IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 20, 2020, accepted June 26, 2020, date of publication July 2, 2020, date of current version July 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3006254

A Review on Interactive Reinforcement Learning From Human Social Feedback

JINYING LIN[1], ZHEN MA[1], RANDY GOMEZ[2], (Member, IEEE), KEISUKE NAKAMURA[2], (Member, IEEE), BO HE[1], (Member, IEEE), AND GUANGLIANG LI[1], (Member, IEEE)

[1]Department of Electronic Engineering, Ocean University of China, Qingdao 266100, China
[2]Honda Research Institute Japan Company Ltd., Wako 351-0188, Japan

Corresponding author: Guangliang Li (guangliangli@ouc.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 51809246, and in part by the Natural Science Foundation of Shandong Province under Grant ZR2018QF003.

ABSTRACT Reinforcement learning agent learns how to perform a task by interacting with the environment. The use of reinforcement learning in real-life applications has been limited because of the sample efficiency problem. Interactive reinforcement learning has been developed to speed up the agent's learning and facilitate to learn from ordinary people by allowing them to provide social feedback, e.g, evaluative feedback, advice or instruction. Inspired by real-life biological learning scenarios, there could be many ways to provide feedback for agent learning, such as via hardware delivered, natural interaction like facial expressions, speech or gestures. The agent can even learn from feedback via unimodal or multimodal sensory input. This paper reviews methods for interactive reinforcement learning agent to learn from human social feedback and the ways of delivering feedback. Finally, we discuss some open problems and possible future research directions.

INDEX TERMS Human agent/robot interaction, interactive reinforcement learning, interactive shaping, social interaction.

I. INTRODUCTION

Reinforcement learning (RL) has achieved remarkable successes in many practical problems [1], [2]. With recent advances in deep learning, RL has attracted more attention and been combined as deep RL to solve end-to-end learning in sequential decision tasks [3]. However, the problem of sample efficiency has largely limited the application of RL and deep RL to real-life situations. For example, it might take an RL agent millions of training samples for learning a good policy to play a video game [3]. In practice, RL and deep RL will be mostly applied to robots or agents operating in human living environments. The interaction between agent and human is essential and will increase as well. Therefore, enormous knowledge and experience from human users could be used to guide the agent's learning.

There are many ways that human trainers can direct an agent to learn, such as by providing demonstration, instruction/advice, and evaluative feedback [4]–[12]. A human user can provide a demonstration to an agent by remote control or by his own body [10], [13]. One form of learning from demonstration is inverse reinforcement learning [14], in which an agent optimizes the policy by learning a reward function from provided demonstrations. And demonstrations are mostly used for initializing the agent's policy or solving the RL task with one time interaction using inverse RL. However, in complex task domains, it might be very difficult for non-expert human trainers to provide high-quality demonstrations.

Another way that human trainers direct agents to learn is to provide instruction or advice via natural languages [15]. When learning from advice, the advice usually needs to be encoded into a programming language or mapped from natural language to a formal language, which can be used to improve the reinforcement agent learning [16], [17]. The agent can also learn from instruction by mapping free-form natural language instructions to intermediate shaping rewards [18] or learn to follow language instructions by learning a reward function from them [19]. The effects of different types of advice such as optimal action advice and optimal gain-risk advice on the agent's learning performance are investigated [20]. The human user can also provide evaluative feedback to train the agents. Agent learning from human evaluative feedback is termed human-centered reinforcement

The associate editor coordinating the review of this manuscript and approving it for publication was Pedro Neto.

learning [21]. The interpretation of evaluative feedback can be different, such as numeric reward, discrete categorical feedback or policy feedback, resulting in different learning algorithms. However, in most studies human feedback are provided via button presses or mouse clicks. Inspired by real-life biological learning scenarios, they could deliver the feedback more naturally via emotions, gestures, or even natural languages to train the agent. The agent can even learn from both these naturally delivered evaluative feedback and other social feedback, like demonstration, advice and instruction.

Since there are already some survey papers on learning from demonstration and observation [10], [22], the objective of this paper is to investigate the most recent work on using different human social feedback (evaluative feedback, advice/instruction) to train agents to solve reinforcement learning tasks. The methods on learning from human feedback can be model-based or model-free as in traditional RL. The way of providing feedback by human users can be unimodal or multimodal. In addition, the agent can learn from both human feedback and environmental rewards, or from different sources of human social feedback. Encouraging results have been shown by these reviewed approaches in one or more challenging reinforcement learning tasks, such as RL benchmarking domains [23], [24], Atari games [25], simulated robotic control [8], [26] and real robot navigation [27].

II. BACKGROUND

In this section, we first describe reinforcement learning, which constitutes the foundation of all the algorithms presented in this paper. We then introduce interactive reinforcement learning, where agent learn from feedback provided by human trainers.

A. REINFORCEMENT LEARNING

Reinforcement Learning [2] is a framework in which agents learn to solve sequential decision-making problems. A sequential decision problem can be modeled as an Markov decision process (MDP), represented by a tuple $< S, A, T, R, \gamma >$, where S represents a set of states and A represents an action set. T is the transition probability function $T : S \times A \times S \rightarrow [0, 1]$, R is the reward function $R : S \times A \times S \rightarrow \Re$. $\gamma \in [0, 1]$ is the discount factor, determining the present value of rewards received in the future. The agent's learned behavior is represented by a *policy*, $\pi : S \times A$, where $\pi(s, a) = Pr(a_t = a|s_t = s)$ is the probability of selecting a possible action $a \in A$ in a state s. The objective of the agent is to learn an optimal policy π^* by maximizing the expected cumulative reward. RL algorithms are usually divided into three categories: policy search methods, value function methods and actor-critic methods. Policy search methods learn the policy directly. Value function methods estimate the value function—state value function $V^\pi(s)$ and action value function $Q^\pi(s, a)$, and derive the policy from it. Actor-critic methods learn the policy and value function at the same time. The policy and value function can be approximated and optimized. In deep RL, deep neural networks are usually used as function approximation. The standard RL framework can be shown in Figure 1.

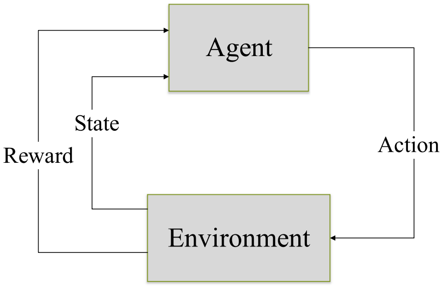

FIGURE 1. Illustration of an agent learning with standard reinforcement learning (adapted from [2]).

B. INTERACTIVE REINFORCEMENT LEARNING

Inspired by potential-based reward shaping [28], interactive reinforcement learning is proposed as one solution to the sample efficiency problem in RL and deep RL (Figure 2). Meanwhile, an interactive RL agent can also learn from a human observer, especially non-experts in agent design and programming. In interactive reinforcement learning, an agent learns from human evaluative feedback, i.e., evaluations of the quality of the agent's behavior provided by a human user, or advice/instruction. The interpretations of evaluative feedback can be different, e.g., a comment on the agent's behavior based on the expected agent policy in the human trainer's mind or the policy the agent is following, discrete categorical feedback strategy, numeric reward etc., which result in many interactive RL algorithms. The human advice/instruction can also be used to aid a standard RL agent learner or an agent learns how to follow instructions by directly learning a reward function from them [16]–[19].

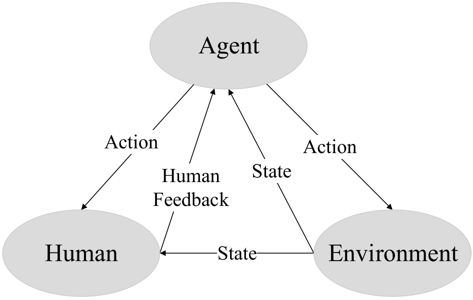

FIGURE 2. Interactive reinforcement learning framework.

III. INTERACTIVE REINFORCEMENT LEARNING FROM HUMAN FEEDBACK

As in standard RL, interactive RL algorithms in the literature are mainly divided into two categories: (1) model-based algorithms, and (2) model-free algorithms. All current model-based methods for interactive RL from human feedback are reward-based methods, which take human feedback

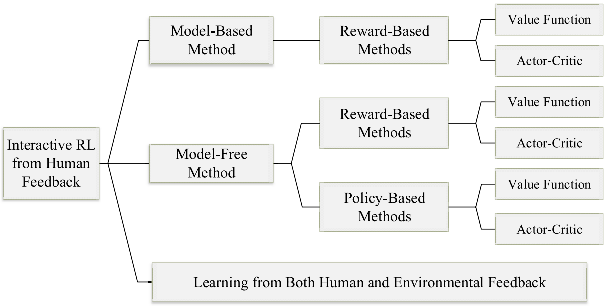

FIGURE 3. A diagrammatic representation of classification of learning methods from human feedback.

as numeric reward as in standard RL. While model-free methods can be reward-based and policy-based methods. Policy-based methods take human feedback as policy feedback which is an evaluation on the agent's policy. In each category, according to whether an agent learns a value function alone or both value function and policy separately at the same time, we can also group them into: value function method and actor-critic method. In the following, we will discuss the features of algorithms in both groups and present relevant example algorithms from the literature. As a small supplement, we will discuss and elaborate on ways of learning from both human and environmental feedback.

A. MODEL-BASED METHOD

Model-based methods are generally considered to be sample efficient, since the learning speed can be improved and the amounts of interactions needed for learning can be decreased once the model of the environment is obtained. Knox and Stone first proposed the TAMER framework [9] which learns and selects actions with an estimated reward function. In TAMER, the human teacher observes the agent's behavior and can give reward corresponding to its quality. There are three key modules for an agent learning with TAMER: 1) a predictive model of human reward from the agent's experienced state-action pairs and the reward instances provided by the human trainer; 2) a credit assigner to deal with the time delay of human reward caused by evaluation of the agent's behavior and delivering it; 3) an action selector with the predictive reward function. The reason why a TAMER agent can learn from myopic human reward is that the human trainer already takes a long-term consequence of the agent's behavior in mind when she is providing the evaluation [9]. VI-TAMER was further proposed to allow a TAMER agent learning non-myopically from human reward [29]. A VI-TAMER agent learns from the discounted human rewards while modeling the human rewards. It learns a value function from the learned human reward function via value iteration and select actions with the value function to get the most accumulated discounted human reward. The VI-TAMER agent can even update the value function via planning with dynamic programming or Monte Carlo tree search strategy. Vien and Ertel extended the TAMER framework to train agents in continuous state and action domains by proposing actor-critic TAMER [30]. In actor-critic TAMER, the agent learns a human reward function — the critic, and a parametrized policy to select actions — the actor, at the same time. To solve complex problems with high-dimensional state space, [25] proposed deep TAMER by using deep neural network to approximate the reward function. In [19], an agent learns to follow language-based instruction by generating a reward function via distinguishing a fixed set of instruction pairs from instruction pairs generated by the current policy with adversarial learning method.

All above methods take human feedback as numeric reward — reward-based methods, and consider to model the human reward function from provided human feedback. This is useful when human trainers get tired of providing feedback for further training. In this case, the learned reward function can be used for learning. In addition to modeling the reward function, VI-TAMER can improve its learning by planning with a known transition function and learned reward function. When the transition function is not available, the agent can also learn the transition model from the interaction with the environment and human trainer. Moreover, except for actor-critic TAMER, most above methods can only learn in tasks with discrete action space which limits their applications. Since the action space for many tasks in the real world is continuous, it would be immensely useful to extend these methods to tasks with continuous actions. Furthermore, for learning in high dimensional state space, actor-critic TAMER can be powered with deep learning to learn the state representation autonomously.

B. MODEL-FREE METHOD

When it is difficult to model the reward function and transition from the environment and human trainer, the agent can also learn from human provided feedback in a model-free manner. Actually most interactive RL methods from human evaluative feedback are model-free methods. Depending on the different interpretations of human feedback, they can be grouped into two categories: reward-based methods and policy-based methods.

1) REWARD-BASED METHODS

In reward-based interactive RL methods, human feedback is taken as numeric reward as in standard RL. Instead of modeling the reward function, an agent can also learn directly from human reward. To our knowledge, *Clicker training* was the first proposed concept using only positive reward to train an agent [31]. The first software agent called Cobot learns from both reward and punishment by applying reinforcement learning in an online text-based virtual world where people interact [4]. The agent learns to take proactive verbal actions (e.g. proposing a topic for conversation) from 'reward and punish' text-verbs invoked by multiple users. A Q-value function [32] can also be learned by taking human rewards with the same way as environmental rewards in traditional reinforcement learning [33], [34]. The agent can also learn the policy directly by optimizing it with a function approximator. Pilarski *et al.* [8] proposed a continuous action actor-critic reinforcement learning algorithm [35] that learns an optimal control policy for a simulated upper-arm robotic prosthesis using only human-delivered reward signals.

2) POLICY-BASED METHODS

While reward-based methods interpret human feedback as a numeric reward, an agent can also learn from human feedback by taking it as policy feedback. In this case, the human feedback is taken as evaluation based on the agent's behavior. Reference [12] take human feedback as policy-dependent on the agent's current policy and use it to replace an advantage function which describes how much better or worse an action selection is compared to the current expected behavior. Temporal Difference (TD) in standard reinforcement learning is an unbiased estimate of the advantage function. The advantage function can better capture a diminishing returns strategy, which means the initial human feedback for taking the optimal action a in state s will be positive, but goes to zero as the probability of selecting action a in state s goes to 1. They proposed the COACH algorithm by using human feedback directly to calculate the policy gradient in an actor-critic algorithm. Arumugam *et al.* further extend COACH to deep COACH using deep neural network as function approximator for the policy [36]. Reference [37] propose 'policy shaping' by formalizing human feedback as a label on the optimality of actions and using it as policy advice, instead of converting feedback signals into numeric rewards.

In addition, [11] interpreted human feedback as discrete categorical feedback strategies that depend both on the behavior the trainer is trying to teach and the trainer's teaching strategy. They inferred knowledge about the desired behavior from cases where no feedback is provided. The experimental results of Loftin *et al.*'s work show that their algorithms could learn faster than algorithms that treat the feedback as a numeric reward. The debate over the interpretation of human feedback is not over yet. In fact, human feedback could be interpreted differently by different trainers especially when they interpret the instruction differently in the task [38]. They might even change the training strategy over time.

C. LEARNING FROM BOTH HUMAN AND ENVIRONMENTAL FEEDBACK

Learning solely from human feedback is useful when there is no objective measure in the task or it is difficult to define an effective reward function for the task. In this case, human trainers can use their feedback to customize the agent's behavior according to their expertise in the task and the agent's optimal behavior is solely decided by the human trainer. When the reward function of the task is available, it would be helpful for the agent to learn from both the environmental reward of the defined reward function and human feedback, especially when the environmental reward is very sparse. In this case, human feedback can be used to guide the agent's exploration and speed up its learning from environmental rewards. For example, in [17], reinforcement learners can learn from both the reinforcement provided by the environment and the human-generated advice. In [16], the advice was represented by creating new hidden nodes in a neural network for approximating the Q function. They showed an improvement in the agent performance compared to an agent learning without advice or making use of advice using naive technique. In [33], the agent learns by maximizing its total discounted sum of human reward and environmental reward. In addition, in the TAMER+RL framework, an agent can learn from both human and environmental rewards while modeling the reward function at the same time, which can lead to a better agent performance than learning from either alone. The agent can learn sequentially first from human evaluative feedback, then environmental reward [39] and from both rewards simultaneously, which allows the human teacher to provide evaluative feedback at any time during the training process [40]. Reference [41] further proposed the the DQN-TAMER framework by combining Deep Q-Network with deep TAMER. In DQN-TAMER, the agent estimates an action value function—$\hat{Q}$—from the environmental feedback and the reward function—$\hat{H}$—from human feedback. The final policy is obtained by weighted averaging the two policies from the DQN agent and TAMER agent trained in parallel.

In summary, while learning from both human feedback and environmental feedback provides a way for solving the exploration problem in reinforcement learning, it also enables an agent to learn according to the trainer's preference from her immediate feedback and a long-term behavior from the environmental feedback.

IV. FEEDBACK SOURCE

Recent research in Human Robot Interaction (HRI) has focused on developing robots that can detect common human communication cues for more natural interactions. Social HRI is a subset of HRI that encompasses robots which interact using natural human communication modalities, including speech, facial expressions and gestures like body language. This allows humans to interact with robots without any extensive prior training, permitting desired tasks to be completed more quickly and requiring less work to be performed by the human user [42]. From the above consideration, we will mainly study the way robots interact with humans from the perspective of feedback sources, as shown in Figure 4.

A. UNIMODAL SENSORY FEEDBACK

In the current study, most interactive RL agent learns from feedback communicated by the human trainer in one single mode. Under the circumstances, human feedback can be delivered via hardware such as keypress on the keyboard, mouse clicks etc, or via natural interactions like facial expressions, natural languages and gestures.

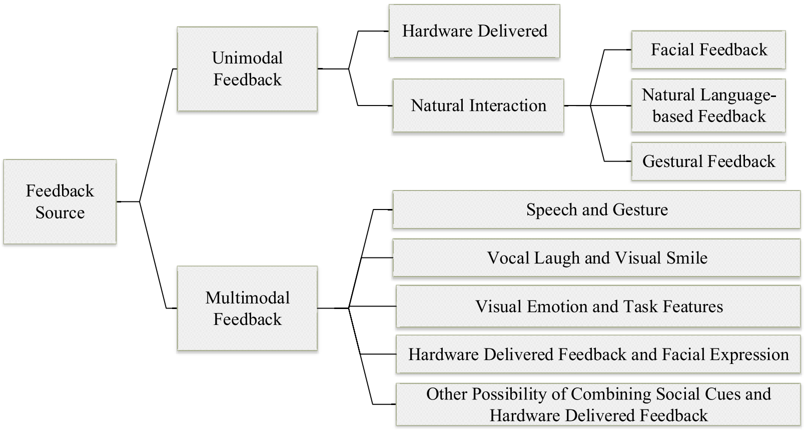
Feedback Source
Unimodal Feedback
Multimodal Feedback
Hardware Delivered
Natural Interaction
Facial Feedback
Natural Language-based Feedback
Gestural Feedback
Speech and Gesture
Vocal Laugh and Visual Smile
Visual Emotion and Task Features
Hardware Delivered Feedback and Facial Expression
Other Possibility of Combining Social Cues and Hardware Delivered Feedback

FIGURE 4. A diagrammatic representation of source of human feedback.

1) HARDWARE DELIVERED

In interactive RL, human trainers can form their feedback in the mind and then deliver the feedback intentionally and explicitly to agents via hardware facilities, mainly including but not limited to keyboard keys, mouse clicks with a slider or bar, or other sensors [9], [11], [12], [25], [36], [43]–[45]. Although this precise feedback can train agents to learn an effective policy, the reaction time of human trainers causes the delay in delivering their feedback, so the agent might be uncertain about which actions the human feedback is targeting at especially for agents with frequent actions. Knox and Stone proposed a credit assign technique to solve this problem with a probability density function to estimate the probability of the teacher's feedback delay [23]. However, the time delay might be very different for different trainers. In addition, the trainers need to learn how to operate the hardware before they start training the robots and most studies have a practice session to allow trainers getting familiar with giving feedback. Moreover, these interfaces are quite tedious and impractical for non-expert trainers in home-like environments. Therefore, it is desirable to develop more natural communication interfaces between the trainer and robot, e.g., using speech, emotion or gestures like caregiver teaching infants.

2) NATURAL INTERACTION

Instead of learning from explicit feedback provided by human trainers intentionally, an interactive RL agent can also learn from implicit feedback provided via natural interactions. Especially for long-term behavior learning with interactive RL, in order to avoid the fatigue caused by the cognitive burden of providing explicit feedback, training agents with natural feedback will be very useful and important. For example, facial expression can be extracted as evaluative feedback for personalizing the interaction process for users with different abilities. Human feedback given without the intention to teach or otherwise affect behavior—possibly derived from smiles, attention, tone of voice, or other social cues are more abundantly broadcast and can be observed without adding any cognitive load to the human [23]. Ideally, human trainers can convey their feedback via emotions, natural languages, gestures, etc., just as naturally as human-human interaction in the real life.

a: FACIAL FEEDBACK

Gadanho proposed an emotion-based architecture (EB architecture) by combining the traditional reinforcement learning with an emotion system. The emotion system is used to calculate a well-being value that was used as social reinforcement. The EB architecture can learn to decide when to switch and reinforce behavior with Q-learning [46]. Broekens examined the relationship between Emotion, Adaptation and Reinforcement Learning by proposing the EARL framework [47]. In EARL, human's real emotional expressions were analyzed in real-time as additional social reinforcement signals to train a ‘‘social robot''. Their results show that affective facial expressions facilitate robot learning significantly faster compared to a robot trained without social reinforcement. Veeriah *et al.* proposed to allow an agent to learn a value function that maps facial features extracted from a camera image to expected future reward [48]. Their preliminary results suggest that an agent can quickly adapt to a user's changing preferences and reduce the amount of explicit feedback required to complete a grip selection task. With a fully autonomous social robotic learning companion for affective child-robot tutoring, Gordon *et al.* used the measured children's valence and engagement via an automatic facial expression analysis system as reward signal for the robot's affective reinforcement learning [49]. They evaluate their system with 34 children in preschool classrooms for a duration of two months. Their results show the robot can personalize its motivational strategies to each student using verbal and non-verbal actions. Arakawa *et al.* also trained a DQN-TAMER agent with facial expressions obtained via a camera as implicit human reward [41].

In the above work, facial expressions were predefined as positive and negative feedback to train the agent, e.g., "happy" as positive feedback (+1), "angry" as negative feedback (−1). However, the positiveness and negativeness of emotions can be dynamic in the training process. Li *et al.* trained a prediction model mapping the facial feedback to explicit keypress feedback with collected data. Their simulated experiment showed that with enough recognition accuracy, agents can learn a comparative performance from solely facial feedback compared to learning from explicit keypress feedback [50].

b: NATURAL LANGUAGE-BASED FEEDBACK

When autonomous agents learn from human users, giving instruction or advice via natural languages is an intuitive and promising way for teaching agents to perform a task, especially for non-technical users. Reference [16] first proposed the RATLE (Reinforcement and Advice-Taking Learning Environment) system to incorporate programming language-based advice provided by external observer into a Q value function. Reference [17] translated natural language-based advice in English into formal language and use them to influence agent's learning policy. Reference [18] proposed the LEARN (languageE-Action Reward Network) framework, which maps free-form natural language instructions to intermediate shaping rewards based on actions taken by the agent. In addition, Tenorio *et al.* used predefined natural language-based verbal commands to communicate human evaluative feedback to train a real autonomous mobile robot learning to perform navigation tasks in a simulated environment [26]. Their experimental results show that even though human rewards delivered by verbal commands are noisy, faster convergence was achieved compared to traditional reinforcement learning from only environmental rewards.

Instead of using natural language to communicate feedback for aiding RL agent learning from environmental reward, an agent can also learn a policy directly from natural language-based instructions. Reference [51] mapped language to a reward function in an object-oriented MDP framework. Reference [52] used raw visual observations and natural language-based text input to learn a policy for instruction execution in contextual bandit setting. In [19], an adversarial learning framework is proposed to improve policy learning by generating a reward function via distinguishing a fixed set of instruction pairs from instruction pairs generated by the current policy.

c: GESTURAL FEEDBACK

Human often use gestures such as hand and body movement as communication cues in human-human interaction, especially when speech is not allowed or cannot be understood. Therefore, human gestures have the potential to be used for extracting feedback to train agents. Kuno *et al.* used gestures to control the direction of an intelligent wheelchair and proposed to recognize unknown gestures by interaction with the human user [53]. To facilitate a robot to learn from task experts rather than programming experts, Voyles and Khosla proposed to use gesture-based programming methods for providing demonstrations to train robots [54]. In addition, gestures can also be used to provide advice feedback or command feedback to aid an RL agent learning [55], [56].

B. MULTIMODAL SENSORY FEEDBACK

The systems and techniques discussed above focus on the recognition of one single input mode in order to determine human affect. The use of multimodal inputs over a single input provides two main advantages: when one modality is not available due to disturbances such as occlusion or noise, a multimodal recognition system can estimate using the remaining modalities, and when multiple modalities are available, the complementarity and diversity of information can provide feedback with increased robustness and performance.

To understand the interplay between gesture and speech and the way in which they support communication, Quek *et al.* proposed a multimodal interaction framework for discourse segmentation in free-form gesticulation accompanying speech in natural conversation [57]. Cruz *et al.* integrate dynamic multimodal audiovisual interaction with interactive reinforcement learning [56]. They allow human trainers to provide predefined advice to agents in either speech, gesture, or a combination of the two. Their results show multimodal integration facilitates the robot with interactive reinforcement learning to obtain a better performance in a smaller number of training episodes compared to unimodal scenarios. References [58], [59] used the audience's vocal laughs and visual smiles to calculate the reward as implicit social evaluative feedback to shape the humor of a robot. To endow a chess companion robot for children with empathic capabilities, Leite *et al.* use a multimodal framework to model the user's affective states and allow the robot to adapt its empathic responses to the particular preferences of the child who is interacting with it [60]. They combine visual and task-related features to measure the user's valence of feeling. The change of valence before and after the robot taking the empathic strategy is calculated as rewards for a multi-armed bandit reinforcement learning algorithm. Their preliminary study with 40 children show that robot's empathic behavior has a positive effect on users.

Almost all above methods only combine two modal inputs as feedback for agent training and the combined inputs are limited to speech, gestures, vocal laugh and visual emotions and task-related features. Actually, there are more other social cues from the human trainer that can be used as feedback, e.g., attention, speech prosody, gaze direction etc. And even more than two modal inputs can be used to deliver the trainer's feedback. In addition, these natural interactive feedback can even be combined with hardware delivered feedback to train agents. For example, Li *et al.* mapped the facial expressions to explicit keystroke feedback and proposed to allow an agent to learn from both the predicted facial feedback and keystroke feedback [50].

V. CONCLUSION AND FUTURE DIRECTIONS

This paper aims at reviewing the progress in leveraging different types of human social feedback to solve reinforcement learning tasks. In this section, we briefly discuss several promising future research directions.

A. LEARNING FROM NATURAL IMPLICIT FEEDBACK

A general problem for interactive RL is that the interface between the trainer and the robot for providing feedback has not been developed in a natural manner for domestic scenarios. Most of them used keyboard buttons or mouse clicks to provide feedback which are quite tedious and impractical for non-expert trainers in home-like environments. Although some researchers studied using facial expressions, speech or gestures to provide evaluative feedback and advice to train agents, these feedback are usually predefined and intentionally provided by trainers. Human social feedback derived from smiles, speech, attention, prosody, or other social cues are more abundantly broadcast and can be taken as implicit feedback for agent learning without adding any cognitive load to the human [23]. An open problem is that of finding ways to allow robots to learn from these free-form communicated implicit feedback. For example, affect and emotion detected in speech prosody [61] and conversations [62], [63].

B. INTERACTION DESIGN

From the perspective of robots, an understanding of how to design the interaction between the robot and the trainer allows for the design of the algorithms that support how people can teach effectively and be actively engaged in the training process at the same time. This is useful for personalizing interaction with a socially assistive robotics. In the transparent learning mechanism [64]–[66], facial expressions and body languages are used to express the robot's learning state and solicit feedback from the human teacher. What information and behavior should be communicated or expressed by the robot to elicit training of higher quality or longer duration is still a problem remaining to be investigated.

C. LEARNING FROM MULTIPLE INSTRUCTIVE MODALITIES

The literatures reviewed in this paper are mostly focused on learning from evaluative feedback or learning from advice/instruction. However, to obtain a fully autonomous interactive RL agent, algorithms for learning from human demonstration, evaluative feedback and advice/instruction should be integrated, even with standard RL learning paradigms. Much work has been done in terms of combining learning from demonstration [67], [68], evaluative feedback [37] and advice [18] with standard RL respectively. However, agents also need to learn from multiple instructive modalities, including primarily demonstration, verbal advice/instruction, evaluative feedback, attentional cues, or gestures which human teachers rely on. While there is some previous work allowing a robot learning from demonstrations and natural feedback cues provided by the teacher through speech [69], and learning from both human demonstration and evaluative feedback [70], there is still much work to be done in this respect.

ACKNOWLEDGMENT

(Jinying Lin and Zhen Ma contributed equally to this work.)

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] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] W. B. Knox and P. Stone, "Interactively shaping agents via human reinforcement: The TAMER framework," in *Proc. 5th Int. Conf. Knowl. Capture*, 2009, pp. 9–16.

[10] 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.

[11] 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.

[12] 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.

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

[14] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proc. 21st Int. Conf. Mach. Learn. (ICML)*, 2004, p. 1.

[15] J. Luketina, N. Nardelli, G. Farquhar, J. Foerster, J. Andreas, E. Grefenstette, S. Whiteson, and T. Rocktäschel, "A survey of reinforcement learning informed by natural language," 2019, *arXiv:1906.03926*. [Online]. Available: http://arxiv.org/abs/1906.03926

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

[17] G. Kuhlmann, P. Stone, R. Mooney, and J. Shavlik, "Guiding a reinforcement 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.

[18] 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

[19] 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

[20] F. Benavent and B. Zanuttini, "An experimental study of advice in sequential decision-making under uncertainty," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–9.
[21] G. Li, R. Gomez, K. Nakamura, and B. He, "Human-centered reinforcement learning: A survey," *IEEE Trans. Human-Machine Syst.*, vol. 49, no. 4, pp. 337–349, Aug. 2019.
[22] 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
[23] W. B. Knox, "Learning from human-generated reward," Ph.D. dissertation, Dept. Comput. Sci., Univ. Texas at Austin, Austin, TX, USA, 2012.
[24] G. Li, "Socially intelligent autonomous agents that learn from human reward," Ph.D. dissertation, Inform. Inst., Univ. Amsterdam, Amsterdam, The Netherlands, 2016.
[25] 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.
[26] 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.
[27] W. B. Knox, P. Stone, and C. Breazeal, "Training a robot via human feedback: A case study," in *Proc. Int. Conf. Social Robot.* Cham, Switzerland: Springer, 2013, pp. 460–470.
[28] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Proc. ICML*, vol. 99, 1999, pp. 278–287.
[29] W. B. Knox and P. Stone, "Framing reinforcement learning from human reward: Reward positivity, temporal discounting, episodicity, and performance," *Artif. Intell.*, vol. 225, pp. 24–50, Aug. 2015.
[30] N. Anh Vien and W. Ertel, "Reinforcement learning combined with human feedback in continuous state and action spaces," in *Proc. IEEE Int. Conf. Develop. Learn. Epigenetic Robot. (ICDL)*, Nov. 2012, pp. 1–6.
[31] B. Blumberg, M. Downie, Y. Ivanov, M. Berlin, M. P. Johnson, and B. Tomlinson, "Integrated learning for interactive synthetic characters," in *Proc. 29th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 2002, pp. 417–426.
[32] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.
[33] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," *Artif. Intell.*, vol. 172, nos. 6–7, pp. 716–737, Apr. 2008.
[34] H. B. Suay and S. Chernova, "Effect of human guidance and state space size on interactive reinforcement learning," in *Proc. RO-MAN*, Jul. 2011, pp. 1–6.
[35] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Trans. Syst., Man, Cybern., C (Appl. Rev.)*, vol. 42, no. 6, pp. 1291–1307, Nov. 2012.
[36] D. Arumugam, J. Ki Lee, S. Saskin, and M. L. Littman, "Deep reinforcement learning from policy-dependent human feedback," 2019, *arXiv:1902.04257*. [Online]. Available: http://arxiv.org/abs/1902.04257
[37] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, and A. L. Thomaz, "Policy shaping: Integrating human feedback with reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2625–2633.
[38] T. Cederborg, I. Grover, C. L. Isbell, and A. L. Thomaz, "Policy shaping with human teachers," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 3366–3372.
[39] W. B. Knox and P. Stone, "Combining manual feedback with subsequent MDP reward signals for reinforcement learning," in *Proc. 9th Int. Conf. Auton. Agents Multiagent Syst.*, vol. 1, 2010, pp. 5–12.
[40] W. B. Knox and P. Stone, "Reinforcement learning from simultaneous human and MDP reward," in *Proc. 11th Int. Conf. Auton. Agents Multiagent Syst.*, vol. 1, 2012, pp. 475–482.
[41] R. Arakawa, S. Kobayashi, Y. Unno, Y. Tsuboi, and S.-I. Maeda, "DQN-TAMER: Human-in-the-Loop reinforcement learning with intractable feedback," 2018, *arXiv:1810.11748*. [Online]. Available: http://arxiv.org/abs/1810.11748
[42] D. McColl, A. Hong, N. Hatakeyama, G. Nejat, and B. Benhabib, "A survey of autonomous human affect detection methods for social robots engaged in natural HRI," *J. Intell. Robotic Syst.*, vol. 82, no. 1, pp. 101–133, Apr. 2016.
[43] A. L. Thomaz, G. Hoffman, and C. Breazeal, "Real-time interactive reinforcement learning for robots," in *Proc. AAAI Workshop Hum. Comprehensible Mach. Learn.*, 2005, pp. 1–5.
[44] A. L. Thomaz, G. Hoffman, and C. Breazeal, "Reinforcement learning with human teachers: Understanding how people want to teach robots," in *Proc. 15th IEEE Int. Symp. Robot Hum. Interact. Commun. (ROMAN)*, Sep. 2006, pp. 352–357.
[45] G. Li, S. Whiteson, W. B. Knox, and H. Hung, "Using informative behavior to increase engagement while learning from human reward," *Auto. Agents Multi-Agent Syst.*, vol. 30, no. 5, pp. 826–848, Sep. 2016.
[46] S. C. Gadanho, "Learning behavior-selection by emotions and cognition in a multi-goal robot task," *J. Mach. Learn. Res.*, vol. 4, no. Jul, pp. 385–412, 2003.
[47] J. Broekens, "Emotion and reinforcement: Affective facial expressions facilitate robot learning," in *Artifical Intelligence for Human Computing*. Berlin, Germany: Springer, 2007, pp. 113–132.
[48] V. Veeriah, P. M. Pilarski, and R. S. Sutton, "Face valuing: Training user interfaces with facial expressions and reinforcement learning," 2016, *arXiv:1606.02807*. [Online]. Available: http://arxiv.org/abs/1606.02807
[49] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, "Affective personalization of a social robot tutor for children's second language skills," in *Proc. AAAI*, 2016, pp. 3951–3957.
[50] G. Li, H. Dibeklio lu, S. Whiteson, and H. Hung, "Facial feedback for reinforcement learning: A case study and offline analysis using the TAMER framework," *Auto. Agents Multi-Agent Syst.*, vol. 34, no. 1, pp. 1–29, Apr. 2020.
[51] E. C. Williams, N. Gopalan, M. Rhee, and S. Tellex, "Learning to parse natural language to grounded reward functions with weak supervision," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1–7.
[52] D. Misra, J. Langford, and Y. Artzi, "Mapping instructions and visual observations to actions with reinforcement learning," 2017, *arXiv:1704.08795*. [Online]. Available: http://arxiv.org/abs/1704.08795
[53] Y. Kuno, T. Murashima, N. Shimada, and Y. Shirai, "Interactive gesture interface for intelligent wheelchairs," in *Proc. IEEE Int. Conf. Multimedia Expo. ICME. Latest Adv. Fast Changing World Multimedia*, Jul./Aug. 2000, pp. 789–792.
[54] R. Voyles and P. Khosla, "A multi-agent system for programming robotic agents by human demonstration," in *Proc. AI Manuf. Res. Planning Workshop*, 1998, pp. 184–190.
[55] P. M. Yanik, J. Manganelli, J. Merino, A. L. Threatt, J. O. Brooks, K. E. Green, and I. D. Walker, "A gesture learning interface for simulated robot path shaping with a human teacher," *IEEE Trans. Human-Machine Syst.*, vol. 44, no. 1, pp. 41–54, Feb. 2014.
[56] F. Cruz, G. I. Parisi, J. Twiefel, and S. Wermter, "Multi-modal integration of dynamic audiovisual patterns for an interactive reinforcement learning scenario," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 759–766.
[57] F. Quek, D. McNeill, R. Bryll, S. Duncan, X.-F. Ma, C. Kirbas, K. E. McCullough, and R. Ansari, "Multimodal human discourse: Gesture and speech," *ACM Trans. Comput.-Hum. Interact.*, vol. 9, no. 3, pp. 171–193, 2002.
[58] K. Weber, H. Ritschel, F. Lingenfelser, and E. André, "Real-time adaptation of a robotic joke teller based on human social signals," in *Proc. 17th Int. Conf. Auton. Agents MultiAgent Syst.*, 2018, pp. 2259–2261.
[59] K. Weber, H. Ritschel, I. Aslan, F. Lingenfelser, and E. Andre, "How to shape the humor of a robot-social behavior adaptation based on reinforcement learning," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, 2018, pp. 154–162.
[60] I. Leite, A. Pereira, G. Castellano, S. Mascarenhas, C. Martinho, and A. Paiva, "Modelling empathy in social robotic companions," in *Proc. Int. Conf. Modeling, Adaptation, Personalization*. Berlin, Germany: Springer, 2007, pp. 135–147.
[61] E. S. Kim and B. Scassellati, "Learning to refine behavior using prosodic feedback," in *Proc. IEEE 6th Int. Conf. Develop. Learn.*, Jul. 2007, pp. 205–210.
[62] D. Hazarika, S. Poria, R. Zimmermann, and R. Mihalcea, "Conversational transfer learning for emotion recognition," 2019, *arXiv:1910.04980*. [Online]. Available: http://arxiv.org/abs/1910.04980
[63] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," 2019, *arXiv:1908.11540*. [Online]. Available: http://arxiv.org/abs/1908.11540
[64] A. L. Thomaz and C. Breazeal, "Transparency and socially guided machine learning," in *Proc. 5th Int. Conf. Develop. Learn. (ICDL)*, 2006, pp. 1–6.

[65] C. Breazeal and A. L. Thomaz, "Learning from human teachers with socially guided exploration," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2008, pp. 3539–3544.

[66] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2005, pp. 708–713.

[67] M. E. Taylor, H. B. Suay, and S. Chernova, "Integrating reinforcement learning with human demonstrations of varying ability," in *Proc. 10th Int. Conf. Auton. Agents Multiagent Syst.*, vol. 2, 2011, pp. 617–624.

[68] T. Brys, A. Harutyunyan, H. B. Suay, S. Chernova, M. E. Taylor, and A. Nowé, "Reinforcement learning from demonstration through shaping," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2015, pp. 3352–3358.

[69] M. N. Nicolescu and M. J. Mataric, "Natural methods for robot task learning: Instructive demonstrations, generalization and practice," in *Proc. 2nd Int. Joint Conf. Auton. Agents Multiagent Syst.*, 2003, pp. 241–248.

[70] G. Li, B. He, R. Gomez, and K. Nakamura, "Interactive reinforcement learning from demonstration and human evaluative feedback," in *Proc. 27th IEEE Int. Symp. Robot Human Interact. Commun. (RO-MAN)*, Aug. 2018, pp. 1156–1162.

JINYING LIN received the bachelor's degree in communication engineering from the School of Communication and Electronic Engineering, Qingdao University of Technology, Qingdao, China, in 2018. She is currently pursuing the master's degree with the School of Information Science and Engineering, Ocean University of China, Qingdao. Her current research interests include reinforcement learning, human agent/robot interaction, and robotics.

ZHEN MA received the bachelor's degree in electronic information engineering from the College of Aeronautical Engineering, Binzhou University, Binzhou, China, in 2019. She is currently pursuing the master's degree with the School of Information Science and Engineering, Ocean University of China, Qingdao, China. Her current research interests include reinforcement learning, human agent/robot interaction, and robotics.

RANDY GOMEZ (Member, IEEE) received the M.Eng.Sci. degree in electrical engineering from the University of New South Wales, Sydney, NSW, Australia, in 2002, and the Ph.D. degree in information science from the Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan, in 2006.

He was a Researcher with Kyoto University, Kyoto, Japan, in 2012, under the auspices of the Japan Society for the Promotion of Science Research Fellowship. He is currently a Senior Scientist with Honda Research Institute Japan Company Ltd., Wako, Japan. His research interests include robust speech recognition, acoustic modeling and adaptation, multimodal interaction, and robotics.

KEISUKE NAKAMURA (Member, IEEE) received the B.E. degree in control and system engineering from the Department of Control and System Engineering, Tokyo, Japan, in 2007, the M.E. degree in mechanical and control engineering from the Department of Mechanical and Control Engineering, Tokyo Institute of Technology, Tokyo, in 2010, and the Ph.D. degree in informatics from Kyoto University, Kyoto, Japan, in 2013.

From 2007 to 2008, he was with the Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, U.K. He is currently a Senior Scientist with Honda Research Institute Japan Company Ltd., Wako, Japan. His research interests include robotics, control systems, and signal processing.

BO HE (Member, IEEE) received the Ph.D. degree in control theory and control engineering from the Harbin Institute of Technology, Harbin, China, in 1999.

From 2000 to 2002, he was a Researcher with Nanyang Technological University, Singapore. He is currently a Full Professor with the Ocean University of China, Qingdao, China. His research interests include SLAM, machine learning, and robotics.

GUANGLIANG LI (Member, IEEE) received the bachelor's degree in automation and the M.Sc. degree in control theory and control engineering from the School of Control Science and Engineering, Shandong University, Jinan, China, in 2008 and 2011, respectively, and the Ph.D. degree in computer science from the University of Amsterdam, Amsterdam, The Netherlands, in 2016.

He was a Visiting Researcher with the Delft University of Technology, Delft, The Netherlands. He was a Research Intern with Honda Research Institute Japan Company Ltd., Wako, Japan. He is currently a Lecturer with the Ocean University of China, Qingdao, China. His research interests include reinforcement learning, human agent/robot interaction, and robotics.