Leveraging Human Guidance for Deep Reinforcement Learning Tasks

Ruohan Zhang[1*] , **Faraz Torabi**[1] , **Lin Guan**[1] , **Dana H. Ballard**[1] , **Peter Stone**[1]

[1]Department of Computer Science, University of Texas at Austin

{zharu, guanlin, danab}@utexas.edu; {faraztrb, pstone}@cs.utexas.edu

Abstract

Reinforcement learning agents can learn to solve sequential decision tasks by interacting with the environment. Human knowledge of how to solve these tasks can be incorporated using imitation learning, where the agent learns to imitate human demonstrated decisions. However, human guidance is not limited to the demonstrations. Other types of guidance could be more suitable for certain tasks and require less human effort. This survey provides a high-level overview of five recent learning frameworks that primarily rely on human guidance other than conventional, step-by-step action demonstrations. We review the motivation, assumption, and implementation of each framework. We then discuss possible future research directions.

1 Introduction

In end-to-end learning of sequential decision tasks, algorithms such as imitation learning (IL) [Osa *et al.*, 2018], reinforcement learning (RL) [Sutton and Barto, 2018], or a combination of both [Silver *et al.*, 2016] have achieved remarkable successes in many practical problems. These algorithms enable learning agents to learn an optimal *policy* through imitating a human demonstrator's behaviors (IL), or through trial and error by interacting with the environment (RL). Recent advancement in deep learning has enabled these learning algorithms to solve more challenging tasks [Mnih *et al.*, 2015; Silver *et al.*, 2016]. For these tasks, one major issue with deep learning algorithms is their sample efficiency. For instance, an RL agent may require millions of training samples to learn a good policy to play a video game [Mnih *et al.*, 2015]. In practice different types of human guidance are often introduced as sources of domain knowledge to speed up learning.

The most common form of human guidance is the human *policy* itself. A human trainer communicates the policy by performing the task in person and demonstrating the correct actions to the learning agent. Most imitation learning or learning from demonstration algorithms assume this type of guidance [Schaal, 1999; Argall *et al.*, 2009;

Osa *et al.*, 2018]. Nevertheless, in many cases it is impractical to use human policy as guidance because some of these tasks are too challenging for even humans to perform well. Additionally, IL algorithms may require a large amount of high-quality demonstration data, whereas collecting human behavioral data could be expensive and subject to error.

One possible solution is to leverage other types of human guidance. The types of human guidance we discuss here can be seen as feedback. The intuition is that even for tasks that are too challenging for humans, they still could provide feedback regarding the performance and guide the agent in that respect. These types of guidance could be less expensive than policy demonstration, or can be collected in parallel with policy demonstration to provide additional information.

This survey aims at providing a high-level overview of recent research efforts that primarily rely on these types of human guidance other than conventional, step-by-step action demonstrations to solve complex deep imitation and reinforcement learning tasks. The types of guidance we review include human evaluative feedback, human preference, high-level goals, human attention, and state sequences without actions. The corresponding approaches we discuss vary with regards to the trade-off between the amount of information provided to the agent and the amount of human effort required. All of these approaches have shown promising results in one or more challenging deep reinforcement learning tasks, such as Atari video games [Bellemare *et al.*, 2012] and simulated robotic control [Todorov *et al.*, 2012].

2 Background

A standard reinforcement learning problem is formalized as a Markov decision process (MDP), defined as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ [Sutton and Barto, 2018]. $\mathcal{S}$ is a set of environment states which encode relevant information for an agent's decision. $\mathcal{A}$ is a set of agent actions. $\mathcal{P}$ is the state transition function which describes $p(s'|s, a)$, i.e., the probability of entering state s' when an agent takes action a in state s. $\mathcal{R}$ is a reward function. $r(s, a, s')$ denotes the scalar reward agent received on transition from s to s' under action a. $\gamma \in [0, 1]$ is a discount factor that indicates how much the agent value an immediate reward compared to a future reward.

$\pi : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ is a policy which specifies the probability distribution of selecting actions in a given state. The goal
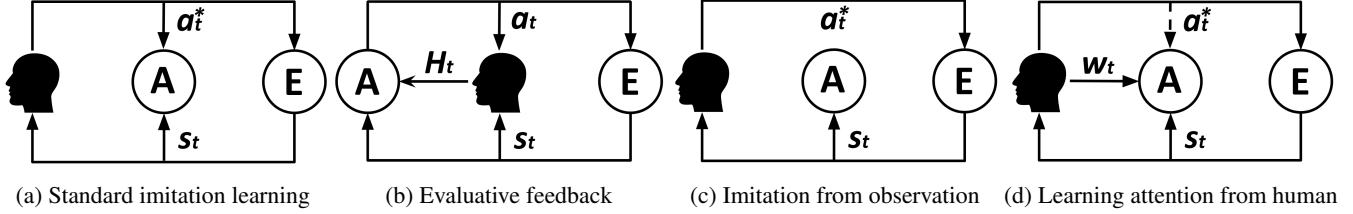
---

*Contact Author

Figure 1: Human-agent-environment interaction diagrams of different approaches discussed in this paper. These diagrams illustrate how different types of human guidance data are collected, including information required by the human trainer and the guidance provided to the agent. Note that the learning process of the agent is not included in these diagrams. **A**: learning agent; **E**: environment; Arrow: information flow direction; Dashed arrow: optional information flow. In (a) standard imitation learning, the human trainer observes state information s_t and demonstrates action a_t^* to the agent; the agent stores this data to be used in learning later. In (b) learning from evaluative feedback, the human trainer watches the agent performing the task, and provides instant feedback H_t on agent decision a_t in state s_t. (c) Imitation from observation is similar to standard imitation learning except that the agent does not have access to human demonstrated action. (d) Learning attention from human requires the trainer to provide attention map w_t to the learning agent.
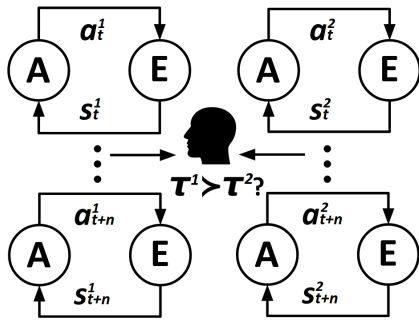

Figure 2: Learning from human preference. The human trainer watches two behaviors generated by the learning agent simultaneously, an decides which behavior is more preferable. $\tau^1 \succ \tau^2$ denotes that the trainer prefers behavior trajectory τ^1 over τ^2.
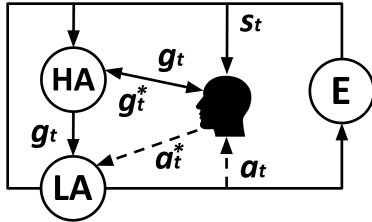

Figure 3: Hierarchical imitation. **HA**: high-level agent; **LA**: low-level agent. The high-level agent chooses a high-level goal g_t for state s_t. The low-level agent then chooses an action a_t based on g_t and s_t. The primary guidance that the trainer provides in this framework is the correct high-level goal g_t^*.

for a learning agent is to find an optimal policy π^* that maximizes the expected cumulative reward. One could optimize π directly, while alternatively many of the algorithms are based on value function estimation, i.e., estimating the state value function $V^\pi(s)$ or the action-value function $Q^\pi(s, a)$. Nowadays, deep neural networks are often used as function approximators to estimate and optimize π, V, and Q.

The standard imitation learning setting (Fig. 1a) can be formulated as MDP\R, i.e. there is no reward function $\mathcal{R}$ avail-

able. Instead, a learning agent records human demonstration in the format of state-action pairs $\{(s_t, a_t^*), (s_{t+1}, a_{t+1}^*) \ldots\}$ at each timestep. Then the agent can learn to directly mimic the human demonstrated policy using supervised learning, an approach known as behavioral cloning [Bain and Sommut, 1999]. Or it can estimate the human reward function from demonstrated policy as an intermediate step followed by standard RL to learn a policy using this reward function which is known as inverse reinforcement learning (IRL) [Abbeel and Ng, 2004]. These two approaches are the major learning frameworks used in imitation learning and comprehensive reviews of these approaches can be found in [Hussein *et al.*, 2017; Osa *et al.*, 2018]. Both approaches assume that (s_t, a_t) pairs are the only learning signal to the agent, and both s_t and a_t are available. In contrast, the learning frameworks we are about to discuss utilize different learning signals.

3 Learning from Evaluative Feedback

The first type of human guidance we consider is real-time evaluative feedback, which is given by a human trainer while watching the agent performing the task (Fig. 1b). The agent adjusts its policy online based on the feedback received. The simplest form of evaluative feedback is a scalar value indicating how desirable an observed action is. Compared to policy demonstration, this approach greatly reduces the needed human effort. It does not necessarily require the human trainer to be an expert at *performing* the task–it only requires the trainer to accurately *judge* agent behaviors.

One of the main challenges in this approach is to interpret human feedback correctly since such interpretation determines how the feedback is used to improve the policy in the MDP framework. Here we compare various methods that assume different interpretations of human feedback.

Policy shaping A policy shaping approach, e.g., the Advise framework, interprets human feedback as direct policy labels [Griffith *et al.*, 2013; Cederborg *et al.*, 2015]. The feedback could simply be "right" or "wrong", indicating whether the observed action is optimal or not. If this feedback is consistent, the agent could simply change its policy to always

choose the optimal action confirmed by the human, and eliminate actions that are wrong. [Griffith *et al.*, 2013] assumes such feedback could be inconsistent. Assuming human trainers provide the correct feedback with probability $\mathcal{C}$. The decision policy thus can be represented using a binomial distribution $\pi(s, a) = \frac{\mathcal{C}^{\Delta s, a}}{\mathcal{C}^{\Delta s, a} + (1-\mathcal{C})^{\Delta s, a}}$ where $\Delta s, a$ is the difference between the number of "right" and "wrong" labels for action a in state s. Explicitly modeling this inconsistency in human feedback allowed Advise to outperform other methods using synthetic feedback [Griffith *et al.*, 2013] and was validated on human trainers [Cederborg *et al.*, 2015].

Reward shaping In contrast to policy shaping, reward shaping interprets human feedback as the value function or a human-specified reward function [Isbell *et al.*, 2001; Tenorio-Gonzalez *et al.*, 2010; Pilarski *et al.*, 2011]. Note that this reward function may completely replace the reward function provided by the environment. The TAMER (Training an Agent Manually via Evaluative Reinforcement) framework adopts the reward function interpretation [Knox and Stone, 2009] and has been recently extended to deep TAMER [Warnell *et al.*, 2018].

TAMER assumes that the human trainer has an internal reward function $H : \mathcal{S} \times \mathcal{A} \mapsto \{-1, 0, +1\}$ that maps an observed agent action in a state to a negative, neutral, or positive feedback. Deep TAMER uses a deep network to represent and learn an estimate of H, $\hat{H}$, via supervised learning. The agent chooses the action $\pi(s) = \max_a \hat{H}(s, a)$. One notable result from [Warnell *et al.*, 2018] is that they asked human trainers to perform the task, and compared a deep TAMER agent's final performance with human trainers'. They found that the agent was able to outperform humans. This result confirms that the feedback-based approach does not require human trainers to be experts at performing the task, and the learning agent's performance is not capped by the trainers' expertise.

A natural extension is to combine human reward signal with environment reward signal, an approach named TAMER+RL [Knox and Stone, 2010; 2012]. Recently deep TAMER has been combined with a Deep Q-Network [Mnih *et al.*, 2015] in the DQN-TAMER framework [Arakawa *et al.*, 2018]. A DQN agent and a TAMER agent are trained in parallel, and the final decision policy is a weighted average of the policies from both agents. Effective methods to combine two sources of reward are yet to be explored although several possibilities have been proposed [Knox and Stone, 2010].

Intervention A slightly different but related approach is Human Intervention RL [Saunders *et al.*, 2018], in which a human supervises the training process of an RL agent and blocks catastrophic actions. The blocked state-action pairs can be used to train a binary classifier to detect and block unsafe actions [Saunders *et al.*, 2018]. Since blocking can be viewed as strong negative feedback, the intervention method can be seen as TAMER+RL without positive feedback.

Policy-dependent feedback TAMER interprets human feedback as a reward function that is independent of the agent's current policy. However, COACH (Convergent Actor-Critic by Humans) [MacGlashan *et al.*, 2017] assumes that such feedback is policy-dependent, and should be better interpreted as the advantage function that specifies how much better or worse when deviating from the agent's current policy. Note that the advantage function gradually decreases to zero or negative as π improves. This interpretation better captures several phenomena observed in human feedback such as diminishing returns. The human feedback can then be used to replace the advantage function in calculating the policy gradient using the actor-critic algorithm [MacGlashan *et al.*, 2017].

The algorithms above interpret human feedback differently, and result in different reinforcement learning objectives and update rules. Using synthetic feedback, MacGlashan *et al.* showed that the convergence of these algorithms depends critically on whether the actual feedback matches the assumed one [2017]. The nature of the feedback could potentially vary across tasks and trainers [Loftin *et al.*, 2016], and can be altered by instruction given to the trainers [Cederborg *et al.*, 2015]. Therefore these factors need to be carefully controlled in practice. A potential further direction is to design a method that is either robust to all types of feedback, or can infer the human feedback type using a probabilistic model and adapt to the trainer's strategy online [Loftin *et al.*, 2016].

4 Learning from Human Preference

The second type of guidance we consider is human preference. Many RL tasks are difficult for humans to provide demonstrations or evaluative feedback. These could be control tasks with many degrees of freedom with non-human morphology, e.g., MuJoCo [Todorov *et al.*, 2012], it is therefore harder for humans to control or to tell whether a particular state-action pair is good or not. It is more natural for the agent to query human trainers for their *preferences* or *rankings* over a set of behaviors. This feedback can be provided over a set of state or action sequences; however, it is much less demanding if it is over trajectories, as the trainer can directly evaluate the outcome of the trajectories. As a result, as shown in Fig. 2, we consider preferences over trajectory segments, or sequences of state-action pairs: $\tau = ((s_0, a_0), (s_1, a_1), \dots)$.

Previous works have used preferences to directly learn policies [Wilson *et al.*, 2012; Busa-Fekete *et al.*, 2013], learn a preference model [Fürnkranz *et al.*, 2012], or learn a reward function [Wirth *et al.*, 2016; Akrour *et al.*, 2014]. A survey on these topics is provided by Wirth *et al.* [2017].

More recent works have extended previous preference-based RL methods to be compatible with deep RL [Christiano *et al.*, 2017; Ibarz *et al.*, 2018]. A pair of agent trajectories of 1-2 seconds long are simultaneously presented to human trainers to query for their preference. The goal is to learn the latent human reward function $r(s, a)$, as in IRL, but from human preference instead of a human policy. Intuitively, the probability of a human preferring a segment could depend on the total reward summed over the trajectory [Christiano *et al.*, 2017]. A model can be trained to find the reward function that minimizes the cross-entropy loss between model's prediction and human preference. Since the targets to be evaluated are trajectories instead of state-action pairs, the feedback can be very sparse. This approach drastically reduces human effort.

The amount of human feedback required can be as little as 1% of the total number of agent actions [Christiano *et al.*, 2017]. Ibarz *et al.* further extended this approach by including human demonstration to pre-train the agent, and including demonstration trajectories in the preference learning, assuming human trajectories are always more preferable than agent trajectories [2018].

An important component in this framework is query selection, in which the agent needs to decide which trajectories to query for preference. Christiano *et al.* simply select trajectories such that an ensemble of learning models have the largest variance in predicting their human preferences [2017]. Ideally, the query should maximize the expected information gain from an active learning perspective, an important research challenge that is closely related to preference elicitation [Zintgraf *et al.*, 2018].

5   Hierarchical Imitation

Many reinforcement learning tasks are hierarchically structured, meaning that they can be decomposed and solved using a divide-and-conquer approach. It is possible to ask human trainers to only provide high-level feedback on these tasks. Similar to preference, this type of feedback also targets trajectory segments, but is provided as option choice [Sutton *et al.*, 1999], i.e., choice of high-level goal in a given state. Due to the task's hierarchical structure, the behavior trajectory can be naturally segmented into these options, instead of arbitrarily segmented in the preference framework.

Le *et al.* [2018] has proposed the hierarchical guidance framework that assumes a two-level hierarchy, in which a high-level agent learns to choose a goal g given a state, while low-level agents learn to execute a sub-policy (option) to accomplish the chosen goal (Fig. 3). Note that an action to terminate the current option needs to be added to the low-level agent's action space, and this termination action can be demonstrated by a human and learned by the agent. Human trainers are asked to provide three types of feedback: 1) a positive signal if the high-level goal g_t and low-level sub-policy a_t are both correct. 2) the correct high-level goal g_t^* for state s_t if the chosen g_t is incorrect. 3) the correct low-level action a_t^* if the high-level goal was chosen correctly but the low-level sub-policy is incorrect. At each level, the learning task becomes a typical imitation learning problem, therefore conventional IL algorithms such as behavioral cloning and DAgger [Ross *et al.*, 2011] can be applied. Perhaps the most exciting result comes from a hybrid of hierarchical imitation learning and RL, in which a DAgger agent learns to choose a high-level goal from human guidance, and a Q-learning agent learns low-level sub-policies. This approach was shown to be substantially more sample efficient than conventional imitation learning on a very difficult Atari game called Montezuma's revenge [Le *et al.*, 2018] .

Andreas *et al.* take a similar hierarchical approach but only require humans to provide policy sketches that are high-level symbolic subtask labels [2017]. The policy of each subtask is learned by the RL agent on its own and no longer requires human demonstration. As mentioned earlier, it is natural to extend hierarchical imitation to incorporate human preferences over the outcome of options, instead of asking humans to provide the correct option labels, as done in Pinsler *et al.* [2018].

6   Imitation from Observation

Imitation from observation (IfO) [Torabi *et al.*, 2019c] is the problem of learning directly by observing a trainer performing the task. The learning agent only has access to visual data of the trainer; therefore, no action data is available (Fig. 1c). The ultimate goal in this framework is to enable agents to utilize existing, rich amount of demonstration data that do not have action labels, such as online videos of humans performing various tasks. There are two general challenges in addressing this problem: (1) perception, and (2) control.

Existing resources are not necessarily generated in ideal conditions; therefore, the agent should be able to deal with the challenges in perception. For instance, the demonstrations could have different viewpoints. One approach developed to address this issue learns a context translation model to translate an observation to predict the observation in the target context [Liu *et al.*, 2018]. Another approach uses a classifier to distinguish between the data that comes from different viewpoints and attempts to maximize the domain confusion in an adversarial setting during the training. Consequently, the extracted features would be invariant with respect to the viewpoint [Stadie *et al.*, 2017]. One other challenge in perception is embodiment mismatch where the imitator and the trainer do not share the same embodiment. One method developed to approach this problem learns a correspondence between the embodiments using autoencoders in a supervised fashion [Gupta *et al.*, 2018]. Another method learns the correspondence in an unsupervised fashion with a small number of human supervision [Sermanet *et al.*, 2018].

The second challenge in IfO is control. This component concerns learning the task from proprioceptive states of the trainer $(s_0, s_1, \dots)$ instead of visual data. The existing IfO control algorithms can be categorized in two broad classes: model-based and model-free.

Model-based algorithms learn a dynamics model during the imitation process. One approach is to learn an inverse dynamics model which estimates the taken action a, given a state transitions (s, s'). Therefore, this model can be used to infer the missing action labels of the expert. Then, the inferred actions can be executed to reproduce the trainers states [Nair *et al.*, 2017]. As an alternative, after inferring the actions, a mapping from states to the actions can be learned and used to improve the learned model and consequently the policy [Torabi *et al.*, 2018a]. If a sparse reward function is available, this algorithm can be combined with an RL algorithm to improve the performance [Guo *et al.*, 2019]. Another type of dynamics model is a forward dynamics model which predicts the next state s', given the current state and action (s, a). A combination of policy with this type of dynamics model can be used to predict the next state from a given current state and then train the imitation policy to make the prediction more accurate [Edwards *et al.*, 2018].

On the other hand, model-free algorithms attempt to imitate the task without explicitly learning any dynamics model. A class of this type of algorithms imitate by designing a re-

ward function as the Euclidean distance of the states of the imitator from that of the trainer at each timestep, and then solving an RL problem [Sermanet *et al.*, 2018; Gupta *et al.*, 2018; Dwibedi *et al.*, 2018]. Another class of approaches, inspired by generative adversarial imitation learning [Ho and Ermon, 2016], imitate the tasks in an adversarial fashion. A method of this type uses GANs [Goodfellow *et al.*, 2014] to bring the state distribution of the imitator closer to that of the trainer [Merel *et al.*, 2017]. However, since equivalence in *single state* distribution does not mean equivalence in policy, getting the *state-transition* distributions closer together may result in better performance [Torabi *et al.*, 2018b; 2019a; 2019b].

7   Learning Attention from Humans

During the human demonstration or evaluation process, there are other helpful signals for a learning agent from a social learning perspective. For example, human voice [Tenorio-Gonzalez *et al.*, 2010] and facial expression [Arakawa *et al.*, 2018] can be used as additional sources of evaluative feedback. For visual learning tasks, a particularly useful information source from humans is the gaze position, which can be collected in parallel with demonstration. Researchers have collected human gaze and policy data for daily cooking [Li *et al.*, 2018], Atari game playing [Zhang *et al.*, 2019], and outdoor driving [Palazzi *et al.*, 2018; Kim *et al.*, 2018]. Since human eyes have limited resolution except for the center fovea, humans learn to move their eyes to the correct place at the right time to process urgent state information. The agent could learn to extract useful state features by learning to attend from human. Recent works have trained convolutional neural networks to predict the probability distribution of human gaze given a raw image and were able to generate a gaze heatmap in the pixel space [Li *et al.*, 2018; Palazzi *et al.*, 2018; Kim *et al.*, 2018; Zhang *et al.*, 2019]. This heatmap serve as a strong feature extractor that is particularly helpful for tasks with a high-dimensional state space.

Existing works have proposed to use a learned gaze prediction model to guide the process of learning a human policy [Li *et al.*, 2018; Zhang *et al.*, 2018; Kim *et al.*, 2018]. Intuitively, knowing where humans would look provides useful information on what action they will take. Therefore the predicted gaze heatmap helps select important features in the given state. Doing so results in higher accuracy in imitating human actions [Li *et al.*, 2018; Zhang *et al.*, 2018; Kim *et al.*, 2018] and incorporating learned gaze model into behavioral cloning leads to an average performance increase of 198% in Atari games [Zhang *et al.*, 2018].

For future work, attention learning is closely related to hierarchical imitation, since gaze is a good indicator of the current high-level behavioral goal which might help an imitator to infer the correct low-level action. Furthermore, gaze data can be collected in parallel with evaluation. This data might reveal more information to the learning agent to explain why the human gives a particular evaluative feedback.

8   Conclusion and Future Directions

This survey aims at providing a literature review for progress in leveraging different types of human guidance to solve deep reinforcement learning tasks. Here we briefly discuss several promising future research directions.

Shared datasets and reproducibility In general, researchers collect their own human guidance data. However, this type of data is often expensive to collect. An effort that could greatly facilitate research in this field is to create publicly available benchmark datasets. Collecting dataset would be difficult for interactive learning methods; but for other approaches, data can be collected in advance and shared. Another concern is reproducibility in RL [Henderson *et al.*, 2018]. When collecting human guidance data, factors such as individual expertise, experimental setup, data collection tools, dataset size, and experimenter bias could introduce large variances in final performance. Therefore, evaluating algorithms using a standard dataset could save the effort and assure a fair comparison between algorithms. But it would not allow for feedback that is dependent on the changing policy as it is being learned.

Understanding human trainers Leveraging human guidance to train an agent naturally follows a teacher-student paradigm. Much effort has been spent on making the student more intelligent. However, understanding the behavior of human teachers is equally important. Thomaz and Breazeal pioneered the effort in understanding human behavior in teaching learning agents [2008]. As RL agents become more powerful and attempt to solve more complex tasks, the human teachers' guiding behaviors could become more complicated and require further study.

Studying the behaviors, especially the limitations of human teachers, allows one to design a teaching environment that is more effective and produces more useful guidance data. Amir *et al.* [2016] studied human attention limits while monitoring the learning process of an agent and proposed an algorithm for the human and the agent to jointly identify states where feedback is most needed to reduce human monitoring cost. Ho *et al.* showed the differences in behavior when a human trainer is intentionally teaching (showing) versus merely doing the task. They found that humans modify their policies to reveal the goal to the agent when in the showing mode but not in doing mode. They further showed that imitation learning algorithm can benefit substantially more from the data collected in the showing mode [Ho *et al.*, 2016].

Understanding the nature of human guidance allows algorithms to learn more effectively. We have already seen the debate on how to interpret human evaluative feedback in complex tasks. A helpful way to resolve this debate is to conduct human studies with diverse subject pools to investigate whether real-life human feedback satisfies their algorithmic assumptions and what factors affect the human feedback strategy [Cederborg *et al.*, 2015; Loftin *et al.*, 2016; MacGlashan *et al.*, 2017].

A unified lifelong learning paradigm The learning frameworks discussed in this paper are often inspired by real-life biological learning scenarios that correspond to differ-

ent learning stages and strategies in lifelong learning. Imitation and reinforcement learning correspond to learning completely by imitating others and learning completely through self-generated experience, where the former may be used more often in early stages of learning and the latter could be more useful in late stages. The other learning strategies discussed fill in the gap between those two extremes. In reality, an animal is likely to utilize all possible intermediate learning signals to learn to perform a challenging task to gain reward. We have compared these approaches within an imitation and reinforcement learning framework. Under this framework, it is possible to develop a unified learning paradigm that accepts multiple types of human guidance, as explored by Abel *et al.* [2017].

Acknowledgement

References

[Abbeel and Ng, 2004] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1. ACM, 2004.

[Abel *et al.*, 2017] David Abel, John Salvatier, Andreas Stuhlmüller, and Owain Evans. Agent-agnostic human-in-the-loop reinforcement learning. *NeurIPS Workshop on the Future of Interactive Learning Machines*, 2017.

[Akrour *et al.*, 2014] Riad Akrour, Marc Schoenauer, Michèle Sebag, and Jean-Christophe Souplet. Programming by feedback. In *International Conference on Machine Learning*, volume 32, pages 1503–1511. JMLR. org, 2014.

[Amir *et al.*, 2016] Ofra Amir, Ece Kamar, Andrey Kolobov, and Barbara J Grosz. Interactive teaching strategies for agent training. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 804–811. AAAI Press, 2016.

[Andreas *et al.*, 2017] Jacob Andreas, Dan Klein, and Sergey Levine. Modular multitask reinforcement learning with policy sketches. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 166–175. JMLR. org, 2017.

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

[Argall *et al.*, 2009] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[Bain and Sommut, 1999] Michael Bain and Claude Sommut. A framework for behavioural claning. *Machine intelligence*, 15(15):103, 1999.

[Bellemare *et al.*, 2012] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 2012.

[Busa-Fekete *et al.*, 2013] Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, Weiwei Cheng, and Eyke Hüllermeier. Preference-based evolutionary direct policy search. In *ICRA Workshop on Autonomous Learning*, 2013.

[Cederborg *et al.*, 2015] Thomas Cederborg, Ishaan Grover, Charles L Isbell, and Andrea L Thomaz. Policy shaping with human teachers. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[Christiano *et al.*, 2017] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307, 2017.

[Dwibedi *et al.*, 2018] Debidatta Dwibedi, Jonathan Tompson, Corey Lynch, and Pierre Sermanet. Learning actionable representations from visual observations. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1577–1584. IEEE, 2018.

[Edwards *et al.*, 2018] Ashley D Edwards, Himanshu Sahni, Yannick Schroeker, and Charles L Isbell. Imitating latent policies from observation. *arXiv preprint arXiv:1805.07914*, 2018.

[Fürnkranz *et al.*, 2012] Johannes Fürnkranz, Eyke Hüllermeier, Weiwei Cheng, and Sang-Hyeun Park. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine learning*, 89(1-2):123–156, 2012.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[Griffith *et al.*, 2013] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in neural information processing systems*, pages 2625–2633, 2013.

[Guo *et al.*, 2019] Xiaoxiao Guo, Shiyu Chang, Mo Yu, Gerald Tesauro, and Murray Campbell. Hybrid reinforcement learning with expert state sequences. *Association for the Advancement of Artificial Intelligence*, 2019.

[Gupta *et al.*, 2018] Abhishek Gupta, Coline Devin, YuXuan Liu, Pieter Abbeel, and Sergey Levine. Learning invariant feature spaces to transfer skills with reinforcement learning. In *International Conference on Learning Representations*, 2018.

[Henderson *et al.*, 2018] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Ho and Ermon, 2016] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pages 4565–4573, 2016.

[Ho *et al.*, 2016] Mark K Ho, Michael Littman, James MacGlashan, Fiery Cushman, and Joseph L Austerweil. Showing versus doing: Teaching by demonstration. In *Advances in neural information processing systems*, pages 3027–3035, 2016.

[Hussein *et al.*, 2017] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):21, 2017.

[Ibarz *et al.*, 2018] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari. In *Advances in Neural Information Processing Systems*, pages 8022–8034, 2018.

[Isbell *et al.*, 2001] Charles Isbell, Christian R Shelton, Michael Kearns, Satinder Singh, and Peter Stone. A social reinforcement learning agent. In *Proceedings of the fifth international conference on Autonomous agents*, pages 377–384. ACM, 2001.

[Kim *et al.*, 2018] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–578, 2018.

[Knox and Stone, 2009] W Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, pages 9–16. ACM, 2009.

[Knox and Stone, 2010] W Bradley Knox and Peter Stone. Combining manual feedback with subsequent mdp reward signals for reinforcement learning. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 5–12. International Foundation for Autonomous Agents and Multiagent Systems, 2010.

[Knox and Stone, 2012] W Bradley Knox and Peter Stone. Reinforcement learning from simultaneous human and mdp reward. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 475–482. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[Le *et al.*, 2018] Hoang Le, Nan Jiang, Alekh Agarwal, Miroslav Dudik, Yisong Yue, and Hal Daumé. Hierarchical imitation and reinforcement learning. In *International Conference on Machine Learning*, pages 2923–2932, 2018.

[Li *et al.*, 2018] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 619–635, 2018.

[Liu *et al.*, 2018] YuXuan Liu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1118–1125. IEEE, 2018.

[Loftin *et al.*, 2016] Robert Loftin, Bei Peng, James MacGlashan, Michael L Littman, Matthew E Taylor, Jeff Huang, and David L Roberts. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Autonomous agents and multi-agent systems*, 30(1):30–59, 2016.

[MacGlashan *et al.*, 2017] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L Roberts, Matthew E Taylor, and Michael L Littman. Interactive learning from policy-dependent human feedback. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2285–2294. JMLR. org, 2017.

[Merel *et al.*, 2017] Josh Merel, Yuval Tassa, Sriram Srinivasan, Jay Lemmon, Ziyu Wang, Greg Wayne, and Nicolas Heess. Learning human behaviors from motion capture by adversarial imitation. *arXiv preprint arXiv:1707.02201*, 2017.

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

[Nair *et al.*, 2017] Ashvin Nair, Dian Chen, Pulkit Agrawal, Phillip Isola, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Combining self-supervised learning and imitation for vision-based rope manipulation. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2146–2153. IEEE, 2017.

[Osa *et al.*, 2018] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2):1–179, 2018.

[Palazzi *et al.*, 2018] Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara. Predicting the driver's focus of attention: the dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[Pilarski *et al.*, 2011] Patrick M Pilarski, Michael R Dawson, Thomas Degris, Farbod Fahimi, Jason P Carey, and Richard S Sutton. Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning. In *2011 IEEE International Conference on Rehabilitation Robotics*, pages 1–7. IEEE, 2011.

[Pinsler *et al.*, 2018] Robert Pinsler, Riad Akrour, Takayuki Osa, Jan Peters, and Gerhard Neumann. Sample and feedback efficient hierarchical reinforcement learning from human preferences. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 596–601. IEEE, 2018.

[Ross *et al.*, 2011] Stéphane Ross, Geoffrey J Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, pages 627–635, 2011.

[Saunders *et al.*, 2018] William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. Trial without error: Towards safe reinforcement learning via human intervention. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2067–2069. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

[Schaal, 1999] Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3(6):233–242, 1999.

[Sermanet *et al.*, 2018] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1134–1141. IEEE, 2018.

[Silver *et al.*, 2016] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[Stadie *et al.*, 2017] Bradly C Stadie, Pieter Abbeel, and Ilya Sutskever. Third-person imitation learning. *International Conference on Learning Representations*, 2017.

[Sutton and Barto, 2018] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[Sutton *et al.*, 1999] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

[Tenorio-Gonzalez *et al.*, 2010] Ana C Tenorio-Gonzalez, Eduardo F Morales, and Luis Villaseñor-Pineda. Dynamic reward shaping: training a robot by voice. In *Ibero-American conference on artificial intelligence*, pages 483–492. Springer, 2010.

[Thomaz and Breazeal, 2008] Andrea L Thomaz and Cynthia Breazeal. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6-7):716–737, 2008.

[Todorov *et al.*, 2012] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[Torabi *et al.*, 2018a] Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4950–4957. AAAI Press, 2018.

[Torabi *et al.*, 2018b] Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*, 2018.

[Torabi *et al.*, 2019a] Faraz Torabi, Garrett Warnell, and Peter Stone. Adversarial imitation learning from state-only demonstrations. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2229–2231. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[Torabi *et al.*, 2019b] Faraz Torabi, Garrett Warnell, and Peter Stone. Imitation learning from video by leveraging proprioception. In *Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

[Torabi *et al.*, 2019c] Faraz Torabi, Garrett Warnell, and Peter Stone. Recent advances in imitation learning from observation. In *Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

[Warnell *et al.*, 2018] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Wilson *et al.*, 2012] Aaron Wilson, Alan Fern, and Prasad Tadepalli. A bayesian approach for policy learning from trajectory preference queries. In *Advances in neural information processing systems*, pages 1133–1141, 2012.

[Wirth *et al.*, 2016] Christian Wirth, Johannes Fürnkranz, and Gerhard Neumann. Model-free preference-based reinforcement learning. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[Wirth *et al.*, 2017] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *The Journal of Machine Learning Research*, 18(1):4945–4990, 2017.

[Zhang *et al.*, 2018] Ruohan Zhang, Zhuode Liu, Luxin Zhang, Jake A Whritner, Karl S Muller, Mary M Hayhoe, and Dana H Ballard. Agil: Learning attention from human for visuomotor tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 663–679, 2018.

[Zhang *et al.*, 2019] Ruohan Zhang, Zhuode Liu, Lin Guan, Luxin Zhang, Mary M Hayhoe, and Dana H Ballard. Atari-head: Atari human eye-tracking and demonstration dataset. *arXiv preprint arXiv:1903.06754*, 2019.

[Zintgraf *et al.*, 2018] Luisa M Zintgraf, Diederik M Roijers, Sjoerd Linders, Catholijn M Jonker, and Ann Nowé. Ordered preference elicitation strategies for supporting multi-objective decision making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1477–1485. International Foundation for Autonomous Agents and Multiagent Systems, 2018.