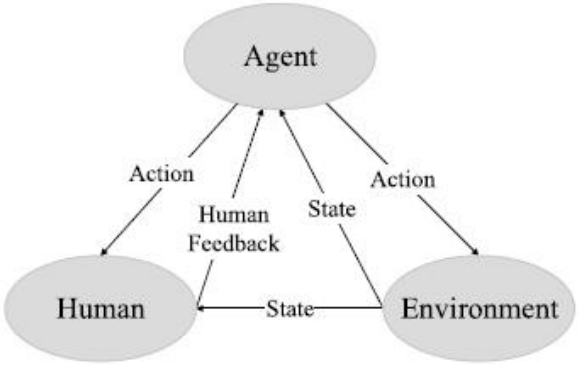



Fig 1:Illustration of an agent learning with standard reinforcement learning (adapted from [2]).

B. INTERACTIVE REINFORCEMENT LEARNING

Figure 2 shows that one possible solution to the sample efficiency issue in RL and deep RL is interactive reinforcement learning, which is inspired by potential-based reward shaping [28]. At the same time, even people who aren't professionals in agent programming or design may teach an interactive RL agent new things. A user's or instructor's assessment of the agent's performance, or evaluative feedback, is the source of knowledge for an agent in interactive reinforcement learning. Many interactive RL algorithms are the product of various interpretations of evaluative feedback, such as a comment on the agent's behaviour based on the expected agent policy in the human trainer's mind or the agent's own policy, a discrete categorical feedback strategy, a numerical reward, etc. Standard RL agent learners may also benefit from human counsel and teaching, and agents can learn to obey instructions by acquiring a reward function directly from humans. from [16]



ISSN: 2366-1313

to [19].

Fig2: Interactive reinforcement learning framework.

## III INTERACTIVE REINFORCEMENT LEARNING FROMHUMAN FEEDBACK

Similar to traditional RL, there are two primary types of interactive RL algorithms found in the literature: those that rely on models and those that do not. As with traditional RL, all existing model-based approaches to interactive RL based on human input are reward-based, meaning they treat human feedback numerically. While approaches that do not rely on models might be either policy- or reward-based. Human input is evaluated in terms of the agent's policy via policy feedback in policy-based approaches. Within each category, we can further classify agents as either learning a value function alone or learning both the value function and policy simultaneously. Here we'll go over the characteristics of algorithms in both categories and provide some examples of algorithms that are applicable from the literature. In this little supplement, we'll talk about and go into more detail on how you learn from both people and the environment.

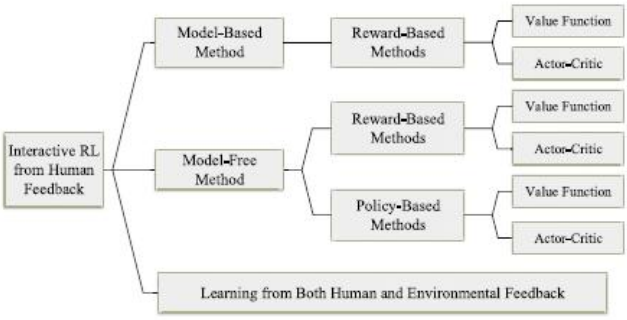



Fig 3: A diagrammatic representation of classification of learningmethods from human feedback.

### A. MODEL-BASED METHOD

Since model-based approaches may enhance learning speed and lower the quantity of interactions required for learning once the environment model is generated, they are typically regarded to be sample efficient. The TAMER framework, first developed by Knox and Stone [9], uses an estimated reward function to learn and choose behaviours. The human instructor in TAMER keeps an eye on the agent's actions and may provide feedback based on how well it does. The TAMER agent learning

ISSN: 2366-1313

framework consists of three main modules:

1) a model that predicts human reward based on the agent's past state-action pairs and the reward instances given by the trainer; 2) a mechanism to handle the time lag between evaluating the agent's behaviour and actually giving the reward; 3) a mechanism to choose actions that are associated with the predictive reward function. Because the human trainer considers the agent's actions in the context of their long-term consequences, TAMER agents are able to learn from short-sighted human rewards [9]. A TAMER agent might learn non-myopically from human reward with the help of VI-TAMER, according to another proposal [29]. A VI-TAMER agent models human incentives and learns from discounted versions of those benefits. A value function is learnt from the learned human reward function using value iteration. Then, actions are selected using the value function in order to maximize the cumulative discounted human reward. Even with dynamic programming and Monte Carlo tree search strategies for planning, the VI-TAMER agent may update the value function. By suggesting actor-critic TAMER, Vien and Ertel expanded the TAMER framework to educate agents in continuous state and action domains [30]. The agent masters both the critic's human reward function and the actor's parametrized policy for action selection in actor-critic TAMER. By approximating the reward function with a deep neural network, deep TAMER was introduced by [25] to tackle complicated issues in high-dimensional state space. In [19], an adversarial learning technique teaches an agent to obey language-based instructions by creating a reward function based on the difference between a predetermined set of instruction pairs and the ones created by the current policy.

The aforementioned approaches are all reward-based, meaning they use human input as a numerical reward, and they all look at using that feedback to represent the human reward function. When humans in the training role become weary of giving feedback, this comes in handy. Here, we may learn using the reward function that has already been taught. By including both the learned reward function and the known transition function into its planning, VI-TAMER is able to enhance its learning capabilities beyond only modeling the reward function. Even in the absence of a transition function, the agent may nevertheless learn the concept of

ISSN: 2366-1313

transition via its interactions with the environment and a human trainer. Furthermore, the majority of the aforementioned approaches are limited in their applicability since they can only learn in tasks with discrete action spaces, with the exception of actor-critic TAMER. It would be very helpful to apply these ideas to projects with continuous actions because the action space for many real-world activities is continuous. Also, actor-critic TAMER may be trained independently to learn the state representation using deep learning in a high dimensional state space.

## B. MODEL-FREE METHOD

Even when it's challenging to model the environment's and the trainer's reward function and transition, the agent may nevertheless learn model-free by analyzing human-provided feedback. The vast majority of RL interactive approaches that rely on human evaluations actually do not use models. They may be classified into two groups: policy-based approaches and reward-based methods, according on how human input is understood.

### 1. REWARD-BASED METHODS

As with traditional RL, human input is treated numerically in reward-based interactive RL approaches. An agent may also learn from seeing human reward, without the need to simulate the reward function. The first notion to teach an agent using just positive rewards was, as far as we are aware, Clicker training [31]. Cobot, the first software agent, applies reinforcement learning in a text-based virtual environment where humans interact, learning from both rewards and punishments [4]. Through repeated invocations of "reward and punish" text-verbs, the agent learns to initiate conversational activities (such as suggesting a subject) via voice. Similar to how environmental incentives are used in classical reinforcement learning [33], [34], human rewards may also be used to train a Q-value function [32]. Alternatively, the agent may optimize the policy using a function approximator, allowing it to learn the policy directly. In order to train a virtual upper-arm robotic prosthesis to respond optimally to reward signals given by humans, Pilarski et al. [8] developed a continuous action actor-critic reinforcement learning system [35].

### 2. POLICY-BASED METHODS

ISSN: 2366-1313

An agent may gain knowledge from human input in two ways: first, by seeing it as a numerical reward, and second, by viewing it as policy feedback. Under these circumstances, the human input is interpreted as an assessment of the agent's performance. Reference [12] substitutes the agent's present policy with human feedback, seeing it as policy-dependent. The reference function explains the relative merits of different action selections in comparison to the anticipated behaviour. An impartial approximation of the benefit function is the Temporal Difference (TD) in conventional reinforcement learning. The advantage function is a better fit for a declining returns strategy; that is, given a positive chance of choosing action an in state *s*, the initial human feedback for taking action an in state *s* will be positive, but it will soon be zero. By integrating human feedback into an actor-critic algorithm to determine the policy gradient, they came up with the COACH algorithm. In order to make COACH even more advanced, Arumugam et al. [36] used a deep neural network to approximate policy functions. This allowed them to create deep COACH. Instead of translating feedback signals into monetary incentives, the authors of reference [37] suggest "policy shaping" via formalizing human feedback as a label on the optimality of acts and using it as policy guidance. Also, trainer-targeted behaviour and trainer-specific instructional approach are two factors that influence the meaning of human feedback, according to [11]. When given no explicit instructions, they deduced the required action. Loftin et al.'s algorithms were able to learn more quickly than algorithms that see feedback as a numerical reward, according to their experimental findings. Human input interpretation is still up for discussion. Indeed, several trainers may assign diverse meanings to the same human feedback, particularly if they use distinct task-specific interpretations of the instructions [38]. They may even decide to switch up the training approach as time goes on.

## IV FEEDBACK SOURCE

Recent efforts in the field of Human Robot Interaction (HRI) have centered on creating robots with the ability to recognize typical human communication signals in order to facilitate more organic interactions. One branch of human-robot interaction (HRI) known as "social HRI" involves robots that mimic human speech, facial emotions, and body language in

ISSN: 2366-1313

their interactions with one another. This paves the way for people to engage with robots in a way that doesn't need a ton of training, which in turn speeds up the completion of desired activities while reducing the amount of effort required from the human user [42]. Based on the previous discussion, we will primarily examine the feedback sources' viewpoint on robot-human interaction (Figure 4).

A. UNIMODAL SENSORY FEEDBACK

According to this research, there is only one way in which human trainers may impart knowledge to their interactive RL agents. Under these conditions, human input may be sent by physical means like keyboard presses, mouse clicks, etc., or through more intangible means like gestures, facial expressions, and natural languages.

As part of interactive RL, human trainers can mentally construct feedback and consciously communicate it to agents using hardware facilities, such as keyboard keys, mouse clicks (slider or bar), or other sensors [9, 11–45]. While this kind of detailed feedback is great for teaching agents good policies, it may be delayed due to human trainers' response times, leaving agents particularly those with a lot of actions in the dark as to which acts the trainer is referring to. To address this issue, Knox and Stone suggested a credit assign method that uses a probability density function to predict the likelihood of a delay in the teacher's response [23]. On the other hand, certain trainers may have vastly varied delays. Also, before teaching the robots, the trainers need to have a feel for the hardware, and most studies include a practice session so that trainers may practice providing feedback. Additionally, in a home-like setting, these interfaces are very cumbersome and unpractical for trainers who are not experts. Consequently, it would be great if robots and trainers could create more natural communication interfaces, similar to how a caretaker teaches a newborn, utilizing things like voice, emotions, and gestures.

Interactions that occur in nature

A natural encounter may offer implicit feedback that an interactive RL agent can use to learn, rather than the purposeful explicit input given by human trainers. Training agents using

ISSN: 2366-1313

natural feedback will be very beneficial and significant, particularly for long-term behavior learning with interactive RL, as it will help to prevent cognitive fatigue induced by delivering explicit feedback. In order to tailor the interaction experience for users with varying skills, it is possible, for instance, to extract facial expressions as evaluation input. Undirected human criticism that does not seek to instruct or influence behavior—perhaps comes from more social indicators, including as smiles, attentiveness, and tone of voice, are transmitted and may be viewed without imposing any cognitive burden on the person [23]. In an ideal world, human trainers would be able to provide feedback in a way that is as natural as real-life human-to-human communication, including via emotions, natural languages, gestures, etc.

a: The Face's Reactions

By integrating an emotion system with conventional reinforcement learning, Gadanho put out the idea of an emotion-based architecture, or EB architecture. As a kind of social reinforcement, the emotion system determines a value for well-being. Using Q-learning, the EB design can figure out when to swap gears and reward good behavior [46]. By presenting the EARL framework, Broekens investigated the interplay of Emotion, Adaptation, and Reinforcement Learning [47]. In EARL, a "social robot" was trained using real-time analysis of human emotional expressions as extra social reinforcement signals. A robot taught without social reinforcement learns far more slowly than one trained with emotive facial expressions, according to their findings. To enable agents to learn a value function that correlates camera-extracted face traits with anticipated future reward, Veeriah et al. suggested [48]. Based on their first findings, it seems that agents may learn user preferences and provide less explicit feedback when choosing a grip. As a reward signal for the robot's emotional reinforcement learning, Gordon et al. used a completely autonomous social robotic learning companion for affective child-robot tutoring. The robot was trained by measuring the children's valence and participation using an automated facial expression analysis system

ISSN: 2366-1313

[49]. They test their method for two months with a group of thirty-four preschoolers. Based on their findings, it seems that the robot may adapt its motivating tactics to suit the needs of individual students by using both verbal and non-verbal cues. As an additional kind of implicit human reward, Arakawa et al. [41] trained a DQN-TAMER bot using camera-captured face expressions.

For example, "happy" would be considered positive feedback (1) and "angry" would be considered negative feedback (1) in the aforementioned study, which used facial expressions to teach the agent. On the other hand, throughout training, the relative importance of good and negative emotions might change. Utilizing the acquired data, Li et al. constructed a prediction model that mapped the facial feedback to explicit keypress feedback. When it comes to recognition accuracy, their simulation experiment shown that agents can learn just as well from face input as they would from explicit keypress feedback [50].

b: LINGUAGE-FRAMEWORK INFORMATIVE

Natural language education and advise is an intuitive and promising method for training agents to complete a job, particularly for non-technical users, when autonomous agents learn from human users. A Q value function that incorporates programming language-based advise delivered by an external observer was first presented in Reference [16] as the RATLE (Reinforcement and advise-Taking Learning Environment) system. In order to impact the agent's learning policy, the guidance in reference [17] was formalized from English-based natural language. The LEARN (languageE-Action Reward Network) architecture, suggested in reference [18], converts unstructured instructions given in plain language into intermediate-shaped rewards depending on the agent's actions. And to teach a genuine autonomous mobile robot to navigate in a simulated environment, Tenorio et al. [26] employed vocal instructions based on preset natural languages to provide human evaluative feedback. Their experimental findings demonstrate that, in contrast to conventional reinforcement learning based only on environmental rewards, quicker convergence was attained when human incentives supplied verbally are considered to be loud.

ISSN: 2366-1313

As an alternative to relying on natural language feedback to assist RL agents in learning from environmental rewards, agents may learn policies directly from instructions based on natural language. In an object-oriented MDP framework, language was mapped to a reward function in reference [51]. One study that learned a strategy for instruction execution in a contextual bandit scenario employed raw visual observations and text input based on natural language [52]. A reward function is constructed by distinguishing a predefined set of instruction pairs from the instruction pairs generated by the present policy. This framework is suggested in [19] as an adversarial learning approach to enhance policy learning.

i: Gestural Input

Hand and body movements are common forms of nonverbal communication between humans, particularly in situations when spoken language is either not permitted or is not understood. Consequently, it is possible to train agents using human gestures as input. In their study, Kuno et al. demonstrated how an intelligent wheelchair could be controlled using hand gestures. They also suggested a way to identify unfamiliar hand motions via user interaction [53]. With the goal of teaching a robot.

developers, Voyles and Khosla advocated for the use of gesture-based programming techniques to teach robots, bypassing the need for programmers altogether [54]. Another way that gestures might help RL agents learn is by giving them feedback in the form of advise or commands [55], [56].

B. Input from Multiple Sensors

All of the aforementioned methods and systems for detecting human emotion rely on identifying a particular input modality. There are two primary benefits to using multimodal inputs instead of a single input: first, a multimodal recognition system can estimate using the remaining modalities in the event that one modality is unavailable owing to noise or occlusion; and second, feedback with improved robustness and performance can be provided by the complementary and diverse information provided by multiple modalities.

A multimodal interaction paradigm for discourse segmentation in free-form gesticulation accompa-nying speech in natural conversation was presented by Quek et al. [57] to get an understanding of the interaction between speech and

ISSN: 2366-1313

gesture and how they facilitate communication. Cruz et al. combine interactive reinforcement learning with dynamic multimodal audiovisual interaction [56]. Agents may be instructed by human trainers using voice, gestures, or a mix of the two, according to specified guidelines. As opposed to unimodal situations, their findings demonstrate that multimodal integration allows the robot to achieve greater performance with a less number of training episodes via interactive reinforcement learning. As an example of implicit social evaluative feedback, references [58], [59] used the audience's verbal and visual grins to determine the reward and influence the robot's comedic performance. Leite et al. [60] adopt a multimodal framework to characterize the user's emotional states and enable the robot to modify its empathic reactions to the individual preferences of the kid engaging with it, endowing a chess companion robot for children with empathy skills. They assess the user's emotional state by integrating visual and task-related characteristics. An method for reinforcement learning with many arms uses the difference in valence before and after the robot adopts an empathetic stance to determine its rewards. Results from their pilot research with 40 kids demonstrate that people are positively affected by robots that exhibit empathetic behaviour.

The majority of the aforementioned approaches to agent training rely on a combination of only two modal inputs, and even then, those inputs are severely confined to visual emotions and task-related aspects, vocal laughter, and gestures. Attention, voice prosody, gaze direction, and other social signals from the human trainer may also be used as feedback. Plus, the trainer may provide feedback via more than two modalities. Furthermore, agents may be trained using a combination of hardware-delivered feedback and this natural interactive feedback. As an example, Li et al. [50] suggested that an agent may learn from both the anticipated facial feedback and the explicit keystroke feedback by mapping the expressions to the keys.

## V CONCLUSION AND FUTURE DIRECTIONS

In this article, we will take a look back at how far we've come in using various forms of human social input to solve RL problems. Several interesting avenues for further study are briefly covered in this section.

ISSN: 2366-1313

A. GAINING INFORMATION FROM IMPLICIT NATURE

A common issue with interactive RL is the lack of a naturally occurring interface for trainer-robot feedback in household scenarios. Most of them relied on clicking buttons or using the mouse to provide feedback, which is very laborious and unpractical for trainers who aren't experts in home-like settings. Although there have been studies looking at the use of nonverbal cues like facial expressions, voice, and gestures to provide agents feedback and guidance during training, most of the time trainers deliver this feedback with a specific goal in mind. Implicit feedback for agent learning may be acquired from human social cues such as smiles, speech, attentiveness, prosody, and other more widely broadcast signals, which do not impose any cognitive burden to the human [23]. Finding strategies to let robots learn from these unstructured, conveyed implicit feedbacks is an open topic. One example is the detection of affect and emotion in conversational prosody [62], [63] and speech itself [61].

B. The Design of Interactions

Looking at it from a robotics standpoint, knowing how to program the trainer-robot relationship is key to developing algorithms that let humans teach well while being fully present during training. Customizing interactions with socially supportive robots may benefit from this. As part of the transparent learning process [64]-[66], the robot communicates its learning status and requests feedback from the human instructor using body language and facial expressions. A issue that has to be researched further is what the robot should say or do in order to get training that lasts longer or is of better quality.

Chapter 5: Using Various Methods of Instruction Evaluative feedback and advice/instruction learning are the primary foci of the literatures surveyed in this article. Even with conventional RL learning paradigms, a completely autonomous interactive RL agent requires algorithms that can learn from human demonstrations, evaluations, and advice/instruction. There has been a lot of effort to integrate normal RL with learning via demonstration[67], evaluative feedback[37], and advise [18]. The same instructional modalities that human instructors depend on—demonstration, verbal advice/instruction,

ISSN: 2366-1313

evaluative feedback, attentional signals, and gestures—are essential for agents to learn as well. There is some prior work that enables a robot to learn from examples and the instructor's natural feedback signals given verbally [69], but there is still a lot of room for improvement in this area, notwithstanding the lessons learned via human demonstration and evaluation feedback [70].

## REFERENCES


[1] J. Kober, J. A. Bagnell, and J. Peters, ‘‘Reinforcement learning in robotics: A survey,’’ *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1238–1274, Sep. 2013.

[2] C. Isbell, C. R. Shelton, M. Kearns, S. Singh, and P. Stone, ‘‘A social reinforcement learning agent,’’ in *Proc. 5th Int. Conf. Auto. Agents*, 2001,pp. 377–384.

[3] R. Maclin, J. Shavlik, L. Torrey, T. Walker, and E. Wild, ‘‘Giving advice about preferred actions to reinforcement learners via knowledge-based kernel regression,’’ in *Proc. Nat. Conf. Artif. Intell.* Cambridge, MA, USA: MIT Press, 1999, 200, p. 819.

[4] A. L. Thomaz and C. Breazeal, ‘‘Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learn- ing performance,’’ in *Proc. AAAI*, vol. 6. Boston, MA, USA, 2006, pp. 1000–1005.

[5] B. Argall, B. Browning, and M. Veloso, ‘‘Learning by demonstration with critique from a human teacher,’’ in *Proc. ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, 2007, pp. 57–64.

[6] P. M. Pilarski, M. R. Dawson, T. Degris, F. Fahimi, J. P. Carey, and R. S. Sutton, ‘‘Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning,’’ in *Proc. IEEE Int. Conf. Rehabil. Robot.*, Jun. 2011, pp. 1–7.

[7] W. B. Knox and P. Stone, ‘‘Interactively shaping agents via human rein- forcement: The TAMER framework,’’ in *Proc. 5th Int. Conf. Knowl. Cap- ture*, 2009, pp. 9–16.

[8] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, ‘‘A survey of robot learning from demonstration,’’ *Robot. Auto. Syst.*, vol. 57, no. 5, pp. 469–483, May 2009.

[9] R. Loftin, B. Peng, J. MacGlashan, M. L. Littman, M. E. Taylor, J. Huang, and D. L. Roberts, ‘‘Learning behaviors via human-delivered discrete feedback: Modeling implicit feedback strategies to speed up learning,’’ *Auto. Agents Multi-Agent Syst.*, vol. 30, no. 1, pp. 30–59, Jan. 2016.

[10] J. MacGlashan, M. K. Ho, R. Loftin, B. Peng, G. Wang, D. L. Roberts,
M. E. Taylor, and M. L. Littman, ‘‘Interactive learning from policy-dependent human feedback,’’ in *Proc. 34th Int. Conf. Mach. Learn.*, Vol. 70, 2017, pp. 2285–2294.

[11] H. Lieberman, *Your Wish is my Command: Programming by Example*. San Mateo, CA, USA: Morgan Kaufmann, 2001.

ISSN: 2366-1313

[12] P. Abbeel and A. Y. Ng, ''Apprenticeship learning via inverse reinforce- ment learning,'' in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, p. 1.

[13] J. Luketina, N. Nardelli, G. Farquhar, J. Foerster, J. Andreas, E. Grefen- stette, S. Whiteson, and T. Rocktäschel, ''A survey of reinforcement learn- ing informed by natural language,'' 2019, *arXiv:1906.03926*. [Online]. Available: http://arxiv.org/abs/1906.03926

[14] R. Maclin and J. W. Shavlik, ''Creating advice-taking reinforcement learn- ers,'' *Mach. Learn.*, vol. 22, nos. 1–3, pp. 251–281, 1996.

[15] G. Kuhlmann, P. Stone, R. Mooney, and J. Shavlik, ''Guiding a rein- forcement learner with natural language advice: Initial results in RoboCup soccer,'' in *Proc. AAAI Workshop Supervisory Control Learn. Adapt. Syst.*, San Jose, CA, USA, 2004, pp. 1–6.

[16] P. Goyal, S. Niekum, and R. J. Mooney, ''Using natural language for reward shaping in reinforcement learning,'' 2019, *arXiv:1903.02020*. [Online]. Available: http://arxiv.org/abs/1903.02020

[17] D. Bahdanau, F. Hill, J. Leike, E. Hughes, A. Hosseini, P. Kohli, and E. Grefenstette, ''Learning to understand goal specifications by modelling reward,'' 2018, *arXiv:1806.01946*. [Online]. Available: http://arxiv.org/abs/1806.01946

[18] F. Benavent and B. Zanuttini, ''An experimental study of advice in sequen- tial decision-making under uncertainty,'' in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–9.

[19] G. Li, R. Gomez, K. Nakamura, and B. He, ''Human-centered reinforce- ment learning: A survey,'' *IEEE Trans. Human-Machine Syst.*, vol. 49, no. 4, pp. 337–349, Aug. 2019.

[20] F. Torabi, G. Warnell, and P. Stone, ''Recent advances in imitation learning from observation,'' 2019, *arXiv:1905.13566*. [Online]. Available: http://arxiv.org/abs/1905.13566

[21] W. B. Knox, ''Learning from human-generated reward,'' Ph.D. disserta- tion, Dept. Comput. Sci., Univ. Texas at Austin, Austin, TX, USA, 2012.

[22] G. Li, ''Socially intelligent autonomous agents that learn from human reward,'' Ph.D. dissertation, Inform. Inst., Univ. Amsterdam, Amsterdam, The Netherlands, 2016.

[23] G. Warnell, N. Waytowich, V. Lawhern, and P. Stone, ''Deep tamer: Interactive agent shaping in high-dimensional state spaces,'' in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–9.

[24] A. C. Tenorio-Gonzalez, E. F. Morales, and L. Villasenor-Pineda, ''Dynamic reward shaping: Training a robot by voice,'' in *Proc. Ibero-Amer. Conf. Artif. Intell.* Berlin, Germany: Springer, 2010, pp. 483–492.

[25] W. B. Knox, P. Stone, and C. Breazeal, ''Training a robot via human feed- back: A case study,'' in *Proc. Int. Conf. Social Robot.* Cham, Switzerland: Springer, 2013, pp. 460–470.

[26] A. Y. Ng, D. Harada, and S. Russell, ''Policy invariance under reward transformations: Theory and application to reward shaping,'' in

ISSN: 2366-1313

*Proc. ICML*, vol. 99, 1999, pp. 278–287.

[27] W. B. Knox and P. Stone, ‘‘Framing reinforcement learning from human reward: Reward positivity, temporal discounting, episodicity, and perfor- mance,’’ *Artif. Intell.*, vol. 225, pp. 24–50, Aug. 2015.