Leveraging Human Guidance for Sequential Decision-Making Tasks
In *Autonomous Agents and Multi-Agent Systems* June, 2021

Noname manuscript No.
(will be inserted by the editor)

Recent Advances in Leveraging Human Guidance for Sequential Decision-Making Tasks

Ruohan Zhang · Faraz Torabi · Garrett Warnell · Peter Stone

Received: date / Accepted: date

Abstract A longstanding goal of artificial intelligence is to create artificial agents capable of learning to perform tasks that require sequential decision making. Importantly, while it is the artificial agent that learns and acts, it is still up to humans to specify the particular task to be performed. Classical task-specification approaches typically involve humans providing stationary reward functions or explicit demonstrations of the desired tasks. However, there has recently been a great deal of research energy invested in exploring alternative ways in which humans may guide learning agents that may, e.g., be more suitable for certain tasks or require less human effort. This survey provides a high-level overview of five recent machine learning frameworks that primarily rely on human guidance apart from pre-specified reward functions or conventional, step-by-step action demonstrations. We review the motivation, assumptions, and implementation of each framework, and we discuss possible future research directions.

Keywords Learning from demonstration · Imitation learning · Reinforcement learning · Human feedback · Hierarchical learning · Imitation from observation · Attention

Ruohan Zhang[1*]
E-mail: zharu@utexas.edu
Faraz Torabi[1*]
E-mail: faraztrb@utexas.edu
Garrett Warnell[2]
E-mail: warnellg@cs.utexas.edu
Peter Stone[1,3]
E-mail: pstone@cs.utexas.edu
[*] Contributed equally to this work
[1] Department of Computer Science, The University of Texas at Austin
[2] U.S. Army Research Laboratory
[3] Sony AI

1 Introduction

Artificial agents require humans to specify the tasks they should perform. With respect to artificial learning agents in particular, humans must provide some specification of what the agent should learn to perform. One method by which humans typically provide this specification is by designing a stationary reward function. This function provides a reward to the agent when it correctly performs the desired task and, perhaps, punishment when the agent does not. Artificial learning agents may then approach the task-learning process using *reinforcement learning* (RL) techniques (Sutton and Barto, 2018) that seek to find a *policy* (i.e., an explicit function that the agent uses to make decisions) that allows the agent to gather as much reward as possible. Another popular way in which humans specify tasks for artificial agents to learn is by demonstrating the task themselves. Typically, this is accomplished by having the human perform the task while the learning agent observes the actions that the human takes (e.g., the human physically moving a robot arm). In these cases, artificial agents may use approaches from *imitation learning* (IL) (Schaal, 1999; Argall et al., 2009; Osa et al., 2018) in order to find policies that allow them to perform the demonstrated task. Both paradigms described above (i.e., RL and IL) have been used with remarkable success (Mnih et al., 2015; Silver et al., 2016; Levine et al., 2016; Silver et al., 2017, 2018; Jaderberg et al., 2019; Vinyals et al., 2019), especially when combined with deep learning (LeCun et al., 2015) to solve challenging sequential decision-making tasks.

While reward functions and explicit action demonstrations currently represent the most common ways in which humans specify tasks for learning agents, recent years have seen a great deal of research energy devoted to studying alternative ways in which humans might perform task specifications. In general, these alternatives are focused on more diverse and creative ways of providing input than the two methods described above, and so we explicitly refer to the resulting types of input as *human guidance*. Because human guidance is less direct compared to specified reward functions or explicit action demonstrations, attempts to leverage it have led to several new research challenges in the machine learning community.

There are many reasons for the recent interest in utilizing human guidance. One reason is the relative ease with which several forms of human guidance can be collected. For some tasks, it may be exceedingly difficult for a human trainer to specify a reward function or provide an action demonstration since both require some level of training and skill that the human may not possess. However, it may still be possible for the human to guide the learning agent. As an analogy from human learning, consider the sports coach that provides guidance in the form of feedback on professional athlete performance. Even though the coach typically can not explicitly demonstrate the skill to be performed at the same skill or performance level as the athlete, their feedback is often useful to the athlete. In these cases, the availability of guidance may even help the learner achieve greater final task performance than if an action demonstration alone was provided. Another reason for the research community's interest in studying machine learning from human guidance lies in the utility of human guidance as a supplemental training signal that can increase the speed of task learning. That is, even in cases for which a reward signal or an action demonstration is available, if the learning agent can leverage available human guidance, the overall amount of time it takes to arrive at an acceptable behavior policy can be greatly reduced compared to if the guidance had not been used at all.

This survey aims at providing a high-level overview of recent research efforts that primarily rely on human guidance as opposed to conventional reward functions or step-by-step action demonstrations. We will define and discuss learning from five forms of human guidance (Zhang et al., 2019), including *(1)* evaluative feedback, *(2)* preferences, *(3)* high-level goals (hierarchical imitation), *(4)* demonstration sequences without actions (imitation from observation), and *(5)* attention. Though the approaches to be discussed vary with regards to the trade-off between the amount of information provided to the agent and the amount of human effort required, all have shown promising results in one or more challenging sequential decision-making tasks.

2 Background

In this section, we provide background relevant to the rest of the paper. More specifically, we first discuss Markov decision processes (*MDP*s), reinforcement learning, and the notation used in this paper. We then provide a short review of imitation learning.

2.1 Markov Decision Processes (*MDP*s)

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), where

- $\mathcal{S}$ is a set of environment states which encodes 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 values an immediate reward compared to a future reward.

As a concrete example, Atari Montezuma's Revenge (Bellemare et al., 2013) (Fig. 1) is one of the most challenging video games for reinforcement learning research. The game has rich visual features, complicated game dynamics, and

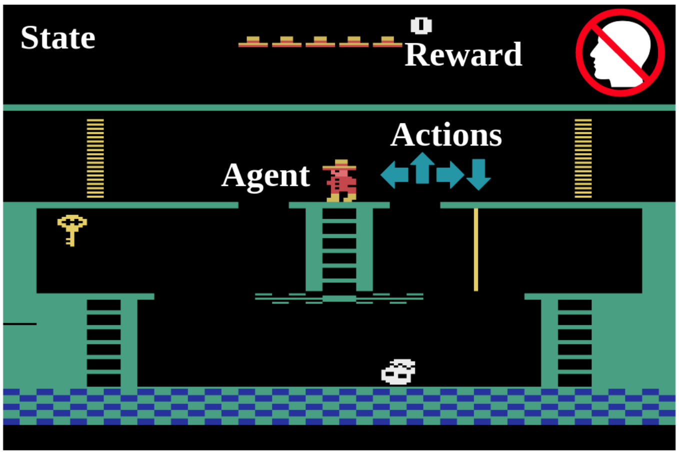


Fig. 1: Atari Montezuma's Revenge is a challenging sequential decision-making task that is widely used in reinforcement learning research. The problem is modeled as a Markov decision process. In a typical reinforcement learning setting, the agent needs to learn to play the game without any human guidance purely based on the score provided by the environment.

very sparse rewards. Modeled as an MDP, the state is the game image frame, or a stack of frames to capture temporal information. The agent controls the avatar by choosing an action from a discrete set of actions at every timestep. The agent receives the reward from the game engine in the form of game scores. We will use this game as a running example throughout the survey.

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

The state value function for a given policy π is defined as (Sutton and Barto, 2018)

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0} \gamma^t R(s_t, a_t) \mid s_0 = s \right] \tag{1}$$

A corresponding action value function, $Q^\pi(s, a)$, also exists and is given by

$$Q^\pi(s, a) = E_\pi \left[R(s_t, a_t) + V^\pi(s_{t+1}) \mid s_t = s, a_t = a \right] \tag{2}$$

and the advantage function $A^\pi(s, a)$, is defined as

$$A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s). \tag{3}$$

The state value function $V^\pi(s)$ measures the expected cumulative reward to be in a particular state s and following policy π afterward. The action-value function $Q^{\pi}(s, a)$ defines the same quantity but for taking a particular action a when in state s and following policy π afterward. The advantage function tells us the relative gain ("advantage") that could be obtained by taking a certain action compared to the average action taken at that state (Wang et al., 2016).

Several successful RL algorithms that seek to estimate these quantities directly have been developed, including Q Learning (Watkins and Dayan, 1992) and advantage actor-critic (see, e.g., Sutton and Barto (2018)). For example, Q-learning seeks to learn the state-action value function for the optimal policy, $Q^{\pi^*}(s, a)$, and the policy is then given by $\pi^*(s) = \arg\max_a Q^{\pi^*}(s, a)$. Nowadays, deep neural networks are often used as function approximators to estimate and optimize π, V, and Q.

An important challenge in RL is to balance exploration vs. exploitation when an agent selects its action. Exploration allows the agent to improve its current knowledge. Exploitation chooses the greedy action to maximize reward by exploiting the agents current knowledge. A simple strategy (ϵ-greedy) chooses a random action with probability ϵ and chooses the greedy action (the action with the highest Q value) with probability $1 - \epsilon$ (Sutton and Barto, 2018). A more sophisticated strategy uses a Boltzmann distribution for selecting actions based on the current estimate of Q function (Sutton and Barto, 2018):

$$P(a|s, Q, \tau) = \frac{e^{Q(s,a)}/\tau}{\sum_{a' \in \mathcal{A}} e^{Q(s,a')}/\tau} \tag{4}$$

where τ is a temperature constant that controls the exploration rate.

2.2 Imitation Learning

The learning frameworks surveyed in this paper are inspired by, an extension of, or combined with traditional imitation learning algorithms. The standard imitation learning setting (Fig. 2 and Fig. 4a) can be formulated as MDP$\backslash R$, i.e. there is no reward function $\mathcal{R}$ available. Instead, a learning agent (the *imitator*) records expert (the *demonstrator*, could be a human expert or an artificial agent) demonstrations in the format of state-action pairs $\{(s_t, a_t^*)\}$ at each timestep, and then attempts to learn the task using that data.

One approach is for the agent to learn to mimic the demonstrated policy using supervised learning, which is known as behavioral cloning (Bain and Sommut, 1999). A second approach to imitation learning is called inverse reinforcement learning (*IRL*) (Abbeel and Ng, 2004) which involves learning a reward function based on the demonstration data and learning the imitation policy using RL with the learned reward function. These two approaches constitute the major learning frameworks used in imitation learning. Comprehensive reviews of these two approaches can be found in Argall et al. (2009); Hussein et al. (2017); Osa et al. (2018); Arora and Doshi (2018); Fang et al. (2019). More recently, generative adversarial imitation learning (*GAIL*) (Ho and Ermon, 2016) has been proposed, which utilizes the notion of generative adversarial networks (*GAN*) (Goodfellow et al., 2014).

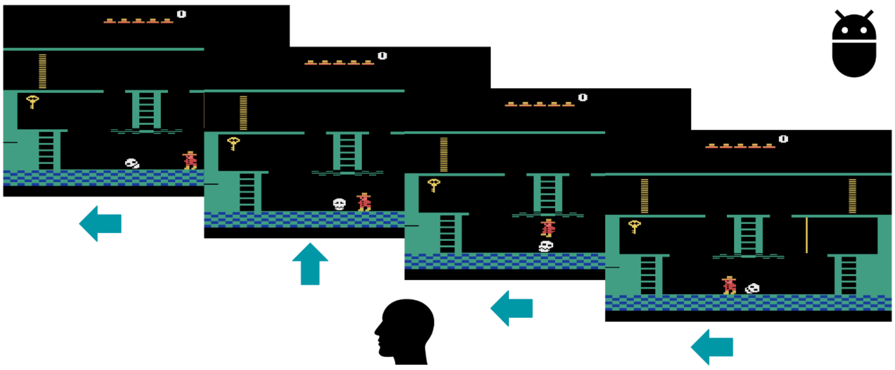


Fig. 2: In standard imitation learning, a human trainer demonstrates a sequence of actions, and the agent learns to imitate the trainer's actions using behavioral cloning, inverse reinforcement learning, or adversarial imitation.

Importantly, all of these approaches assume that (s_t, a_t^*) pairs are the only learning signal to the agent and that both s_t and a_t^* are available to the agent. Unfortunately, access to the optimal actions, a_t^*, is not always plausible, as the task might be too complex for human demonstrators to perform. Therefore, there has recently been a great deal of interest in the research community in learning frameworks that utilize learning signals other than optimal action information, and it is these techniques that we review in this survey. Before doing so, however, we briefly describe the three IL frameworks described above, i.e., *(1)* behavioral cloning (*BC*), *(2)* inverse reinforcement learning (*IRL*), and *(3)* generative adversarial imitation learning (*GAIL*).

*2.2.1 Behavioral Cloning (*BC*)*

Behavioral cloning (Pomerleau, 1989; Bain and Sommut, 1999) is one of the main methods to approach an imitation learning problem. The agent receives as training data both the encountered states and actions of the demonstrator, then uses supervised learning techniques such as classification or regression to estimate the demonstrator's policy. This method is powerful in the sense that it is capable of imitating the demonstrator without having to interact with the environment, and it has been successfully applied in many application domains. For instance, it has been used to train a quadrotor to fly down a forest trail (Giusti et al., 2016). There, the training data consists of images of the forest trail gathered by cameras mounted on a human hiker and labeled with the actions (walking directions) that the human used. The policy is modeled as a convolutional neural network classifier, and trained using supervised learning. In the end, the quadrotor managed to fly down the trail successfully. BC has also been used in autonomous driving (Bojarski et al., 2016). The training data is acquired using a human demonstrator, and a convolutional neural network is trained to map raw pixels from a single front-facing camera directly to platform steering commands. After training, the vehicle was capable of driving in traffic on local roads. BC has also been successfully used to teach robotic manipulators complex, multi-step, real-world tasks using kinesthetic demonstrations (Niekum et al., 2015).

One of BC's major drawbacks is potential performance degradation due to the well-studied compounding error caused by covariate shift (Ross and Bagnell, 2010; Ross et al., 2011), i.e., that training and testing data distribution mismatch results in deviation of the learned behavior from the demonstration (Torabi et al., 2018a). Ross et al. (2011) proposed an interactive training method to correct the shift called DAgger (Dataset Aggregation) which attempts to bring the distribution of demonstration data closer to that of the learned behavior. It does so by collecting demonstration data on the states observed by the imitator at each iteration. Retraining the policy on the aggregated dataset ultimately prevents the imitator from deviating from the demonstration behavior.

*2.2.2 Inverse Reinforcement Learning (*IRL*)*

Inverse reinforcement learning (Abbeel and Ng, 2004; Ziebart et al., 2008) is a second category of imitation learning. IRL techniques seek to learn a reward function that has the maximum value for the demonstrated actions. The learned reward function is then used in combination with RL methods to find an imitation policy. To be more specific, most IRL algorithms first initialize a random policy. Next, the agent executes that policy in the environment to collect state-action data, and then the algorithms estimate the expert's reward function based on the data generated by the policy and the demonstration data. Finally, standard RL algorithms are used to learn an optimal policy for that reward function. The process of reward learning and policy learning is repeated until the agent policy becomes sufficiently close to the demonstrator's policy. Like BC techniques, IRL methods usually assume that state-action pairs are available (Finn et al., 2016), and also that the reward is a function of both states and actions. The algorithms developed in this category have shown impressive results in a variety of tasks such as autonomous helicopter aerobatics (Abbeel et al., 2010), robot object manipulation (Finn et al., 2016), and autonomous navigation in complex unstructured terrains (Silver et al., 2010), etc.

One major drawback of most algorithms developed for IRL is that at each iteration, they have to solve a complete RL problem to find an optimal policy given the currently estimated reward function which is computationally very expensive. However, the learned policies are often more robust than the policies learned by BC algorithms as they do not suffer from the covariate shift problem. This shift does not happen in the case of IRL because the agent can interact with the environment while training and the distribution mismatch diminishes during the process.

2.2.3 Adversarial Imitation Learning

Recently an imitation learning algorithm, generative adversarial imitation learning (*GAIL*) (Ho and Ermon, 2016), has been developed that alleviates the IRL's drawback just set forth. This algorithm directly learns the policy given demonstration bypassing the optimal reward recovery. *GAIL* formulates the problem of finding an imitating policy as that of solving the following optimization problem:

$$\begin{aligned}\min_{\pi \in \prod} \max_{D \in (0,1)^{\mathcal{S} \times \mathcal{A}}} &- \lambda_H H(\pi) + \mathbb{E}_{\pi}[\log(D(s,a)] + \\ &\mathbb{E}_{\pi_E}[\log(1 - D(s,a))] \ ,\end{aligned} \tag{5}$$

where $\prod$ is the set of all stationary stochastic policies, π_E is the demonstrator's policy, λ_H is a weight factor, H is the entropy function, and the discriminator function $D : \mathcal{S} \times \mathcal{A} \to (0,1)$ can be thought of as a classifier trained to differentiate between the state-action pairs provided by the demonstrator and those experienced by the imitator. The objective in (5) is inspired by the one used in generative adversarial networks (*GAN*s) (Goodfellow et al., 2014). A *GAN* system is trained in a competitive process: the generator tries to fool the classifier while the classifier tries to distinguish the generated data from the real data. This competitive training process makes both models do better by trying to beat the other. In *GAIL* the associated algorithm can be thought of as trying to induce an imitator state-action occupancy measure that is similar to that of the demonstrator. π and D are often parameterized in practice and that *GAIL*seeks to find the saddle point of Eq. 5 by sequentially making gradient steps with respect to the parametrization of D and π. Trust Region Policy Optimization (TRPO) (Schulman et al., 2015) is often used to update the policy. Maximizing the entropy term $H(\pi)$ follows the maximum causal entropy IRL (Ziebart et al., 2008, 2010; Bloem and Bambos, 2014). The entropy serves as a policy regularizer to account for the noise and suboptimality in the demonstrated behavior (Ziebart et al., 2008). Even more recently, there has been research on methods that seek to improve on *GAIL* by, e.g., increasing sample efficiency (Kostrikov et al., 2019; Sasaki et al., 2019) and improving reward representation (Fu et al., 2018; Qureshi et al., 2019).

2.3 Common Task Domains

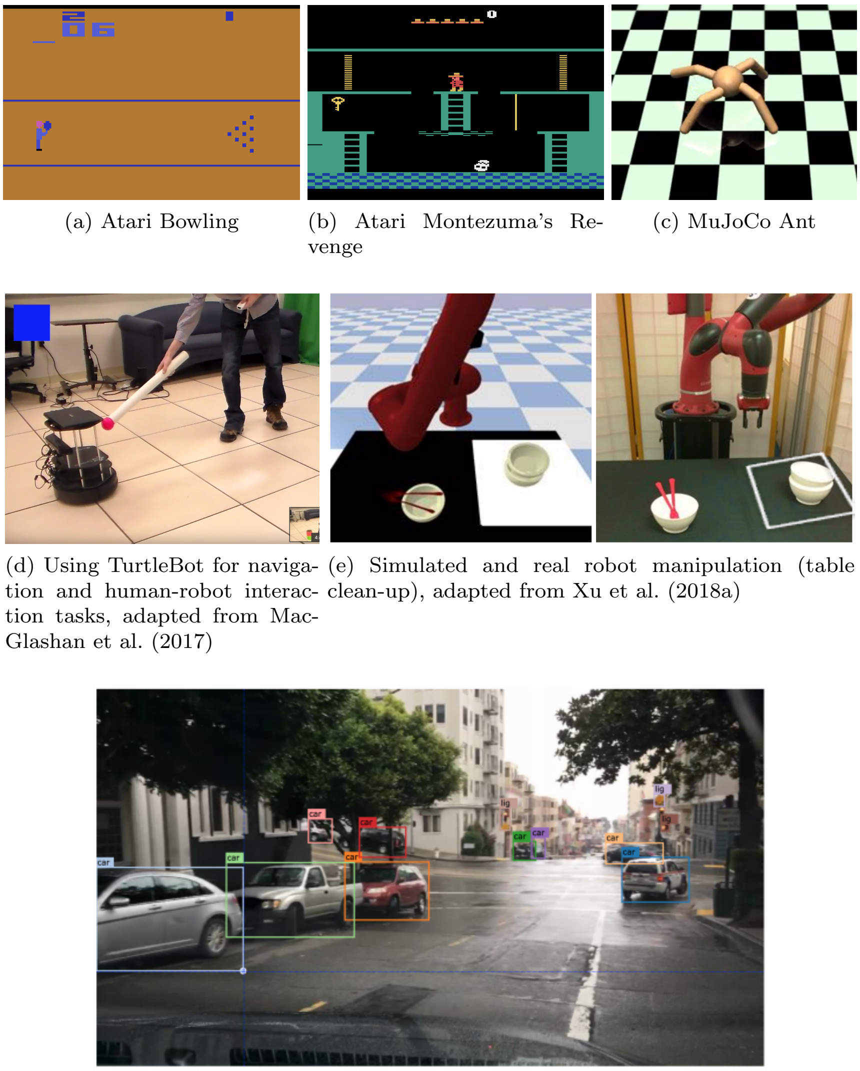
Next, we introduce several sequential decision-making tasks that are commonly used today to test the algorithms discussed above. Before deep learning, sequential decision-making models and learning algorithms were often confined to task domains with low dimensional state space such as 2D gridworld and mountain car (Sutton and Barto, 1998). The emergence of deep neural networks (LeCun et al., 2015) has enabled these models and algorithms to solve significantly more challenging tasks. These tasks include Atari 2600 video games (Bellemare et al., 2013; Machado et al., 2018) in which the state space could be high-dimensional raw game images. The platform has 60 unique video games that span a variety of dynamics, visual features, reward mechanisms, and difficulty levels for both humans and AIs. Montezuma's Revenge (Fig. 1) is one of the most difficult games due to very sparse rewards. Hence it is one of the most challenging games in terms of exploration.

Another example is robotic locomotion tasks using MuJoCo (Todorov et al., 2012), a simulator with a physics engine and multi-joint dynamics to study complex dynamical systems in contact-rich behaviors. It is the first full-featured simulator designed from the ground up for the purpose of model-based optimization, and in particular optimization through contacts (Todorov et al., 2012).

Recently, much effort has been spent on moving from simulation to real-world applications, from the navigation robots (e.g., TurtleBot (MacGlashan et al., 2017)), robotic manipulators (e.g., Sawyer robot arm (Xu et al., 2018a)), to autonomous driving vehicles (Yu et al., 2018). These are typically tasks with high-dimensional state space at which humans are particularly good. Examples of the tasks can be seen in Fig. 3. In some of the tasks such as board games, reinforcement learning agents have already surpassed human expert performance (Silver et al., 2016), and could perform even better without human knowledge (Silver et al., 2017, 2018). For example, Silver et al. (2017) have shown that a pure RL agent that learns to play Go by itself from scratch can outperform an IL/RL hybrid agent (Silver et al., 2016) which first learns to imitate expert human Go players' moves. However, RL agents and algorithms still face significant challenges in solving many of the tasks we discuss in this survey.

3 Overview

Given the models and notations defined above, we now provide formal definitions of the five learning frameworks surveyed that leverage human guidance. Diagrams that visualize the interactions between the human trainers, the learning agents, and the task environment for imitation learning together with these five learning frameworks can be found in Fig. 4. 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 does not perform the task, instead, he or she watches the agent performing the task, and provides instant feedback H_t on agent decision a_t in state s_t. In (c) learning from human preference. The human trainer watches two behaviors generated by the learning agent simultaneously and decides which behavior is more preferable. In (d) hierarchical imitation, 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^*. Imitation from observation (e) is similar to standard imitation learning except that the agent does not have access to human



(a) Atari Bowling

(b) Atari Montezuma's Revenge

(c) MuJoCo Ant

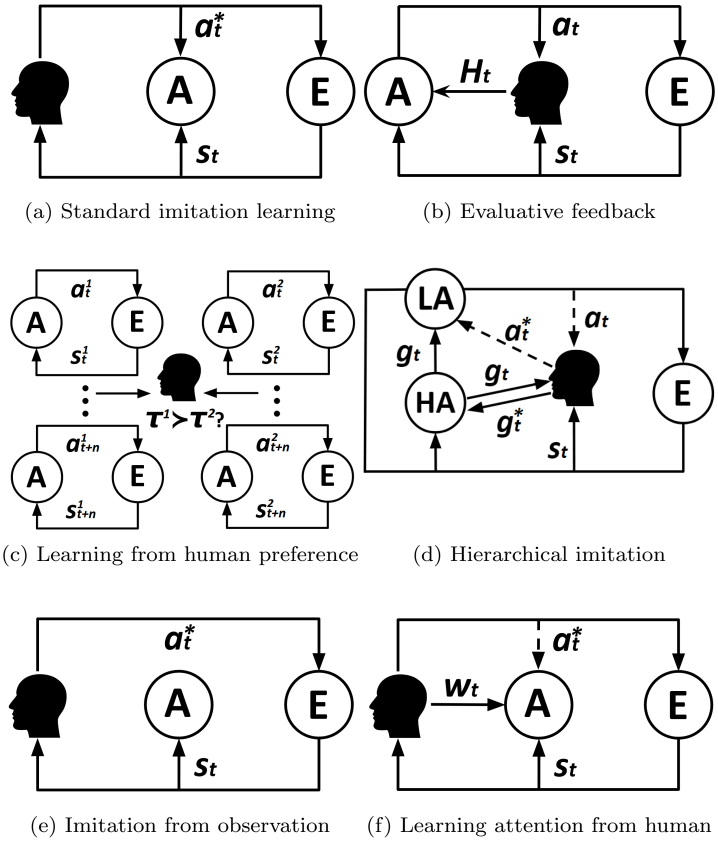
(d) Using TurtleBot for navigation and human-robot interaction tasks, adapted from MacGlashan et al. (2017)

(e) Simulated and real robot manipulation (table clean-up), adapted from Xu et al. (2018a)

(f) Autonomous driving, adapted from Yu et al. (2018)

Fig. 3: Example sequential decision tasks that are commonly used in recent research that leverages human guidance. The state spaces in these tasks are typically high-dimensional, such as raw images and robot joints with multiple degrees of freedom.

demonstrated action – it only observes the state sequence demonstrated by the human. Learning attention from humans (f) requires the trainer to provide attention information w_t that indicates important task features to the learning agent. For each learning framework, a summary and comparison of selected papers surveyed can be found in Table 1.



(a) Standard imitation learning (b) Evaluative feedback

(c) Learning from human preference (d) Hierarchical imitation

(e) Imitation from observation (f) Learning attention from human

Fig. 4: Human-agent-environment interaction diagrams of five learning frameworks surveyed. 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. Arrow: information flow direction; Dashed arrow: optional information flow. **A**: learning agent; **E**: environment; s_t: the state at time t; a_t: agent action. (a) a_t^*: human demonstrated action. (b) H_t: human evaluative feedback on agent decision a_t in state s_t. (c) $\tau^1 \succ \tau^2$: human trainer prefers agent behavior trajectory τ^1 over τ^2. (d) **HA**: a high-level agent that chooses a high-level goal g_t for state s_t; **LA**: a low-level agent that chooses an action a_t based on g_t and s_t; g_t^*: high-level goal provided by human. (e) Note that the human demonstrated action a_t^* is not available to the agent. (f) w_t: human attention information.

Paper	Human guidance	Task domain(s)	On-line?	Dataset?	Section
Cederborg et al. (2015)	evaluative feedback	Pac-Man	Yes	No	4
Warnell et al. (2018)	evaluative feedback	Atari Bowling	Yes	No	4
Arumugam et al. (2019)	evaluative feedback	Minecraft	Yes	No	4
Saunders et al. (2018)	evaluative feedback, actions	3 Atari games	Yes	No	4
Akinola et al. (2020)	evaluative feedback	simulated robot navigation	Yes	No	4
Christiano et al. (2017)	preference	8 MuJoCo tasks, 7 Atari games	Yes	No	5
Sadigh et al. (2017)	preference	simulated driving	Yes	No	5
Ibarz et al. (2018)	preference, actions	9 Atari games	Yes	No	5
Bestick et al. (2018)	preference	simulated and physical robot handover	Yes	No	5
Cui and Niekum (2018)	preference	simulated robot manipulation	Yes	No	5
Palan et al. (2019)	preference, action	simulated driving, Lunar Lander, simulated and physical robot manipulation	Yes	No	5
Le et al. (2018)	high-level and low-level actions	Atari Montezuma's Revenge, maze navigation	Yes	No	6
Andreas et al. (2017)	high-level actions	crafting, maze navigation, MuJoCo	No	No	6
Gupta et al. (2020)	low-level actions	simulated robot manipulation	No	No	6
Krishnan et al. (2017)	low-level actions	robot manipulation	No	No	6
Codevilla et al. (2018)	high-level and low-level actions	driving	No	Link	6
Xu et al. (2018a)	high-level and low-level actions	robot manipulation	No	No	6
Fox et al. (2019)	high-level and low-level actions	robot manipulation	No	Link	6
Torabi et al. (2018a)	state	MuJoCo	No	No	7
Liu et al. (2018)	state	MuJoCo, physical robot manipulation	No	No	7
Sermanet et al. (2018)	state	physical robot manipulation	No	Link	7
Torabi et al. (2018b)	state	MuJoCo	No	No	7
Yang et al. (2019)	state	MuJoCo	No	No	7
Palazzi et al. (2018)	gaze	driving	No	Link	8
Deng et al. (2019)	gaze	driving	No	Link	8
Liu et al. (2019)	gaze, action	driving	No	No	8
Xia et al. (2020)	gaze, action	driving	No	Link	8
Zuo et al. (2018)	gaze, action	non-verbal interaction	No	No	8
Li et al. (2018)	gaze, action	meal preparation	No	Link	8
Zhang et al. (2020b)	gaze, action	20 Atari games	No	Link	8

Table 1: A comparison of selected papers surveyed. This table only includes recent works that aimed to solve task domains with high-dimensional state space. The "On-line" column specifies whether the learning is done on-line or off-line, where on-line means that a human trainer must be available during the agent's learning process. "Dataset" indicates whether associated human guidance data is published. If so the link to the dataset is provided.

4 Learning from Evaluative Feedback

We begin with one of the most natural forms of human guidance that have been studied: *evaluative feedback*. Proposed paradigms for learning from evaluative feedback typically involve human trainers watching artificial agents attempt to execute tasks and those humans providing a scalar signal that communicates the desirability of the observed agent behavior, as shown in Fig. 4b and 5. Using this type of human guidance, the learning problem for the agent is that of determining how to adjust its policy such that its future behavior becomes more desirable to the human.

Evaluative feedback is an attractive form of human guidance due to the relative ease with which humans can provide it. For example, for cases in which the human trainer cannot provide a demonstration of the task (because, e.g., the task is too difficult), the human typically still knows what constitutes good behavior and can therefore provide evaluative feedback. Moreover, even

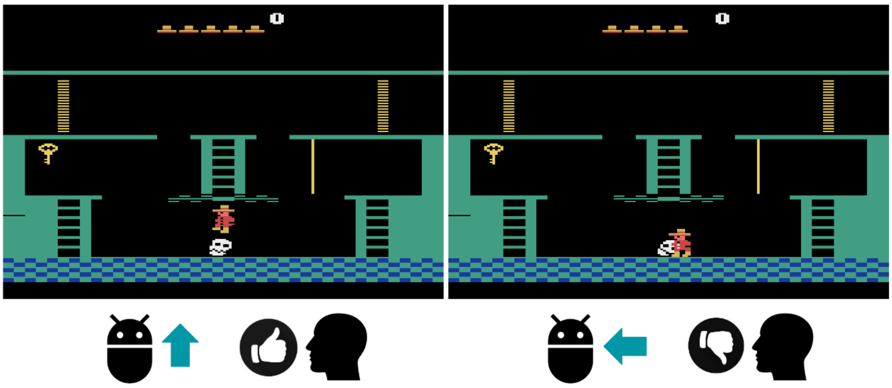


Fig. 5: In learning from evaluative feedback, a human trainer watches the agent's learning process, and provides positive feedback for a desirable action (jumping over the skull), and negative feedback for an undesirable action (running into the skull).

when the human can provide a demonstration, providing additional evaluative feedback during the learning process may allow the artificial agent to achieve a task performance that exceeds that of the human demonstrator.

One of the main challenges faced by machines that seek to learn from human-provided evaluative feedback is that of correctly interpreting the feedback signal. Indeed, several interpretations have been proposed by members of the research community, each leading to a different type of machine learning method. Typically, the particular feedback interpretation manifests as equating the feedback with a particular quantity derived from the RL setting. Here, we group the proposed methods into two categories: those that assume the feedback given communicates *reward-like* information, and those that interpret the feedback as a *value-like* quantity.

4.1 Human Feedback as Reward

In the RL setting (a fixed MDP), a stationary reward function $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is defined as a means by which to specify a fixed task. Due to this stationarity, RL algorithms are able to seek policies that exhibit a notion of optimality with respect to a statistic dependent upon this distribution. That is, RL algorithms seek $\pi^* = \arg\max_\pi J(\pi)$, where $J(\pi) = \mathbb{E}_\pi \left[\sum_t \gamma^t R(s_t, a_t)\right]$ is well-defined. The RL community has proposed several algorithms to accomplish this task, including policy gradient techniques (Sutton et al., 2000) and actor-critic techniques (Grondman et al., 2012).

Due to the success of RL, some researchers have proposed techniques for learning from human feedback that interpret the feedback as the reward function itself (Isbell et al., 2001; Tenorio-Gonzalez et al., 2010). Intuitively, this interpretation amounts to assuming that the human feedback provides an instantaneous rating of the agent's current decision. For example, Pilarski et al. (2011) propose a technique for learning from human feedback that uses the feedback as the reward function in an actor-critic algorithm. During training, the human trainer provides a positive ($r = +0.5$) or negative ($r = -0.5$) reward to the learning agent. In the absence of a human-delivered reward, they simply assume the feedback is neural ($r = 0$). For their experimental setting, they find that the proposed algorithm can learn useful policies when the human feedback is consistent, but that policy quality degrades when the feedback becomes inconsistent (i.e., becomes less stationary). Another such technique is Advise (Griffith et al., 2013; Cederborg et al., 2015), in which the human feedback is used as the reward signal in a policy-gradient-like algorithm. More specifically, the probability of an action being a good action is:

$$P_c(a) = \frac{\mathcal{C}^{\Delta s,a}}{\mathcal{C}^{\Delta s,a} + (1-\mathcal{C})^{\Delta s,a}} \tag{6}$$

where $\Delta s, a$ is the difference between the number of "right" and "wrong" labels that the human provided for action a in state s. Notably, this approach coarsely takes human error into consideration using the $\mathcal{C}$ parameter. Let $\mathcal{C}$ denote the probability that an evaluation of an action choice is correctly provided by the human teacher ($\mathcal{C} = 0.5$ is a random non-informative teacher, and $\mathcal{C} = 1$ is a flawless teacher) (Cederborg et al., 2015). Suppose $P_q(a)$ is the probability of selecting action a by an RL agent (e.g., according to Boltzmann distribution, Eq. 4), the final policy during learning is determined using both P_c and P_q:

$$\pi(s,a) = \frac{P_q(a)P_c(a)}{\sum_{a' \in \mathcal{A}} P_q(a')P_c(a')} \tag{7}$$

In this way the algorithm combines knowledge it learned from interacting with the environment and knowledge it gained from human evaluative feedback.

4.2 Human Feedback as Value

Alternatively, several methods interpret the feedback signal as a value-like quantity (Knox and Stone, 2009; MacGlashan et al., 2017). Intuitively, this interpretation amounts to assuming that the human feedback provides a rating of the agent's current decision with respect to some forecast of future behavior.

One such technique is the TAMER algorithm (training an agent manually via evaluative reinforcement) (Knox and Stone, 2009), in which it is assumed that the human has in mind a desired policy π_H, and the feedback given at a time instant t, $H(s_t, a_t)$ roughly corresponds to $Q^{\pi_H}(s_t, a_t)$ (defined in Eq. 2). TAMER agents use supervised learning with all the feedback collected up to time t to calculate the current estimate of H, $\hat{H}$, e.g., through minimizing a standard squared loss (Warnell et al., 2018):

$$\hat{H}^* = \arg\min_{\hat{H}} \sum_t \left[\hat{H}(s_t, a_t) - H(s_t, a_t)\right]^2 \tag{8}$$

Then the agent acts, in the next state, according to the policy

$$a_{t+1} = \arg\max_{a} \hat{H}^*(s_{t+1}, a) \tag{9}$$

in a fashion similar to Q Learning since we interpret $\hat{H}^*$ as an approximation for Q^{π_H}. Because the TAMER algorithm interprets the human feedback to be the value corresponding to a fixed (ideal) policy π_H, the implicit assumption made is that the feedback given is independent of the agent's current policy and depends only on the quality of an agents action selection. This type of human feedback model is called *policy-independent* models.

Alternatively, we could have *policy-dependent* models in which the feedback depends on the agents current policy. An action selection may be rewarded or punished more depending on how often the agent would typically be inclined to select it. For example, the human may greatly reward the agent for deviating from its current policy to take a slightly better action (though this action may still be sub-optimal), and stop rewarding this action as the agent consistently adopts this action (MacGlashan et al., 2017). This phenomenon is known as diminishing returns and is policy-dependent (MacGlashan et al., 2017). The COACH (convergent actor-critic by humans) framework has leveraged the idea of policy-dependent feedback and assumes instead that the human feedback corresponds to the *advantage* (Eq. 3) for the current policy (MacGlashan et al., 2017). Intuitively, the advantage function communicates how much better or worse the agent's behavior is when deviating from its current policy. Algorithmically, COACH uses the feedback to replace the advantage function in calculating the policy gradient in an advantage actor-critic algorithm. Note that the human trainers do not need to provide feedback at every timestep like other evaluative feedback approaches.

With the advent of deep learning, several researchers in the community have recently begun attempting to use these techniques in the context of more challenging, high-dimensional state spaces. For example, Warnell et al. (2018) propose a technique that enables the use of TAMER for pixel-level state spaces in Atari games. To overcome the difficulty faced by trying to learn functions over such state spaces from sparse feedback, the authors propose to use a combination of a pre-trained deep autoencoder for state representation and a feedback replay buffer to allow for off-policy updates. Arumugam et al. (2019) report that a similar approach is successful when applying COACH to pixel-level state spaces. Aside from demonstrating the utility of learning from human feedback algorithms in high-dimensional state spaces, Warnell et al. (2018) also reported that agents trained using human-provided feedback ultimately learned policies that outperformed that of the human trainers themselves. This result would seem to support the hypothesis that the performance of agents that can learn from human feedback is not capped by the trainer's expertise to perform the task. However, such performance could be affected by the trainer's expertise in providing evaluative feedback.

4.3 Extensions and Outlook

Several extensions to the above algorithms have been proposed in the literature. Notably, several have studied combining human-provided evaluative feedback with existing reward functions (Cederborg et al., 2015; Knox and Stone, 2010, 2012; Arakawa et al., 2018) with the goal of augmenting reinforcement learning. Saunders et al. (2018) look explicitly at situations in which humans block catastrophic actions, and interpret these blocking actions as evaluative feedback when learning in combination with an existing reward function. This method is particularly useful for RL tasks that require safe exploration.

Evaluative feedback is usually communicated through button presses by humans, some other works have sought to infer feedback signals from multimodal evaluative signals humans naturally emit during social interactions, including gestures (Najar et al., 2020), facial expressions (Broekens, 2007; Arakawa et al., 2018; Cui et al., 2020), electroencephalogram (EEG) based brain waves signals (Xu et al., 2020; Akinola et al., 2020), and implied feedback when humans refrain from giving explicit feedback (Loftin et al., 2014; Joachims et al., 2017). Other body-language and vocalization modalities not aimed at explicit communication, such as tone of voice, subtle head gestures, and hand gestures, could also be modeled and leveraged during training in the future (Cui et al., 2020). Other works have looked at methods by which to elicit more feedback from human trainers (Li et al., 2016b) or to explicitly account for situations in which the human trainer may not be paying attention (Kessler Faulkner et al., 2019).

Each of the algorithms presented above interprets human feedback in slightly different ways, resulting in different policy update rules. Using synthetic feedback, MacGlashan et al. (2017) showed that the convergence of these algorithms depends critically on whether the actual feedback matches the assumed one. Critically, the nature of the feedback could potentially vary across tasks and trainers. Loftin et al. (2016) has shown that instead of providing balanced feedback, human trainers could be more reward-focused (provides explicit rewards for correct actions and ignore incorrect ones), punishment-focused, or even inactive. Factors such as previous experience in training pets and feedback from the agent can affect the trainer's strategies (Loftin et al., 2016). Additionally, the nature of the feedback can be altered by the instruction given to the trainers. For example, Cederborg et al. (2015) has shown that they can manipulate the meaning of human trainer's silence by differing the instructions given: The trainers were told that their silence meant positive/negative to the agent. Not surprisingly, the agent needs to adjust their interpretations of silence accordingly to perform well (Cederborg et al., 2015). Therefore, these factors need to be carefully controlled in practice. One potential future research direction is to study methods that explicitly attempt to be robust to many types of feedback or methods that attempt to infer the human feedback type and adapt to that type in real time (Grizou et al., 2014; Loftin et al., 2016; Najar et al., 2020).

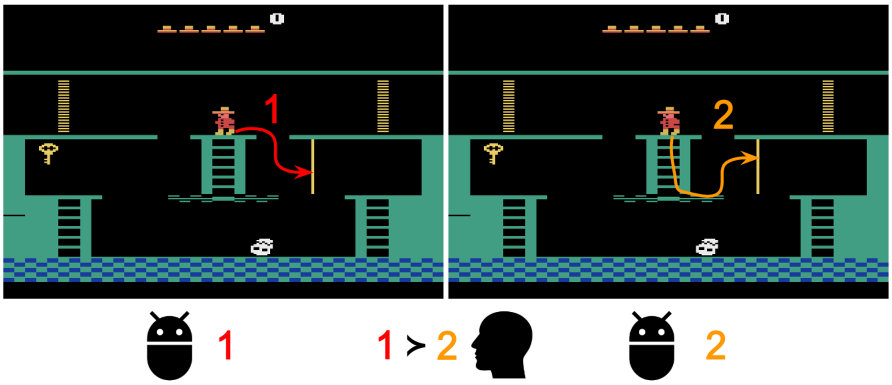


Fig. 6: In learning from human preference, the learning agent presents two learned behavior trajectories to the human trainer, and the human tells the agent which trajectory is preferable. Here the human trainer prefers trajectory 1.

5 Learning from Human Preference

The second type of human guidance we discuss is that communicated in the form of a preference. As with evaluative feedback, for many tasks that we may wish artificial agents to learn, it may be difficult or impossible for humans to provide demonstrations due to challenges such as embodiment mismatch. For example, consider control tasks with many degrees of freedom in which the artificial agent exhibits non-human morphology, as are commonly present in the MuJoCo environment (Todorov et al., 2012). Further, because of the complexity of the state space, it may also prove difficult for humans to provide fine-grained evaluative feedback on any particular portion of the behavior.

For such cases, some in the research community have posited that it is more natural for the agent to query human trainers for their *preferences*, or *rankings*, over a set of exhibited behaviors. This feedback can be provided for a set of state or action sequences; however, it is much less demanding if it is over trajectories as the trainer can directly evaluate outcomes. Here, as shown in Fig. 4c and 6, we consider preferences over trajectory segments, or sequences of state-action pairs: $\tau = ((s_0, a_0), (s_1, a_1), \dots)$. Using this type of human guidance, the learning problem is to learn a policy or an external reward function from human preference.

Preference learning has long been a topic of interest in the research community. 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 learning methods to be compatible with deep RL. The goal is to learn a hypothesized latent human reward function, $r(s,a)$ (as in IRL) from the communicated preferences (Christiano et al., 2017; Sadigh et al., 2017; Bestick et al., 2018; Cui and Niekum, 2018). In Christiano et al. (2017), a pair of agent trajectories approximately 1-2 seconds in duration is simultaneously presented to human trainers to query for their preference. Under the model developed by Christiano et al. (2017), the probability of a human preferring a segment depends exponentially on the total reward summed over the trajectory:

$$P[\tau^1 \succ \tau^2] = \frac{\exp \sum r(s_t^1, a_t^1)}{\exp \sum r(s_t^1, a_t^1) + \exp \sum r(s_t^2, a_t^2)} \tag{10}$$

This model provides a training objective that can be used to find the reward function by minimizing the cross-entropy loss between the model's prediction and the human's preferences. Since the targets to be evaluated are trajectories instead of state-action pairs, the feedback is typically very sparse compared to the amount of state-action data, resulting in a drastic reduction in 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).

Preference learning problems are generally formulated as IRL problems. Hence it is natural to integrate preference and action demonstration via a joint IRL framework (Palan et al., 2019; Bıyık et al., 2020), with a nice insight that these two sources of information are complementary under the IRL framework: "demonstrations provide a high-level initialization of the human's overall reward functions, while preferences explore specific, fine-grained aspects of it" (Bıyık et al., 2020). Therefore they use demonstrations to initialize a reward distribution, and refine the reward function with preference queries (Palan et al., 2019; Bıyık et al., 2020). Ibarz et al. (2018) takes a different approach to combine demonstration and preference information, by using human demonstrations to pre-train the agent. Further, they include demonstration trajectories when learning preferences, assuming human trajectories are always more preferable than agent trajectories.

A key aspect of learning methods designed to leverage preferences is that of query selection, i.e., the decision the agent makes regarding which trajectories to query for the human's preference. Christiano et al. (2017) select trajectories such that an ensemble of their learning models have the largest variance, i.e., uncertainty, in predicting the human's preference. Ideally, however, the query should maximize the expected information gain from an active learning perspective (Cui and Niekum, 2018), an important research challenge that is closely related to preference elicitation (Zintgraf et al., 2018). Sadigh et al. (2017) have shown that query selection can be done by actively synthesizing preference queries. The reward learning procedure can be facilitated if selected queries can remove the maximal amount of hypotheses in the space of possible reward functions (Sadigh et al., 2017). Follow-up works have extended this approach to batch-active methods (Biyik and Sadigh, 2018), using rankings instead of pairwise comparisons (Bıyık et al., 2019), and modeling the reward using more expressive models such as Gaussian processes (Biyik et al., 2020).

Query selection in preference learning falls into the general active learning paradigm and worth further investigation.

5.1 Extensions and Outlook

In learning from human preferences, the targets to be evaluated are trajectories instead of state-action pairs as in learning from human evaluative feedback. The advantage of evaluating trajectories is that the feedback is typically very sparse, resulting in a drastic reduction in human effort. However, there are two potential concerns in leveraging human preference. First, although the amount of feedback is less, the time or cognitive efforts in watching the two trajectories to make a preference choice could be more. Second, selecting the optimal trajectory length is challenging. Shorter trajectories allow humans to provide feedback of high granularity at the cost of more frequent human interactions. From the human trainer's perspective, the ideal trajectory length should be *subjective*, meaning that human trainers could select and adjust their preferred trajectory length during the training process. From the learning agent's perspective, the ideal length could also be *adaptive*, meaning that the agent could adaptively adjust the trajectory length to query humans to receive feedback of desired granularity.

Recent work by Bhatia et al. (2020) raised an important issue in preference learning: the human preference could be multi-criteria in nature, i.e., different behaviors are preferred under different criteria. For example, a driving policy τ_1 is preferred in terms of comfort, while τ_2 is preferred when speed is the only concern. Interestingly, a third policy τ_3 which is a linear combination of τ_1 and τ_2 may be preferred among all three when considering both criteria (Bhatia et al., 2020). The authors propose a novel framework to solve this problem by decomposing the single overall comparison and ask humans to provide preferences along simpler criteria (Bhatia et al., 2020). The authors lay the groundwork from a game-theoretic perspective but many questions are yet to be answered.

6 Hierarchical Imitation

Many sequential decision-making tasks are hierarchically structured, meaning that they can be decomposed into subtasks and solved using a divide-and-conquer approach. As an example from behavioral psychology, case studies with non-human primates have shown that fine-grained, low-level actions are mostly learned without imitation. In contrast, coarse, high-level "programs" learning is pervasive in imitation learning (Byrne and Russon, 1998). Program-level imitation is defined as imitating the high-level structural organization of a complex process, by observation of the behavior of another individual, while furnishing the exact details of actions by individual learning (Byrne and Russon, 1998), perhaps through reinforcement learning.

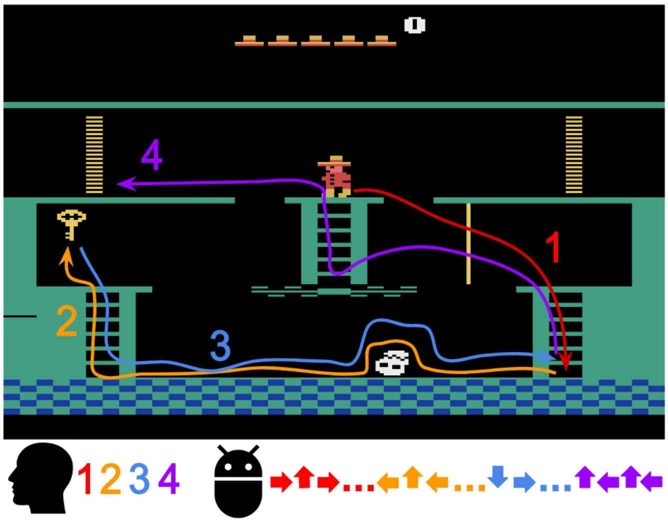


Fig. 7: In hierarchical imitation, the basic idea is to have the human trainer specify high-level goals. For example, red goal one is to reach the bottom of the ladder. An agent will learn to accomplish each high-level goal by performing a sequence of low-level actions potentially through reinforcement learning by itself.

Therefore, an interesting form of guidance can be provided by asking human trainers to provide only high-level feedback on these tasks. Similar to preference, this type of feedback also targets trajectory segments but is provided as choices of high-level goal in a given state, such as options[1]. Due to the hierarchical structure of the task, the behavior trajectory can be naturally segmented into options, instead of arbitrary segments in the preference framework. As shown in Fig. 4d and 7, using this type of human guidance, the learning problem for the agent is to learn a policy for choosing high-level goals in addition to learning a policy for low-level action selection.

6.1 Simulated Agents

Le et al. (2018) has proposed a 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. 4d). 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 were asked to provide three types of feedback: 1) a

[1] An option is a temporally extended action, or macro-action, which is composed of a policy, a termination condition, and an initiation set (Sutton et al., 1999).

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 if the chosen one 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. One approach is to train the agent to choose high-level goals via imitation learning, and let the agent learn low-level policies via RL by itself. This approach was shown to be substantially more sample efficient than conventional imitation learning. For example, Andreas et al. (2017) only required humans to provide policy sketches which are high-level symbolic subtask labels. The policy of each subtask is learned by the RL agent on its own and no longer requires human demonstration. A similar approach has been shown to be successful on Atari Montezuma's Revenge (Le et al., 2018). Another approach is to train both high-level and low-level policies via imitation learning, then fine-tune them using RL later (Gupta et al., 2020).

6.2 Physical Robots

Compared to simulated learning agents, in physical robot learning safety and sample efficiency are critical issues for IL and RL algorithms. Therefore, incorporating human prior knowledge through hierarchical task structuring to make robot learning tractable has been long studied and implemented. The classic approach is to define and represent the learning task at a more abstract level using human knowledge. As an example, actions can be defined as high-level goals such as “turn 90 degrees clockwise”, meanwhile fine-grained motor commands that accomplish these goals can be handled by low-level controllers. An early survey of previous robotic research on this topic is provided by Kober et al. (2013).

In contrast to providing only high-level goals for simulated agents (Andreas et al., 2017), in robotic tasks humans often need to provide low-level demonstrations due to safety and sample efficiency concerns. The works that leverage this type of demonstration can be roughly classified into *segmentation-based* or *non-segmentation-based* approaches depending on whether the task hierarchy is provided by humans.

In segmentation-based approaches, the task hierarchy is not provided, hence the aim is to extract meaningful segments from the low-level demonstration trajectories. For example, in contrast to Andreas et al. (2017), Krishnan et al. (2017) and Henderson et al. (2018a) set up the learning task in the opposite way which attempts to discover high-level options from low-level demonstration data. In this setting, only low-level demonstrations are collected, the options are latent variables of the trainer that can be inferred in a fashion similar to Expectation Maximization (Krishnan et al., 2017). Similarly, other methods aim to learn low-level primitives (Kipf et al., 2019; Sharma et al., 2018), latent conditioned policies (Hausman et al., 2017), goal-conditioned policies (Gupta et al., 2020), or skills (Konidaris et al., 2012; Kroemer et al., 2015) which meaningfully segment the low-level demonstrations. If information about high-level goals is also provided, they can be used to help infer meaningful segmentation boundaries. For example, Codevilla et al. (2018) has successfully combined high-level navigation commands with low-level control signals in a framework named conditional imitation learning for autonomous driving tasks.

In contrast, the task hierarchy can be provided explicitly or implicitly to eliminate the need for task segmentation. Humans can explicitly define the task hierarchy and feed such information to the robots (Mohseni-Kabir et al., 2015). Alternatively, humans can provide demonstrations subtask by subtask, therefore options, subtasks, subprograms, subroutines, or individual skills are learned first in isolation and then combined (Friesen and Rao, 2010).

A fixed, rigid task hierarchy has a poor ability to generalize. Recently, a framework named neural programming (NTP) has been developed that can decompose a demonstrated task into modular and reusable neural programs in a hierarchical manner (Reed and De Freitas, 2015; Li et al., 2016a; Xu et al., 2018a; Fox et al., 2018). The task hierarchy is only provided by humans during the training phase until the agent has learned to do task segmentation on its own. Neural programs are structured policies that perform algorithmic tasks by controlling the behavior of a computation mechanism (Fox et al., 2018). They are represented by neural networks that can learn to represent and execute compositional programs from demonstrations (Reed and De Freitas, 2015). The demonstration is still provided as low-level actions, the learning algorithm attempts to learn and reuse primitive network modules from the demonstration. A task manager, which is often a trainable task-agnostic network, decides which subprogram to run next and feeds task specification to the next program. Training this high-level manager requires ground-truth high-level task labels provided by human (Xu et al., 2018a). The low-level policy is represented as a neural program that takes a task specification as its input argument.

The benefit of this kind of hierarchical modular approach comes from the observation that in many robotic tasks there are shared components, and a learned task component (e.g., a particular skill) can often generalize across tasks. In a multitask setting, learned task components can be transferred between tasks so the required human demonstration effort could be drastically reduced (Fox et al., 2019; Xu et al., 2018b).

6.3 Extensions and Outlook

In some of the above robot learning works, asking humans to provide high-level actions requires additional human effort. However, in physical robot experiments, human annotation is often less costly and risky than demonstrations or teleoperations. If providing extra annotation can reduce the cost of using physical robots, such extra effort is justified and desirable (Fox et al., 2019).

We have seen that humans can either provide low-level action demonstrations, or high-level goals, or both types of guidance together. The choice that is suitable for a particular task domain depends on at least two factors. The first concern is the relative effort in specifying goals vs. providing demonstrations. High-level goals are often clear and easy to be specified in tasks such as navigation (Andreas et al., 2017). On the contrary in tasks like Tetris providing low-level demonstration is easier since high-level goals are not easy to represent and be communicated. The second concern is safety and sample efficiency. Only providing high-level goals requires the agents to learn low-level policies by themselves through trial-and-error, perhaps with many more samples, which is suitable for simulated agents but not for physical robots. Therefore in robotic tasks, low-level action demonstrations are often required.

One way to further reduce human effort in hierarchical imitation learning is to leverage evaluative feedback. Evaluative feedback can be naturally incorporated in the hierarchical imitation learning framework, in which human trainers provide evaluative feedback (yes or no) on either high-level or low-level actions of the learning agent, an approach that has been partially explored by Mohseni-Kabir et al. (2015) and (Le et al., 2018). Moreover, 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).

Hierarchical imitation learning is naturally related to hierarchical reinforcement learning (Sutton et al., 1999; Dietterich, 2000; Barto and Mahadevan, 2003), which is an active research field with its own exciting progress (Kulkarni et al., 2016; Bacon et al., 2017; Vezhnevets et al., 2017; Nachum et al., 2018). Another closely related research field here is multi-agent reinforcement learning (Ghavamzadeh et al., 2006) since the multi-agent setting implicitly contains a two-level hierarchy: One at the individual agent's level and the other at the group level. For a recent survey on this topic, please see Hernandez-Leal et al. (2019). A potential research direction is to leverage human guidance, in the forms of demonstration or feedback, in the settings of hierarchical RL or multi-agent learning systems. As we have seen in Le et al. (2018), a reasonable starting point is to ask humans to demonstrate or evaluate high-level decisions.

7 Imitation from Observation

Imitation from observation (IfO) (Torabi et al., 2019d) is the problem of learning directly by observing a trainer performing the task. The learning agent only has access to state demonstrations (e.g. in the form of visual observations) of the trainer (Fig. 4e and 8). Using this type of human guidance, the learning problem for the agent is to learn a policy from the state sequences demonstrated by the human.

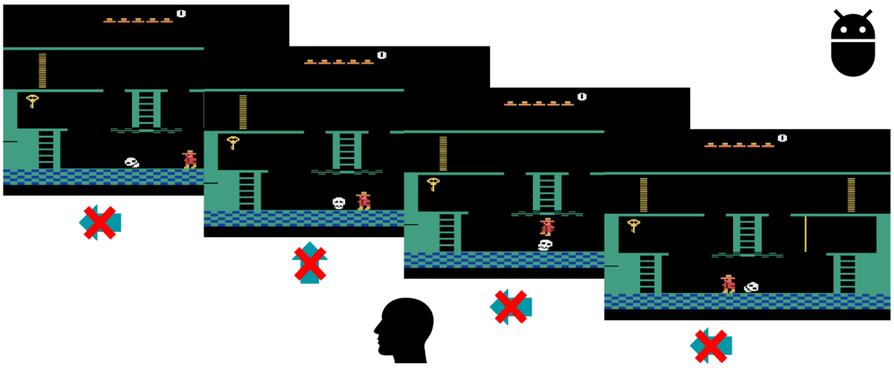


Fig. 8: In imitation from observation, the setting is very much like standard imitation learning (Fig. 2), except the agent does not have access to the actions demonstrated by the human trainer.

This framework is different from conventional imitation learning in the sense that it eschews the requirement for action labels in demonstrations. Removing this constraint enables imitating agents to use a large amount of previously ignored available demonstration data such as videos on YouTube. The ultimate goal in this framework is to enable agents to utilize the existing, rich amount of demonstration data that do not have action labels, such as the human guidance provided through online videos of humans performing various tasks.

Broadly speaking, there are two major components of the IfO problem: *(1)* perception, and *(2)* control.

7.1 Perception

Because IfO depends on observations of an expert agent, processing these observations perceptually is extremely important. Previous works have used multiple approaches for this part of the problem. One such approach is to record the expert's movements using sensors placed directly on the expert agent (Ijspeert et al., 2011). Using these recordings, techniques have been proposed that can allow humanoid or anthropomorphic robots to mimic human motions, e.g., arm-reaching movements (Ijspeert et al., 2002; Bentivegna et al., 2002), biped locomotion (Nakanishi et al., 2004), and human gestures (Calinon and Billard, 2007). Another approach to the perception problem is that of motion capture (Field et al., 2009), which typically uses visual markers on the demonstrator to infer movement. IfO techniques built upon this approach have been used for a variety of tasks, including locomotion, acrobatics, and martial arts (Peng et al., 2018a; Merel et al., 2017; Setapen et al., 2010). The methods discussed above often require costly instrumentation and preprocessing (Holden et al., 2016), and therefore cannot be used in conjunction with more passive resources such as YouTube videos.

Recently, however, convolutional neural networks and advances in visual recognition have provided promising tools to work towards visual imitation where the expert demonstration consists of raw video information (e.g., pixel color values) alone. Even with such tools, the imitating agent is still faced with several challenges: *(1)* embodiment mismatch, and *(2)* viewpoint difference. Embodiment mismatch arises when the demonstrating agent has a different embodiment from that of the imitator. For example, the video could be of a human performing a task, but the goal may be to train a robot to do the same. Since humans and robots do not look exactly alike (and may look quite different), the challenge is in how to interpret the visual information such that IfO can be successful. One IfO method developed to address this problem learns a correspondence between the embodiments using autoencoders in a supervised fashion (Gupta et al., 2018). The autoencoder is trained in such a way that the encoded representations are invariant with respect to the embodiment features. Another method learns the correspondence in an unsupervised fashion with a small amount of human supervision (Sermanet et al., 2018). The second IfO perceptual challenge is the viewpoint difference that arises when demonstrations are not recorded in a controlled environment. For instance, the video background may be cluttered, or there may be a mismatch in the point of view present in the demonstration video and that with which the agent sees itself. One IfO approach that attempts to address this issue learns a context translation model to translate an observation by predicting it in the target context (Liu et al., 2018). The translation is learned using data that consists of images of the target context and the source context, and the task is to translate the frame from the source context to that of the target. 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 (Stadie et al., 2017). Consequently, the extracted features can be invariant with respect to the viewpoint.

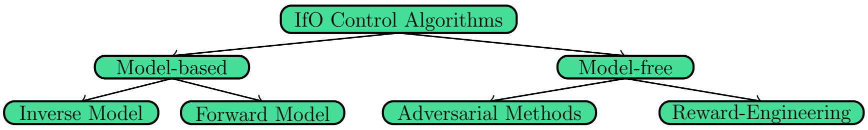


Fig. 9: A diagrammatic representation of categorization of the IfO control algorithm. The algorithms can be categorized into two groups: (1) model-based algorithms in which the algorithms may use either a forward dynamics model (Edwards et al., 2018) or an inverse dynamics model (Torabi et al., 2018a; Nair et al., 2017). (2) Model-free algorithms, which itself can be categorized into adversarial methods (Torabi et al., 2018b; Merel et al., 2017; Stadie et al., 2017) and reward engineering (Sermanet et al., 2018; Gupta et al., 2018; Liu et al., 2018).

7.2 Control

Another main component of IfO is control, i.e., the approach used to learn the imitation policy, typically under the assumption that the agent has access to clean state demonstration data $\{s_t\}$. Since the action labels are not available, this is a very challenging problem, and many approaches have been discussed in the literature. Previously, this problem was referred to as *trajectory tracking* where the goal was to follow a time parameterized reference (Yang and Kim, 1999; Aguiar and Hespanha, 2007). Most of the algorithms developed for trajectory tracking do not involve machine learning at all and require the state features and reference points to be well-defined such as joint angles, velocities, etc. (Yang and Kim, 1999; Caracciolo et al., 1999). Therefore, it is not clear how to directly scale these algorithms to visual imitation (imitating directly from raw pixel data).

With the rise of deep learning, however, many new learning-based algorithms have recently been proposed to tackle the IfO control problem. We organize them here into two general groups: *(1)* model-based algorithms, and *(2)* model-free algorithms. Model-based approaches to IfO are characterized by the fact that they learn some type of dynamics model during the imitation process. Most of these algorithms learn an inverse dynamics model which is a mapping from state-transitions $\{(s_t, s_{t+1})\}$ to actions $\{a_t\}$ (Hanna and Stone, 2017). The goal of these algorithms is to retrieve the missing demonstration action labels. To do so, they interact with the environment, collect state action data, and then learn an inverse dynamics model. Applying this learned model on two consecutive demonstrated states would output the missing taken action that had resulted in that state transition. After retrieving the actions, the learning problem can be treated as a conventional imitation learning problem. Recently, many algorithms are developed with this high-level idea (Nair et al., 2017; Torabi et al., 2018a; Pavse et al., 2020; Pathak et al., 2018; Guo et al., 2019; Robertson and Walter, 2020; Jiang et al., 2020; Radosavovic et al., 2020). Some other algorithms use forward dynamics model instead which is a mapping from state-action pairs, $\{(s_t, a_t)\}$, to the next states, $\{s_{t+1}\}$. One algorithm of this type is developed by Wu et al. [2020] in which a forward dynamics model is learned which is used to predict the future state of the agent and then future state similarity is used to learn an imitation policy. There is another algorithm (Edwards et al., 2018) that learns *forward* dynamics model. This algorithm hypothesizes that the state transitions are caused by the actions taken by the agent. The actions are unknown and therefore the algorithm considers a latent (unreal) action space and learns a policy in that latent space that best describes the state transitions. Since the actions generated by this learned policy are not real, next the agent takes a few interactions with the environment to make corrections to the action labels. To be more specific, this algorithm creates an initial hypothesis for the imitation policy by learning a latent policy $\pi(z|s_t)$ that estimates the probability of latent (unreal) action z given the current state s_t. Since actual actions are not needed, this process can be done offline without any interaction with the environment. To learn the latent policy, they use a latent forward dynamics model which predicts s_{t+1} and a prior over z given s_t. Then they use a limited number of environment interactions to learn an action-remapping network that associates the latent actions with their corresponding correct actions. Since most of the process happens offline, the algorithm is efficient with regard to the number of interactions needed.

The other broad category of IfO control approaches is that of model-free algorithms. Model-free techniques attempt to learn the imitation policy without any sort of model-learning step. Within this category, there are two fundamentally different types of algorithms. One is adversarial methods which are inspired by the generative adversarial imitation learning (*GAIL*) algorithm described in Section 2.2. In *GAIL*, the goal is to bring the state-action distribution of the imitator close to that of the demonstrator. However, since in IfO the imitator does not have access to the actions, the proposed algorithms attempt to bring the state distribution (Merel et al., 2017; Henderson et al., 2018a), or state transition distribution (Stadie et al., 2017; Torabi et al., 2018b, 2019b,c,a; Zolna et al., 2018; Sun et al., 2019; Yang et al., 2019; Chaudhury et al., 2019) of the imitator close to that of the demonstrator. The overall scheme of these algorithms is as follows. They use a *GAN*-like architecture in which the imitation policy is interpreted as the generator. The imitation policy is executed in the environment to collect data, $\{(s_t^i, a_t^i)\}$, and either the states or the state transitions are fed into the discriminator, which is trained to differentiate between the data that comes from the imitator and data that comes from the demonstrator. The output value of the discriminator is then used as a reward to update the imitation policy using RL. Another class of model-free approaches developed for IfO control is that utilizes reward engineering. Here, reward engineering means that, based on the expert demonstrations, a manually designed reward function is used to find imitation policies via *RL*. Importantly, the designed reward functions are not necessarily the ones that the demonstrator used to produce the demonstrations—rather, they are simply estimates inferred from the demonstration data. Most of the algorithms of this type use the negative of the Euclidean distance of the states of the imitator and the demonstrator (or an embedded version of them) as the reward at each time step (Kimura et al., 2018; Sermanet et al., 2018; Dwibedi et al., 2018; Gupta et al., 2018; Liu et al., 2018, 2020). Another approach of this type is developed by Goo and Niekum (2019) in which the algorithm uses a formulation similar to shuffle-and-learn Misra et al. (2016) to train a neural network that learns the order of frames in the demonstration. The network in a supervised fashion gets two observations and outputs a value between zero and one. The closer the value to one, the higher the chance of observations being in the right order. This neural network is then used as a surrogate reward function to train a policy. Aytar et al. (2018) also take a similar approach, learning an embedding function for the video frames based on the demonstration. They use the closeness between the imitator's embedded states and some checkpoint embedded features as the reward function.

7.3 Extensions and Outlook

Regarding the perception component of the IfO problem, adversarial training techniques have led to several recent and exciting advances in the computer vision community. One such advance is in the area of pose estimation (Cao et al., 2017; Wang et al., 2019), which enables detection of the position and orientation of the objects in a cluttered video through keypoint detection—such keypoint information may also prove useful in IfO. While there has been a small amount of effort to incorporate these advances in IfO (Peng et al., 2018b), there is still much to investigate.

Another recent advancement in computer vision is in the area of visual domain adaptation (Wang and Deng, 2018), which is concerned with transferring learned knowledge to different visual contexts. For instance, the recent success of CycleGAN (Zhu et al., 2017) suggests that modified adversarial techniques may be applicable to IfO problems that require solutions to embodiment mismatch, though it remains to be seen if such approaches will truly lead to advances in IfO.

Regarding the control component of the IfO problem, very few of the mentioned IfO algorithms discussed have been successfully tested on physical robots, such as Sermanet et al. (2018); Liu et al. (2018). That is, most discuss results only in simulated domains. For instance, while adversarial control methods currently provide state-of-the-art performance for several baseline experimental IfO problems, these methods exhibit high sample complexity and have therefore only been applied to relatively simple simulation tasks. Thus, an open problem in IfO is that of finding ways to adapt these techniques such that they can be used in scenarios for which high sample complexity is prohibitive, i.e., tasks in robotics. Furthermore, there have been few works investigating the combination of IfO with other learning frameworks (Brown et al., 2019; Pavse et al., 2020; Schmeckpeper et al., 2020). There is room to investigate how different types of learning paradigms could be incorporated in IfO to improve the overall task learning performance.

8 Learning Attention from Humans

During the human demonstration or evaluation process, there are other useful learning signals. One useful signal is human visual attention, which can be treated as a form of guidance that reveals important task features to the learning agent. For decision tasks with high-dimensional visual information as input, humans visual attention is revealed by eye movements, i.e., gaze behaviors. Gaze is an informative source of *(1)* important state features in high-dimensional state space at a given time *(2)* the explanatory information that reveals the target or goal of an observed action. For the former, 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. For the latter, knowing which visual object the human trainer

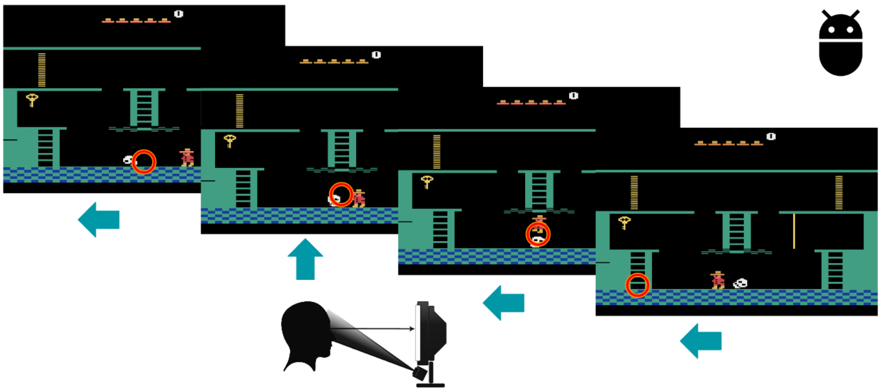


Fig. 10: In learning attention from humans, the agent has access to human attention information in addition to the action demonstrations. The eye movement data indicated by the red circles here can be recorded by an eye tracker. This data reveals the current behavioral goal (such as the object of interest, e.g., the skull and the ladder) when taking an action.

looked at while making a decision can help to explain why a particular decision was made. For these reasons, learning attention from humans could help a learning agent extract useful features from a high-dimensional state space and understand the underlying causes of a human trainer's demonstrated action. This approach has recently become very popular as learning agents migrate from simple tasks to challenging sequential decision-making tasks with high-dimensional inputs (Fig. 3).

The gaze data can be collected with an eye tracker while the human trainer is demonstrating the task (Fig. 4f and 10). Recently, researchers have collected human gaze and policy data for meal preparation (Li et al., 2018), Atari game playing (Zhang et al., 2020b), human-to-human (non-verbal) interactions (Zuo et al., 2018), and outdoor driving (Palazzi et al., 2018). Using this type of human guidance, the learning problem for the agent is to learn the attention mechanism from humans in addition to learning a decision policy.

8.1 Attention Learning

The first learning objective using these datasets could be training an agent to imitate human gaze behaviors, i.e., learning to attend to certain features of a given image. The problem was formalized as a visual saliency prediction problem in computer vision research (Itti et al., 1998). Recently this area has made tremendous progress due to deep learning as large-scale eye-tracking datasets became available for images (Papadopoulos et al., 2014; Li et al., 2014; Xu et al., 2014; Bylinskii et al., 2015b,a; Krafka et al., 2016), videos (Mathe and Sminchisescu, 2014; Wang et al., 2018), and 360-degree videos (Zhang et al., 2018b; Xu et al., 2018c). Visual saliency is a well-developed field in computer vision. We direct interested readers to recent review papers on the topics of saliency evaluation metrics (Bylinskii et al., 2019), saliency model performance analyses (Bylinskii et al., 2016; He et al., 2019) and a closely related field called salient object detection (Borji et al., 2015).

In our context of sequential decision-making tasks, the saliency prediction problem can be formalized as follows:

> Given a state s_t, learn to predict human gaze positions w_t, i.e., learn $P(w|s)$.

Note that w_t could be a set of positions since the human can look at multiple regions of the image. In practice, discrete human gaze positions are converted into a continuous distribution (Bylinskii et al., 2019). So the agent should learn to predict this probability distribution over the given image. This can be done using supervised learning where Kullback-Leibler divergence can be used as the loss function to calculate the difference between the ground truth distribution P and predicted distribution Q (Bylinskii et al., 2019):

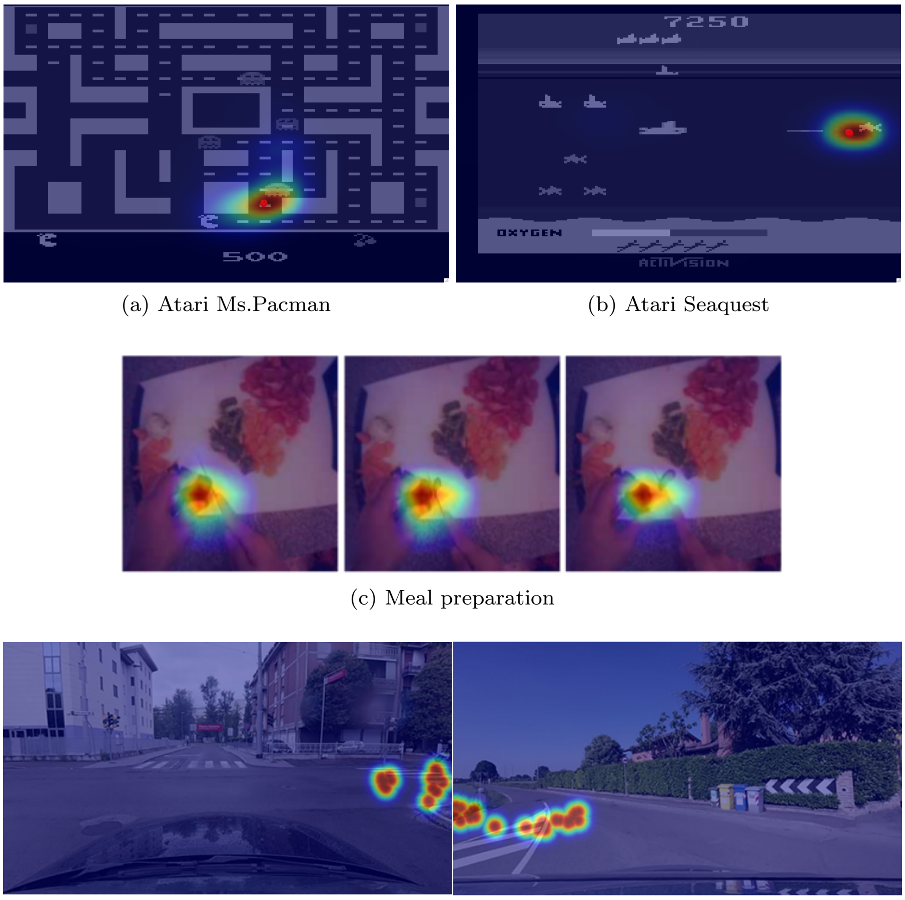
$$KL(P, Q) = \sum_i \sum_j Q(i, j) \log \left(\epsilon + \frac{Q(i, j)}{\epsilon + P(i, j)} \right) \tag{11}$$

where i, j are pixels indices and ϵ is a small regularization constant and determines how much zero-valued predictions are penalized. Recent works have trained convolutional neural networks to accomplish this learning task (Li et al., 2018; Zhang et al., 2020b; Palazzi et al., 2018; Deng et al., 2019; Chen et al., 2020). Example gaze prediction results in the format of saliency maps can be seen in Fig. 11. A notable challenge here is *egocentric* gaze prediction in which the spatial distribution of the gaze is highly biased towards the image center, a problem further addressed by Palazzi et al. (2018); Tavakoli et al. (2019).

8.2 Decision Learning

In computer vision, traditional saliency prediction does not involve active tasks nor human decisions. The humans look at static images or videos in a free-viewing manner without performing any particular task and only the eye movements are recorded and modeled. Meanwhile, the aforementioned datasets all require humans to perform a task while collecting their gaze and action data. From a decision-learning perspective, human attention may provide additional information about their decisions, therefore it is intuitive to leverage learned attention models to guide the learning process of human decisions. The learning problem can be formalized as follows:

> Given a state s_t and human gaze positions w_t, learn to predict human action a_t, i.e., learn $P(a|s, w)$.



(a) Atari Ms.Pacman

(b) Atari Seaquest

(c) Meal preparation

(d) Driving

(e) Driving

Fig. 11: Learning attention from human in game playing (Zhang et al., 2018a), meal preparation (Li et al., 2018), and driving (Palazzi et al., 2018). The heatmaps show the agent's prediction of human attention, represented as saliency maps (probability distribution of attention) overlayed on the images. Red indicates regions that have high predicted probability to be attended by humans.

Intuitively, knowing where humans would look provides useful information on what action they will take. To incorporate human attention into action learning, there are at least three common methods: as an additional channel of information, as a mask on the input to filter out unimportant information, or as a secondary optimization objective. For example, in training a neural network, the above methods correspond to concatenating a gaze map with the input image, masking the input image with the gaze map, and adding gaze prediction as an auxiliary loss term in the objective function, respectively (Zhang et al., 2020a). The most popular way is to treat the predicted gaze distribution of an image as a filter or a mask. This mask can be applied to the image to generate a representation of the image that highlights the attended visual features.

Experimental results have shown that including gaze information leads to higher accuracy in recognizing or predicting human actions, in reaching (Ravichandar et al., 2018), human-to-human interaction (Zuo et al., 2018), driving (Xia et al., 2018; Liu et al., 2019; Chen et al., 2019; Xia et al., 2020), meal preparation (Li et al., 2018; Shen et al., 2018; Sudhakaran et al., 2019; Huang et al., 2020), and video game playing (Zhang et al., 2018a, 2020b).

Once the agent has learned both the attention and decision models from human data, it can perform the task on its own. It has been shown that incorporating a learned gaze model into imitation learning agents leads to a large performance increase, comparing to agents without attention information (Zhang et al., 2020b; Saran et al., 2020; Chen et al., 2020). For real-world tasks like autonomous driving, it is reasonable to expect a similar improvement when incorporating human attention models. Due to physical constraints and safety reasons, this is yet to be explored but preliminary tests in simulated environments are possible.

8.3 Extensions and Outlook

In general, human gaze is a good indicator of the underlying decision-making mechanism, it bridges perception and decision-making by indicating the current behavioral target. The gaze data can be collected in parallel with actions. One concern with this approach is the hardware and software required to collect human gaze data. Recent progress in computer vision has improved eye tracker accuracy and portability by a significant margin. Appearance-based algorithms using convolutional neural networks have been shown to have better tracking accuracy and are more robust to visual appearance variations (Zhang et al., 2015; Wood et al., 2015; Krafka et al., 2016; Shrivastava et al., 2017; Zhang et al., 2017; Park et al., 2018), compared to more traditional approaches like hand-crafted feature-based or model-based algorithms. Advanced tracking software can estimate gaze in real-time from head poses and appearance without specialized hardware on low-cost devices such as webcams (Papoutsaki et al., 2016) and mobile tablets and phones (Huang et al., 2017; Krafka et al., 2016).

Gaze data can be collected in parallel when providing other types of feedback, and potentially be combined with previously introduced learning methods. Saran et al. (2020) has shown that incorporating gaze information into imitation from observation (IfO) and inverse reinforcement learning can lead to a large performance increase in Atari games. Since attention is an intermediate mechanism between perception and action, it becomes very useful when action information is missing in the case of IfO. In learning evaluative feedback and preference, gaze data might reveal more information to the learning agent to explain why the human gives a particular evaluation. 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 this goal. However, the problem of inferring behavioral goals from human attention needs to be solved first.

9 Conclusion and Future Directions

In this survey, we have provided a literature review of progress in leveraging five different types of human guidance (i.e., human inputs that do not involve explicitly defining a reward function or providing an action demonstration) to solve sequential decision-making tasks. In particular, we discussed techniques that have been proposed in the literature that learn from human-provided evaluative feedback, preference, goals, action-free demonstrations, and attention. In each section above, we have discussed future research directions for each approach. Here we briefly discuss several issues and associated potential research questions that are common to all the approaches that leverage human guidance as learning signals.

9.1 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 and reusing such datasets may be difficult for some interactive learning methods, in which the guidance (such as evaluative feedback) depends on the changing policy as it is being learned. But, for other approaches, data can be collected in advance and shared. In Table 1 we provide links to existing datasets that are publicly available. Another concern is reproducibility in RL (Henderson et al., 2018b). 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 the final performance. Therefore, evaluating algorithms using a standard dataset could save effort and assure a fair comparison between algorithms.

9.2 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 (2008) pioneered the effort in understanding human behavior in teaching learning agents. 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 this aspect of human behavior, 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. (2016) 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 algorithms can benefit substantially more from the data collected in the showing mode (Ho et al., 2016). An important factor to consider is the human trainer's knowledge of the task. Laskey et al. (2016) have shown that using a hierarchy of human supervisors with different expertise levels can substantially reduce the burden on the experts.

Understanding the variations in human guidance signals 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).

9.3 An Interactive Paradigm

The best learning results often come from an interactive teaching and learning process in a teacher-student paradigm which involves active instruction by the human and active learning by the agent. Two factors justify an interactive learning paradigm in our context: *(1)* We only have a partial understanding of human guidance behaviors; and *(2)* the nature of human guidance may vary during training according to the behaviors of the learning agents. As shown by Cooperative IRL (Hadfield-Menell et al., 2016), an iterative and interactive learning process can greatly enhance learning. Therefore the same idea may benefit the process of learning from human guidance as well.

For evaluative feedback, we have seen a debate on how we interpret human feedback. We have also seen methods that are robust to many types of feedback or that can infer and adapt to different human feedback types (Grizou et al., 2014; Loftin et al., 2016; Najar et al., 2020). However, perhaps an alternative is to allow the agent to actively query humans for a certain type of feedback that best informs the agent. Learning from human preference is naturally an interactive process when the learning agents actively query for human preferences as we have discussed. Additionally, the aforementioned challenge of selecting optimal trajectory length for query likely requires two-way communication between the human and the agent. In hierarchical imitation, imitation from observation, and attention learning we rarely see examples of interactive learning (Mohseni-Kabir et al., 2015), since the human data is often collected off-line (see Table 1) before training as in standard imitation learning. From DAgger (Ross et al., 2011) it is clear that interactive training can also benefit imitation learning, however, making the three learning paradigms above interactive remains to be explored.

9.4 A Unified Learning Framework

The learning frameworks discussed in this paper are often inspired by real-life biological learning scenarios that correspond to different 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 the early stages of learning and the latter could be more useful in the late stages. The other learning strategies discussed are often mixed with these two to allow an agent to utilize signals from all possible sources. For example, it is widely known that children learn largely by imitation and observation (Bandura et al., 1961) at their early stage of learning. Then the children gradually learn to develop joint attention with adults through gaze following (Goswami, 2008). Later children begin to adjust their behaviors based on the evaluative feedback and preference received when interacting with other people. Once they developed the ability to reason abstractly about task structure, hierarchical imitation becomes feasible. At the same time, learning through trial and error from reinforcement is always one of the most common types of learning (Skinner, 1938). The human's ability to learn from all types of resources continue to develop through a lifetime. We have compared these learning strategies 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. We start to notice efforts towards this goal (Abel et al., 2017; Waytowich et al., 2018; Goecks et al., 2019; Woodward et al., 2020; Najar et al., 2020; Bıyık et al., 2020).

In conclusion, the goal of this survey is to serve as a high-level overview of five recent learning frameworks that leverage human guidance to solve sequential decision-making tasks, especially in the context of deep reinforcement learning. We compare and contrast these frameworks by reviewing the motivation, assumption, and implementation of each framework. We hope this will allow researchers in the related areas to see the connections between the works being surveyed, and inspire more research to be done in this field.

References

Abbeel P, Ng AY (2004) Apprenticeship learning via inverse reinforcement learning. In: Proceedings of the Twenty-First International Conference on Machine learning (ICML), ACM, p 1

Abbeel P, Coates A, Ng AY (2010) Autonomous helicopter aerobatics through apprenticeship learning. The International Journal of Robotics Research 29(13):1608–1639

Abel D, Salvatier J, Stuhlmüller A, Evans O (2017) Agent-agnostic human-in-the-loop reinforcement learning. NeurIPS Workshop on the Future of Interactive Learning Machines

Aguiar AP, Hespanha JP (2007) Trajectory-tracking and path-following of underactuated autonomous vehicles with parametric modeling uncertainty. IEEE Transactions on Automatic Control 52(8):1362–1379

Akinola I, Wang Z, Shi J, He X, Lapborisuth P, Xu J, Watkins-Valls D, Sajda P, Allen P (2020) Accelerated robot learning via human brain signals. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 3799–3805

Akrour R, Schoenauer M, Sebag M, Souplet JC (2014) Programming by feedback. In: International Conference on Machine Learning (ICML), JMLR. org, vol 32, pp 1503–1511

Amir O, Kamar E, Kolobov A, Grosz BJ (2016) Interactive teaching strategies for agent training. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, AAAI Press, pp 804–811

Andreas J, Klein D, Levine S (2017) Modular multitask reinforcement learning with policy sketches. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, pp 166–175

Arakawa R, Kobayashi S, Unno Y, Tsuboi Y, Maeda Si (2018) Dqn-tamer: Human-in-the-loop reinforcement learning with intractable feedback. arXiv preprint arXiv:181011748

Argall BD, Chernova S, Veloso M, Browning B (2009) A survey of robot learning from demonstration. Robotics and autonomous systems 57(5):469–483

Arora S, Doshi P (2018) A survey of inverse reinforcement learning: Challenges, methods and progress. arXiv preprint arXiv:180606877

Arumugam D, Lee JK, Saskin S, Littman ML (2019) Deep reinforcement learning from policy-dependent human feedback. arXiv preprint arXiv:190204257

Aytar Y, Pfaff T, Budden D, Paine T, Wang Z, de Freitas N (2018) Playing hard exploration games by watching youtube. In: Advances in Neural Information Processing Systems, pp 2935–2945

Bacon PL, Harb J, Precup D (2017) The option-critic architecture. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp 1726–1734

Bain M, Sommut C (1999) A framework for behavioural cloning. Machine intelligence 15(15):103

Bandura A, Ross D, Ross SA (1961) Transmission of aggression through imitation of aggressive models. The Journal of Abnormal and Social Psychology 63(3):575

Barto AG, Mahadevan S (2003) Recent advances in hierarchical reinforcement learning. Discrete event dynamic systems 13(1-2):41–77

Bellemare MG, Naddaf Y, Veness J, Bowling M (2013) The arcade learning environment: An evaluation platform for general agents. Journal of Artificial

Intelligence Research 47:253–279

Bentivegna DC, Ude A, Atkeson CG, Cheng G (2002) Humanoid robot learning and game playing using pc-based vision. In: IEEE/RSJ international conference on intelligent robots and systems, IEEE, vol 3, pp 2449–2454

Bestick A, Pandya R, Bajcsy R, Dragan AD (2018) Learning human ergonomic preferences for handovers. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 1–9

Bhatia K, Pananjady A, Bartlett P, Dragan A, Wainwright MJ (2020) Preference learning along multiple criteria: A game-theoretic perspective. Advances in Neural Information Processing Systems 33

Biyik E, Sadigh D (2018) Batch active preference-based learning of reward functions. In: Conference on Robot Learning, pp 519–528

Bıyık E, Lazar DA, Sadigh D, Pedarsani R (2019) The green choice: Learning and influencing human decisions on shared roads. In: 2019 IEEE 58th Conference on Decision and Control (CDC), IEEE, pp 347–354

Biyik E, Huynh N, Kochenderfer MJ, Sadigh D (2020) Active preference-based gaussian process regression for reward learning. In: Proceedings of Robotics: Science and Systems (RSS), DOI 10.15607/rss.2020.xvi.041

Bıyık E, Losey DP, Palan M, Landolfi NC, Shevchuk G, Sadigh D (2020) Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences. arXiv preprint arXiv:200614091

Bloem M, Bambos N (2014) Infinite time horizon maximum causal entropy inverse reinforcement learning. In: 53rd IEEE Conference on Decision and Control, IEEE, pp 4911–4916

Bojarski M, Del Testa D, Dworakowski D, Firner B, Flepp B, Goyal P, Jackel LD, Monfort M, Muller U, Zhang J, et al. (2016) End to end learning for self-driving cars. arXiv preprint arXiv:160407316

Borji A, Cheng MM, Jiang H, Li J (2015) Salient object detection: A benchmark. IEEE transactions on image processing 24(12):5706–5722

Broekens J (2007) Emotion and reinforcement: affective facial expressions facilitate robot learning. In: Artifical intelligence for human computing, Springer, pp 113–132

Brown D, Goo W, Nagarajan P, Niekum S (2019) Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In: International Conference on Machine Learning, pp 783–792

Busa-Fekete R, Szörényi B, Weng P, Cheng W, Hüllermeier E (2013) Preference-based evolutionary direct policy search. In: ICRA Workshop on Autonomous Learning

Bylinskii Z, Isola P, Bainbridge C, Torralba A, Oliva A (2015a) Intrinsic and extrinsic effects on image memorability. Vision research 116:165–178

Bylinskii Z, Judd T, Borji A, Itti L, Durand F, Oliva A, Torralba A (2015b) Mit saliency benchmark

Bylinskii Z, Recasens A, Borji A, Oliva A, Torralba A, Durand F (2016) Where should saliency models look next? In: European Conference on Computer Vision, Springer, pp 809–824

Bylinskii Z, Judd T, Oliva A, Torralba A, Durand F (2019) What do different evaluation metrics tell us about saliency models? IEEE transactions on pattern analysis and machine intelligence 41(3):740–757

Byrne RW, Russon AE (1998) Learning by imitation: A hierarchical approach. Behavioral and brain sciences 21(5):667–684

Calinon S, Billard A (2007) Incremental learning of gestures by imitation in a humanoid robot. In: Proceedings of the ACM/IEEE international conference on Human-robot interaction, pp 255–262

Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7291–7299

Caracciolo L, De Luca A, Iannitti S (1999) Trajectory tracking control of a four-wheel differentially driven mobile robot. In: Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C), IEEE, vol 4, pp 2632–2638

Cederborg T, Grover I, Isbell CL, Thomaz AL (2015) Policy shaping with human teachers. In: Proceedings of the 24th International Conference on Artificial Intelligence, AAAI Press, pp 3366–3372

Chaudhury S, Kimura D, Munawar A, Tachibana R (2019) Injective state-image mapping facilitates visual adversarial imitation learning. In: 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), IEEE, pp 1–6

Chen Y, Liu C, Tai L, Liu M, Shi BE (2019) Gaze training by modulated dropout improves imitation learning. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 7756–7761

Chen Y, Liu C, Shi BE, Liu M (2020) Robot navigation in crowds by graph convolutional networks with attention learned from human gaze. IEEE Robotics and Automation Letters 5(2):2754–2761

Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D (2017) Deep reinforcement learning from human preferences. In: Advances in Neural Information Processing Systems, pp 4299–4307

Codevilla F, Miiller M, López A, Koltun V, Dosovitskiy A (2018) End-to-end driving via conditional imitation learning. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 1–9

Cui Y, Niekum S (2018) Active reward learning from critiques. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 6907–6914

Cui Y, Zhang Q, Allievi A, Stone P, Niekum S, Knox WB (2020) The empathic framework for task learning from implicit human feedback. arXiv preprint arXiv:200913649

Deng T, Yan H, Qin L, Ngo T, Manjunath B (2019) How do drivers allocate their potential attention? driving fixation prediction via convolutional neural networks. IEEE Transactions on Intelligent Transportation Systems 21(5):2146–2154

Dietterich TG (2000) Hierarchical reinforcement learning with the maxq value function decomposition. Journal of artificial intelligence research 13:227–303

Dwibedi D, Tompson J, Lynch C, Sermanet P (2018) Learning actionable representations from visual observations. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp 1577–1584

Edwards AD, Sahni H, Schroeker Y, Isbell CL (2018) Imitating latent policies from observation. arXiv preprint arXiv:180507914

Fang B, Jia S, Guo D, Xu M, Wen S, Sun F (2019) Survey of imitation learning for robotic manipulation. International Journal of Intelligent Robotics and Applications pp 1–8

Field M, Stirling D, Naghdy F, Pan Z (2009) Motion capture in robotics review. In: 2009 IEEE International Conference on Control and Automation, IEEE, pp 1697–1702

Finn C, Levine S, Abbeel P (2016) Guided cost learning: Deep inverse optimal control via policy optimization. In: International Conference on Machine Learning, pp 49–58

Fox R, Shin R, Krishnan S, Goldberg K, Song D, Stoica I (2018) Parametrized hierarchical procedures for neural programming. International Conference on Learning Representations 2018

Fox R, Berenstein R, Stoica I, Goldberg K (2019) Multi-task hierarchical imitation learning for home automation. In: 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), IEEE, pp 1–8

Friesen AL, Rao RP (2010) Imitation learning with hierarchical actions. In: 2010 IEEE 9th International Conference on Development and Learning, IEEE, pp 263–268

Fu J, Luo K, Levine S (2018) Learning robust rewards with adverserial inverse reinforcement learning. In: International Conference on Learning Representations, URL `https://openreview.net/forum?id=rkHywl-A-`

Fürnkranz J, Hüllermeier E, Cheng W, Park SH (2012) Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. Machine learning 89(1-2):123–156

Ghavamzadeh M, Mahadevan S, Makar R (2006) Hierarchical multi-agent reinforcement learning. Autonomous Agents and Multi-Agent Systems 13(2):197–229

Giusti A, Guzzi J, Cireşan DC, He FL, Rodríguez JP, Fontana F, Faessler M, Forster C, Schmidhuber J, Di Caro G, et al. (2016) A machine learning approach to visual perception of forest trails for mobile robots. IEEE Robotics and Automation Letters 1(2):661–667

Goecks VG, Gremillion GM, Lawhern VJ, Valasek J, Waytowich NR (2019) Efficiently combining human demonstrations and interventions for safe training of autonomous systems in real-time. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 2462–2470

Goo W, Niekum S (2019) One-shot learning of multi-step tasks from observation via activity localization in auxiliary video. In: 2019 International Conference on Robotics and Automation (ICRA), IEEE, pp 7755–7761

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680

Goswami U (2008) Cognitive development: The learning brain. Psychology Press

Griffith S, Subramanian K, Scholz J, Isbell CL, Thomaz AL (2013) Policy shaping: Integrating human feedback with reinforcement learning. In: Advances in neural information processing systems, pp 2625–2633

Grizou J, Iturrate I, Montesano L, Oudeyer PY, Lopes M (2014) Interactive learning from unlabeled instructions. In: Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, pp 290–299

Grondman I, Busoniu L, Lopes GA, Babuska R (2012) A survey of actor-critic reinforcement learning: Standard and natural policy gradients. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42(6):1291–1307

Guo X, Chang S, Yu M, Tesauro G, Campbell M (2019) Hybrid reinforcement learning with expert state sequences. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 3739–3746

Gupta A, Devin C, Liu Y, Abbeel P, Levine S (2018) Learning invariant feature spaces to transfer skills with reinforcement learning. In: International Conference on Learning Representations

Gupta A, Kumar V, Lynch C, Levine S, Hausman K (2020) Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. In: Conference on Robot Learning, pp 1025–1037

Hadfield-Menell D, Dragan A, Abbeel P, Russell S (2016) Cooperative inverse reinforcement learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, pp 3916–3924

Hanna JP, Stone P (2017) Grounded action transformation for robot learning in simulation. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp 3834–3840

Hausman K, Chebotar Y, Schaal S, Sukhatme G, Lim JJ (2017) Multi-modal imitation learning from unstructured demonstrations using generative adversarial nets. In: Advances in Neural Information Processing Systems, pp 1235–1245

He S, Tavakoli HR, Borji A, Mi Y, Pugeault N (2019) Understanding and visualizing deep visual saliency models. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 10206–10215

Henderson P, Chang WD, Bacon PL, Meger D, Pineau J, Precup D (2018a) Optiongan: Learning joint reward-policy options using generative adversarial inverse reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 32

Henderson P, Islam R, Bachman P, Pineau J, Precup D, Meger D (2018b) Deep reinforcement learning that matters. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 32

Hernandez-Leal P, Kartal B, Taylor ME (2019) A survey and critique of multiagent deep reinforcement learning. Autonomous Agents and Multi-Agent Systems 33(6):750–797

Ho J, Ermon S (2016) Generative adversarial imitation learning. In: Advances in Neural Information Processing Systems, pp 4565–4573

Ho MK, Littman M, MacGlashan J, Cushman F, Austerweil JL (2016) Showing versus doing: Teaching by demonstration. In: Advances in neural information processing systems, pp 3027–3035

Holden D, Saito J, Komura T (2016) A deep learning framework for character motion synthesis and editing. ACM Transactions on Graphics (TOG) 35(4):138

Huang Q, Veeraraghavan A, Sabharwal A (2017) Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. Machine Vision and Applications 28(5-6):445–461

Huang Y, Cai M, Li Z, Lu F, Sato Y (2020) Mutual context network for jointly estimating egocentric gaze and action. IEEE Transactions on Image Processing 29:7795–7806

Hussein A, Gaber MM, Elyan E, Jayne C (2017) Imitation learning: A survey of learning methods. ACM Computing Surveys (CSUR) 50(2):21

Ibarz B, Leike J, Pohlen T, Irving G, Legg S, Amodei D (2018) Reward learning from human preferences and demonstrations in atari. In: Advances in Neural Information Processing Systems, pp 8022–8034

Ijspeert AJ, Nakanishi J, Schaal S (2002) Movement imitation with nonlinear dynamical systems in humanoid robots. In: Proceedings 2002 IEEE International Conference on Robotics and Automation, IEEE, vol 2, pp 1398–1403

Ijspeert AJ, Nakanishi J, Schaal S (2011) Trajectory formation for imitation with nonlinear dynamical systems. In: Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, vol 2, pp 752–757

Isbell C, Shelton CR, Kearns M, Singh S, Stone P (2001) A social reinforcement learning agent. In: Proceedings of the fifth international conference on Autonomous agents, ACM, pp 377–384

Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis & Machine Intelligence (11):1254–1259

Jaderberg M, Czarnecki WM, Dunning I, Marris L, Lever G, Castaneda AG, Beattie C, Rabinowitz NC, Morcos AS, Ruderman A, et al. (2019) Human-level performance in 3d multiplayer games with population-based reinforcement learning. Science 364(6443):859–865

Jiang S, Pang J, Yu Y (2020) Offline imitation learning with a misspecified simulator. Advances in Neural Information Processing Systems 33

Joachims T, Granka L, Pan B, Hembrooke H, Gay G (2017) Accurately interpreting clickthrough data as implicit feedback. In: ACM SIGIR Forum, Acm New York, NY, USA, vol 51, pp 4–11

Kessler Faulkner T, Gutierrez RA, Short ES, Hoffman G, Thomaz AL (2019) Active attention-modified policy shaping: Socially interactive agents track. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '19, pp 728–736, URL http://dl.acm.org/citation.cfm?id=3306127.3331762

Kimura D, Chaudhury S, Tachibana R, Dasgupta S (2018) Internal model from observations for reward shaping. arXiv preprint arXiv:180601267

Kipf T, Li Y, Dai H, Zambaldi V, Sanchez-Gonzalez A, Grefenstette E, Kohli P, Battaglia P (2019) Compile: Compositional imitation learning and execution. In: International Conference on Machine Learning, PMLR, pp 3418–3428

Knox WB, Stone P (2009) Interactively shaping agents via human reinforcement: The tamer framework. In: Proceedings of the fifth international conference on Knowledge capture, ACM, pp 9–16

Knox WB, Stone P (2010) 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, International Foundation for Autonomous Agents and Multiagent Systems, pp 5–12

Knox WB, Stone P (2012) Reinforcement learning from simultaneous human and mdp reward. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1, International Foundation for Autonomous Agents and Multiagent Systems, pp 475–482

Kober J, Bagnell JA, Peters J (2013) Reinforcement learning in robotics: A survey. The International Journal of Robotics Research 32(11):1238–1274

Konidaris G, Kuindersma S, Grupen R, Barto A (2012) Robot learning from demonstration by constructing skill trees. The International Journal of Robotics Research 31(3):360–375

Kostrikov I, Agrawal KK, Dwibedi D, Levine S, Tompson J (2019) Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. In: International Conference on Learning Representations, URL `https://openreview.net/forum?id=Hk4fpoA5Km`

Krafka K, Khosla A, Kellnhofer P, Kannan H, Bhandarkar S, Matusik W, Torralba A (2016) Eye tracking for everyone. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2176–2184

Krishnan S, Fox R, Stoica I, Goldberg K (2017) Ddco: Discovery of deep continuous options for robot learning from demonstrations. In: Conference on Robot Learning, pp 418–437

Kroemer O, Daniel C, Neumann G, Van Hoof H, Peters J (2015) Towards learning hierarchical skills for multi-phase manipulation tasks. In: 2015 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 1503–1510

Kulkarni TD, Narasimhan K, Saeedi A, Tenenbaum J (2016) Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. In: Advances in neural information processing systems, pp 3675–3683

Laskey M, Lee J, Chuck C, Gealy D, Hsieh W, Pokorny FT, Dragan AD, Goldberg K (2016) Robot grasping in clutter: Using a hierarchy of supervisors for learning from demonstrations. In: 2016 IEEE International Conference on Automation Science and Engineering (CASE), IEEE, pp 827–834

Le H, Jiang N, Agarwal A, Dudik M, Yue Y, Daumé H (2018) Hierarchical imitation and reinforcement learning. In: International Conference on Machine

Learning, pp 2923–2932

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. nature 521(7553):436

Levine S, Finn C, Darrell T, Abbeel P (2016) End-to-end training of deep visuomotor policies. The Journal of Machine Learning Research 17(1):1334–1373

Li C, Tarlow D, Gaunt AL, Brockschmidt M, Kushman N (2016a) Neural program lattices. International Conference on Learning Representations 2018

Li G, Whiteson S, Knox WB, Hung H (2016b) Using informative behavior to increase engagement while learning from human reward. Autonomous agents and multi-agent systems 30(5):826–848

Li Y, Hou X, Koch C, Rehg JM, Yuille AL (2014) The secrets of salient object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 280–287

Li Y, Liu M, Rehg JM (2018) 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), pp 619–635

Liu C, Chen Y, Tai L, Ye H, Liu M, Shi BE (2019) A gaze model improves autonomous driving. In: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, ACM, p 33

Liu S, Cao J, Chen W, Wen L, Liu Y (2020) Hilonet: Hierarchical imitation learning from non-aligned observations. arXiv preprint arXiv:201102671

Liu Y, Gupta A, Abbeel P, Levine S (2018) Imitation from observation: Learning to imitate behaviors from raw video via context translation. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 1118–1125

Loftin R, Peng B, MacGlashan J, Littman ML, Taylor ME, Huang J, Roberts DL (2014) Learning something from nothing: Leveraging implicit human feedback strategies. In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication, IEEE, pp 607–612

Loftin R, Peng B, MacGlashan J, Littman ML, Taylor ME, Huang J, Roberts DL (2016) 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

MacGlashan J, Ho MK, Loftin R, Peng B, Wang G, Roberts DL, Taylor ME, Littman ML (2017) Interactive learning from policy-dependent human feedback. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, pp 2285–2294

Machado MC, Bellemare MG, Talvitie E, Veness J, Hausknecht M, Bowling M (2018) Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. Journal of Artificial Intelligence Research 61:523–562

Mathe S, Sminchisescu C (2014) Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. IEEE transactions on pattern analysis and machine intelligence 37(7):1408–1424

Merel J, Tassa Y, Srinivasan S, Lemmon J, Wang Z, Wayne G, Heess N (2017) Learning human behaviors from motion capture by adversarial imitation.

arXiv preprint arXiv:170702201

Misra I, Zitnick CL, Hebert M (2016) Shuffle and learn: unsupervised learning using temporal order verification. In: European conference on computer vision, Springer, pp 527–544

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, et al. (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533

Mohseni-Kabir A, Rich C, Chernova S, Sidner CL, Miller D (2015) Interactive hierarchical task learning from a single demonstration. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, ACM, pp 205–212

Nachum O, Gu SS, Lee H, Levine S (2018) Data-efficient hierarchical reinforcement learning. In: Advances in neural information processing systems, pp 3303–3313

Nair A, Chen D, Agrawal P, Isola P, Abbeel P, Malik J, Levine S (2017) Combining self-supervised learning and imitation for vision-based rope manipulation. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 2146–2153

Najar A, Sigaud O, Chetouani M (2020) Interactively shaping robot behaviour with unlabeled human instructions. Auton Agents Multi Agent Syst 34(2):35

Nakanishi J, Morimoto J, Endo G, Cheng G, Schaal S, Kawato M (2004) Learning from demonstration and adaptation of biped locomotion. Robotics and autonomous systems 47(2-3):79–91

Niekum S, Osentoski S, Konidaris G, Chitta S, Marthi B, Barto AG (2015) Learning grounded finite-state representations from unstructured demonstrations. The International Journal of Robotics Research 34(2):131–157

Osa T, Pajarinen J, Neumann G, Bagnell JA, Abbeel P, Peters J, et al. (2018) An algorithmic perspective on imitation learning. Foundations and Trends® in Robotics 7(1-2):1–179

Palan M, Landolfi NC, Shevchuk G, Sadigh D (2019) Learning reward functions by integrating human demonstrations and preferences. In: Proceedings of Robotics: Science and Systems (RSS), DOI 10.15607/rss.2019.xv.023

Palazzi A, Abati D, Solera F, Cucchiara R, et al. (2018) Predicting the driver's focus of attention: the dr (eye) ve project. IEEE transactions on pattern analysis and machine intelligence 41(7):1720–1733

Papadopoulos DP, Clarke AD, Keller F, Ferrari V (2014) Training object class detectors from eye tracking data. In: European conference on computer vision, Springer, pp 361–376

Papoutsaki A, Sangkloy P, Laskey J, Daskalova N, Huang J, Hays J (2016) Webgazer: scalable webcam eye tracking using user interactions. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pp 3839–3845

Park S, Spurr A, Hilliges O (2018) Deep pictorial gaze estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 721–738

Pathak D, Mahmoudieh P, Luo M, Agrawal P, Chen D, Shentu F, Shelhamer E, Malik J, Efros AA, Darrell T (2018) Zero-shot visual imitation.

In: International Conference on Learning Representations, URL https://openreview.net/forum?id=BkisuzWRW

Pavse BS, Torabi F, Hanna J, Warnell G, Stone P (2020) Ridm: Reinforced inverse dynamics modeling for learning from a single observed demonstration. IEEE Robotics and Automation Letters 5(4):6262–6269

Peng XB, Abbeel P, Levine S, van de Panne M (2018a) Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. ACM Transactions on Graphics (TOG) 37(4):143

Peng XB, Kanazawa A, Malik J, Abbeel P, Levine S (2018b) Sfv: Reinforcement learning of physical skills from videos. In: SIGGRAPH Asia 2018 Technical Papers, ACM, p 178

Pilarski PM, Dawson MR, Degris T, Fahimi F, Carey JP, Sutton RS (2011) Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning. In: 2011 IEEE International Conference on Rehabilitation Robotics, IEEE, pp 1–7

Pinsler R, Akrour R, Osa T, Peters J, Neumann G (2018) Sample and feedback efficient hierarchical reinforcement learning from human preferences. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 596–601

Pomerleau DA (1989) Alvinn: An autonomous land vehicle in a neural network. In: Advances in Neural Information Processing Systems, pp 305–313

Qureshi AH, Boots B, Yip MC (2019) Adversarial imitation via variational inverse reinforcement learning. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=HJlmHoR5tQ

Radosavovic I, Wang X, Pinto L, Malik J (2020) State-only imitation learning for dexterous manipulation. arXiv preprint arXiv:200404650

Ravichandar HC, Kumar A, Dani A (2018) Gaze and motion information fusion for human intention inference. International Journal of Intelligent Robotics and Applications 2(2):136–148

Reed S, De Freitas N (2015) Neural programmer-interpreters. arXiv preprint arXiv:151106279

Robertson ZW, Walter MR (2020) Concurrent training improves the performance of behavioral cloning from observation. arXiv preprint arXiv:200801205

Ross S, Bagnell D (2010) Efficient reductions for imitation learning. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp 661–668

Ross S, Gordon GJ, Bagnell D (2011) A reduction of imitation learning and structured prediction to no-regret online learning. In: International Conference on Artificial Intelligence and Statistics, pp 627–635

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

Saran A, Zhang R, Short ES, Niekum S (2020) Efficiently guiding imitation learning algorithms with human gaze. arXiv preprint arXiv:200212500

Sasaki F, Yohira T, Kawaguchi A (2019) Sample efficient imitation learning for continuous control. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=BkN5UoAqF7

Saunders W, Sastry G, Stuhlmueller A, Evans O (2018) Trial without error: Towards safe reinforcement learning via human intervention. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, pp 2067–2069

Schaal S (1999) Is imitation learning the route to humanoid robots? Trends in cognitive sciences 3(6):233–242

Schmeckpeper K, Rybkin O, Daniilidis K, Levine S, Finn C (2020) Reinforcement learning with videos: Combining offline observations with interaction. arXiv preprint arXiv:201106507

Schulman J, Levine S, Abbeel P, Jordan M, Moritz P (2015) Trust region policy optimization. In: International Conference on Machine Learning, pp 1889–1897

Sermanet P, Lynch C, Chebotar Y, Hsu J, Jang E, Schaal S, Levine S, Brain G (2018) Time-contrastive networks: Self-supervised learning from video. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 1134–1141

Setapen A, Quinlan M, Stone P (2010) Marionet: Motion acquisition for robots through iterative online evaluative training. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1, International Foundation for Autonomous Agents and Multiagent Systems, pp 1435–1436

Sharma M, Sharma A, Rhinehart N, Kitani KM (2018) Directed-info gail: Learning hierarchical policies from unsegmented demonstrations using directed information. In: International Conference on Learning Representations

Shen Y, Ni B, Li Z, Zhuang N (2018) Egocentric activity prediction via event modulated attention. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 197–212

Shrivastava A, Pfister T, Tuzel O, Susskind J, Wang W, Webb R (2017) Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2107–2116

Silver D, Bagnell JA, Stentz A (2010) Learning from demonstration for autonomous navigation in complex unstructured terrain. The International Journal of Robotics Research 29(12):1565–1592

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, et al. (2016) Mastering the game of go with deep neural networks and tree search. Nature 529(7587):484–489

Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, et al. (2017) Mastering the game of go without human knowledge. Nature 550(7676):354

Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, Lanctot M, Sifre L, Kumaran D, Graepel T, et al. (2018) A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. Science 362(6419):1140–1144

Skinner BF (1938) The behavior of organisms: An experimental analysis. BF Skinner Foundation

Stadie BC, Abbeel P, Sutskever I (2017) Third-person imitation learning. International Conference on Learning Representations

Sudhakaran S, Escalera S, Lanz O (2019) Lsta: Long short-term attention for egocentric action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 9954–9963

Sun W, Vemula A, Boots B, Bagnell D (2019) Provably efficient imitation learning from observation alone. In: International Conference on Machine Learning, pp 6036–6045

Sutton RS, Barto AG (1998) Reinforcement learning: An introduction, vol 1. MIT press Cambridge

Sutton RS, Barto AG (2018) Reinforcement learning: An introduction. MIT press

Sutton RS, Precup D, Singh S (1999) Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. Artificial intelligence 112(1-2):181–211

Sutton RS, McAllester DA, Singh SP, Mansour Y (2000) Policy gradient methods for reinforcement learning with function approximation. In: Advances in neural information processing systems, pp 1057–1063

Tavakoli HR, Rahtu E, Kannala J, Borji A (2019) Digging deeper into egocentric gaze prediction. In: 2019 IEEE Winter Conference on Applications of Computer Vision, IEEE, pp 273–282

Tenorio-Gonzalez AC, Morales EF, Villaseñor-Pineda L (2010) Dynamic reward shaping: training a robot by voice. In: Ibero-American conference on artificial intelligence, Springer, pp 483–492

Thomaz AL, Breazeal C (2008) Teachable robots: Understanding human teaching behavior to build more effective robot learners. Artificial Intelligence 172(6-7):716–737

Todorov E, Erez T, Tassa Y (2012) Mujoco: A physics engine for model-based control. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, pp 5026–5033

Torabi F, Warnell G, Stone P (2018a) Behavioral cloning from observation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, AAAI Press, pp 4950–4957

Torabi F, Warnell G, Stone P (2018b) Generative adversarial imitation from observation. arXiv preprint arXiv:180706158

Torabi F, Geiger S, Warnell G, Stone P (2019a) Sample-efficient adversarial imitation learning from observation. arXiv preprint arXiv:190607374

Torabi F, Warnell G, Stone P (2019b) Adversarial imitation learning from state-only demonstrations. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, pp 2229–2231

Torabi F, Warnell G, Stone P (2019c) Imitation learning from video by leveraging proprioception. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, pp 3585–3591

Torabi F, Warnell G, Stone P (2019d) Recent advances in imitation learning from observation. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, pp 6325–6331

Vezhnevets AS, Osindero S, Schaul T, Heess N, Jaderberg M, Silver D, Kavukcuoglu K (2017) Feudal networks for hierarchical reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp 3540–3549

Vinyals O, Babuschkin I, Czarnecki WM, Mathieu M, Dudzik A, Chung J, Choi DH, Powell R, Ewalds T, Georgiev P, et al. (2019) Grandmaster level in starcraft ii using multi-agent reinforcement learning. Nature 575(7782):350–354

Wang K, Lin L, Jiang C, Qian C, Wei P (2019) 3d human pose machines with self-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 42(5):1069–1082

Wang M, Deng W (2018) Deep visual domain adaptation: A survey. Neurocomputing 312:135–153

Wang W, Shen J, Guo F, Cheng MM, Borji A (2018) Revisiting video saliency: A large-scale benchmark and a new model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4894–4903

Wang Z, Schaul T, Hessel M, Hasselt H, Lanctot M, Freitas N (2016) Dueling network architectures for deep reinforcement learning. In: International Conference on Machine Learning, pp 1995–2003

Warnell G, Waytowich N, Lawhern V, Stone P (2018) Deep tamer: Interactive agent shaping in high-dimensional state spaces. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 32

Watkins CJ, Dayan P (1992) Q-learning. Machine learning 8(3-4):279–292

Waytowich NR, Goecks VG, Lawhern VJ (2018) Cycle-of-learning for autonomous systems from human interaction. arXiv preprint arXiv:180809572

Wilson A, Fern A, Tadepalli P (2012) A bayesian approach for policy learning from trajectory preference queries. In: Advances in neural information processing systems, pp 1133–1141

Wirth C, Fürnkranz J, Neumann G (2016) Model-free preference-based reinforcement learning. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp 2222–2228

Wirth C, Akrour R, Neumann G, Fürnkranz J (2017) A survey of preference-based reinforcement learning methods. The Journal of Machine Learning Research 18(1):4945–4990

Wood E, Baltrusaitis T, Zhang X, Sugano Y, Robinson P, Bulling A (2015) Rendering of eyes for eye-shape registration and gaze estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3756–3764

Woodward M, Finn C, Hausman K (2020) Learning to interactively learn and assist. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp

2535–2543

Wu A, Piergiovanni A, Ryoo MS (2020) Model-based behavioral cloning with future image similarity learning. In: Conference on Robot Learning, PMLR, pp 1062–1077

Xia Y, Zhang D, Kim J, Nakayama K, Zipser K, Whitney D (2018) Predicting driver attention in critical situations. In: Asian conference on computer vision, Springer, pp 658–674

Xia Y, Kim J, Canny J, Zipser K, Canas-Bajo T, Whitney D (2020) Periphery-fovea multi-resolution driving model guided by human attention. In: The IEEE Winter Conference on Applications of Computer Vision, pp 1767–1775

Xu D, Nair S, Zhu Y, Gao J, Garg A, Fei-Fei L, Savarese S (2018a) Neural task programming: Learning to generalize across hierarchical tasks. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp 1–8

Xu D, Agarwal M, Fekri F, Sivakumar R (2020) Playing games with implicit human feedback. In: Workshop on Reinforcement Learning in Games, AAAI

Xu J, Jiang M, Wang S, Kankanhalli MS, Zhao Q (2014) Predicting human gaze beyond pixels. Journal of vision 14(1):28–28

Xu J, Liu Q, Guo H, Kageza A, AlQarni S, Wu S (2018b) Shared multi-task imitation learning for indoor self-navigation. In: 2018 IEEE Global Communications Conference (GLOBECOM), IEEE, pp 1–7

Xu Y, Dong Y, Wu J, Sun Z, Shi Z, Yu J, Gao S (2018c) Gaze prediction in dynamic 360 immersive videos. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5333–5342

Yang C, Ma X, Huang W, Sun F, Liu H, Huang J, Gan C (2019) Imitation learning from observations by minimizing inverse dynamics disagreement. In: Advances in Neural Information Processing Systems, pp 239–249

Yang JM, Kim JH (1999) Sliding mode control for trajectory tracking of nonholonomic wheeled mobile robots. IEEE Transactions on robotics and automation 15(3):578–587

Yu F, Xian W, Chen Y, Liu F, Liao M, Madhavan V, Darrell T (2018) Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv preprint arXiv:180504687

Zhang R, Liu Z, Zhang L, Whritner JA, Muller KS, Hayhoe MM, Ballard DH (2018a) Agil: Learning attention from human for visuomotor tasks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 663–679

Zhang R, Torabi F, Guan L, Ballard DH, Stone P (2019) Leveraging human guidance for deep reinforcement learning tasks. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, pp 6339–6346

Zhang R, Saran A, Liu B, Zhu Y, Guo S, Niekum S, Ballard D, Hayhoe M (2020a) Human gaze assisted artificial intelligence: A review. In: IJCAI: proceedings of the conference, NIH Public Access, vol 2020, p 4951

Zhang R, Walshe C, Liu Z, Guan L, Muller K, Whritner J, Zhang L, Hayhoe M, Ballard D (2020b) Atari-head: Atari human eye-tracking and demonstration dataset. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 34, pp 6811–6820

Zhang X, Sugano Y, Fritz M, Bulling A (2015) Appearance-based gaze estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4511–4520

Zhang X, Sugano Y, Fritz M, Bulling A (2017) Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. IEEE transactions on pattern analysis and machine intelligence 41(1):162–175

Zhang Z, Xu Y, Yu J, Gao S (2018b) Saliency detection in 360 videos. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 488–503

Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232

Ziebart BD, Maas AL, Bagnell JA, Dey AK (2008) Maximum entropy inverse reinforcement learning. In: AAAI, Chicago, IL, USA, vol 8, pp 1433–1438

Ziebart BD, Bagnell JA, Dey AK (2010) Modeling interaction via the principle of maximum causal entropy. In: ICML

Zintgraf LM, Roijers DM, Linders S, Jonker CM, Nowé A (2018) Ordered preference elicitation strategies for supporting multi-objective decision making. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, pp 1477–1485

Zolna K, Rostamzadeh N, Bengio Y, Ahn S, Pinheiro PO (2018) Reinforced imitation learning from observations. NeurIPS 2018 Workshop

Zuo Z, Yang L, Peng Y, Chao F, Qu Y (2018) Gaze-informed egocentric action recognition for memory aid systems. IEEE Access 6:12894–12904