How to Train your Decision-Making AIs

Ruohan Zhang, Dhruva Bansal
Department of Computer Science,Stanford University
zharu@stanford.edu

The combination of deep learning and decision learning has led to several impressive stories in decision-making AI research, including AIs that can play a variety of games (Atari video games [11], board games [13], complex real-time strategy game Starcraft II [17]), control robots (in simulation and in the real world), and even fly a weather balloon [1]. These are examples of sequential decision tasks, in which the AI agent needs to make a sequence of decisions to achieve its goal.

Today, the two main approaches for training such agents are reinforcement learning (RL) and imitation learning (IL). In reinforcement learning, humans provide rewards for completing discrete tasks, with the rewards typically being delayed and sparse. For example, 100 points are given for solving the first room of Montezumas revenge (Fig. 1). In the imitation learning setting, humans can transfer knowledge and skills through step-by-step action demonstrations (Fig. 2), and the agent then learns to mimic human actions.

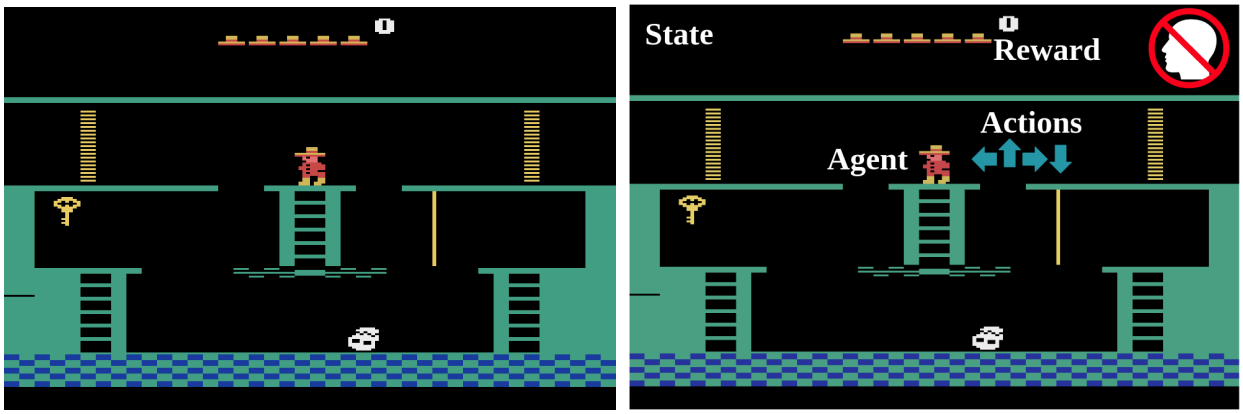

Figure 1: Left: Montezumas Revenge, one of the most difficult Atari 2600 games for both humans and AIs. The player controls the little person (agent). Right: In a typical reinforcement learning setting, the agent needs to learn to play the game without any human guidance purely based on the score provided by the environment.

But, success stories about RL and IL are often based on the fact that we can train AIs in simulated environments with a large amount of training data. What if we dont have a simulator for the learning agent to fool around in? What if these agents need to learn quickly and safely? What if the agents need to adapt to individual human needs? These concerns lead to the key questions we ask, which are: How do humans transfer their knowledge and skills to artificial decision-making agents more efficiently? What kind of knowledge and skills should humans provide and in what format?

Human guidance

For humans, there are many diverse and creative ways of teaching and learning from other people (or animals). In daily social learning settings, we use many learning signals and learning methods, which well refer to as human guidance. Recently, a lot of research has explored alternative ways in which humans may guide learning agents.

My colleagues and I reviewed five types of human guidance to train AIs: evaluation, preference, goals, attention, and demonstrations without action labels [20, 21] (Fig 3). They don't replace imitation or reinforcement learning

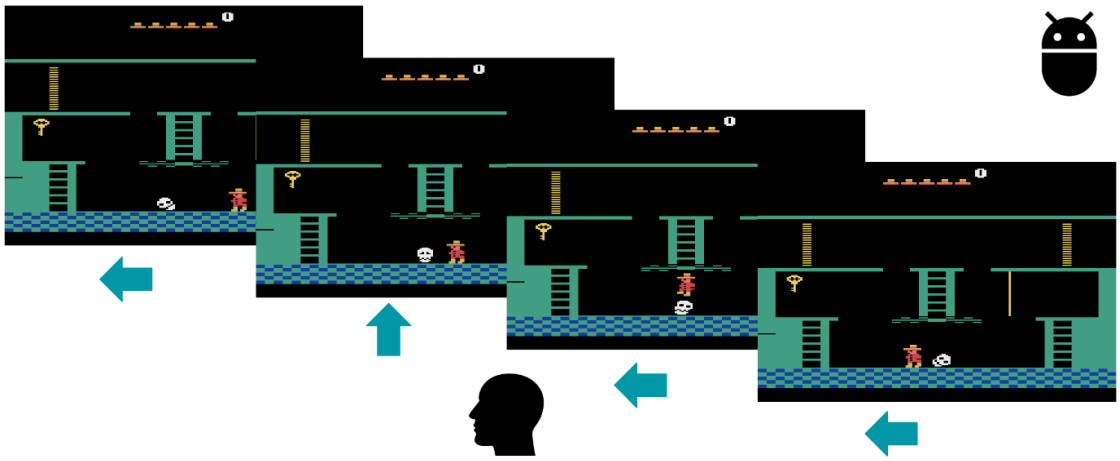

Figure 2: In standard imitation learning, a human trainer demonstrates a sequence of actions, and the agent learns to imitate the trainer's actions.

Figure 3: The author is training his dog Twinkle. How is training dogs different from training AIs using reinforcement learning or imitation learning? It is infeasible to demonstrate the correct actions in most cases. In addition to providing a large piece of treat at the end of the training like in RL, I also provide **evaluative feedback** (good girl! bad dog!) immediately after observing the desired action, using clicker training. I clearly indicate my **preference**: pee on the pee pad, not on the floor. I indicate behavior **goals** to her by pointing to the toy I want her to retrieve. I train her to **pay attention** to me and the objects I am pointing to. She watches me do things like opening the treat jar – while she does not have hands but she can **imitate from observation** to do that with her teeth.

methods, but rather work with them to widen the communication pipeline between humans and learning agents.

Training AIs via evaluative feedback

The first type of guidance is human evaluative feedback. In this setting, a human trainer watches an agent trying to learn and provides positive feedback for desirable actions and negative feedback for undesirable actions (Fig 4).

The benefit of human evaluative feedback is that it is instantaneous and frequent, while the true reward is delayed and sparse. For example, in games like chess or Go, the reward (win or lose) is not revealed until the end of the game. In addition, it does not require the trainers to be experts in performing the task – they just need to be good at judging the performance. This is akin to a sports coach who provides guidance in the form of feedback to professional athletes based on their performance. Although 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 useful to the athlete. Finally, this method is particularly useful for tasks that require safe learning, since we block catastrophic actions when human trainers provide negative feedback.

While evaluative feedback has traditionally been communicated through button presses by humans, recent work has also explored inferring feedback from signals humans naturally emit such as gestures, facial expressions, and even

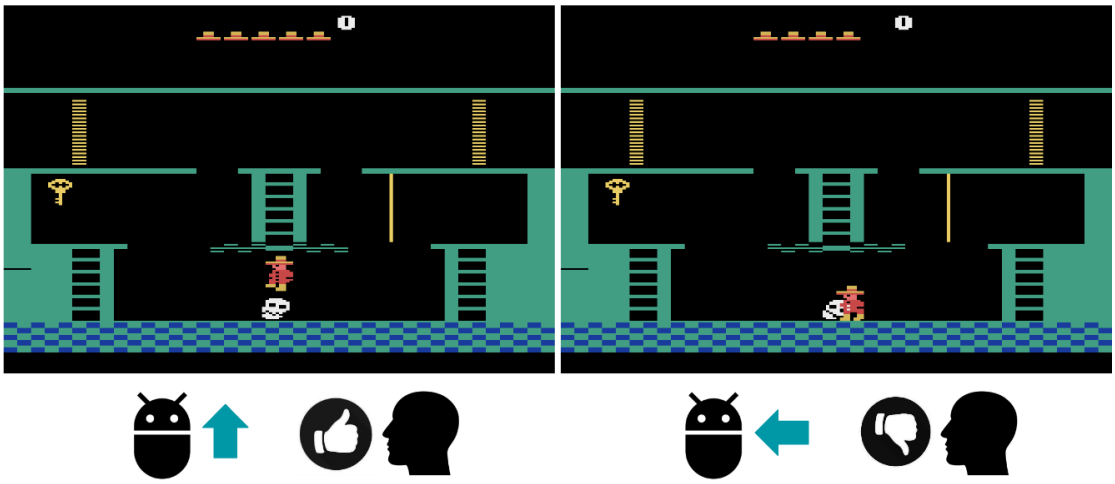

Figure 4: While learning from evaluative feedback, a human trainer watches the agent's learning process, and provides positive feedback for a desirable action (jumping over the skull), and negative feedback for an undesirable action (running into the skull).

brain waves.

In the context of reinforcement learning, researchers have different interpretations of the nature of human feedback. While it is intuitive to think of feedback as additional reward signals, more likely interpretations include the feedback being a form of policy gradient, value function, or advantage function. Different learning algorithms have been developed depending on the interpretations, such as Advise [5], TAMER [8], and COACH [10].

Training AIs using preference

The second type of guidance is human preference [4, 2]. The learning agent presents two of its learned behavior trajectories to the human trainer, and the human tells the agent which trajectory is preferable (Fig. 5). The motivation is that sometimes the evaluation can only be provided at the end of a trajectory, instead of at every time step. For example, in Fig. 5, moving right is a better choice than moving down from the starting position, but this is only clear when we see trajectory 1 is shorter and safer than trajectory 2. Additionally, ranking behavior trajectories is easier than rating them. Typically, preference learning is formalized as an inverse reinforcement learning problem in which the goal is to learn the unobserved reward function from human preference. A particularly interesting research question is to determine which trajectories should be selected to query humans for their preference, such that the agent can gain useful information from humans. This is known as the preference elicitation problem.

Training AIs by providing high-level goals

Many decision-making tasks are hierarchically structured, meaning that they can be solved using a divide-and-conquer strategy. In hierarchical imitation, the idea is to have the human trainer specify high-level goals. For example, in Fig. 6, the first goal is to reach the bottom of this ladder. The agent will then try to accomplish each goal by performing a sequence of low-level actions.

Behavioral psychology studies with non-human primates have shown that fine-grained, low-level actions are mostly learned without imitation. In contrast, coarse, high-level program" learning is pervasive in imitation learning [3]. Program-level imitation (e.g., imitating human high-level goals in Fig. 6) is defined as imitating the high-level structural organization of a complex process by observing the behavior of another individual, while furnishing the exact details of actions by individual learning [3] (perhaps through reinforcement learning).

When training AIs we can be more flexible and not just replicate the learning seen in non-human primates [9]. We can let human trainers provide both high-level and low-level demonstrations, and do imitation learning at both levels. Alternatively, like conventional imitation learning, humans can provide only low-level action demonstrations, and agents must then extract task hierarchy on their own. A promising combination is to learn goal selection using

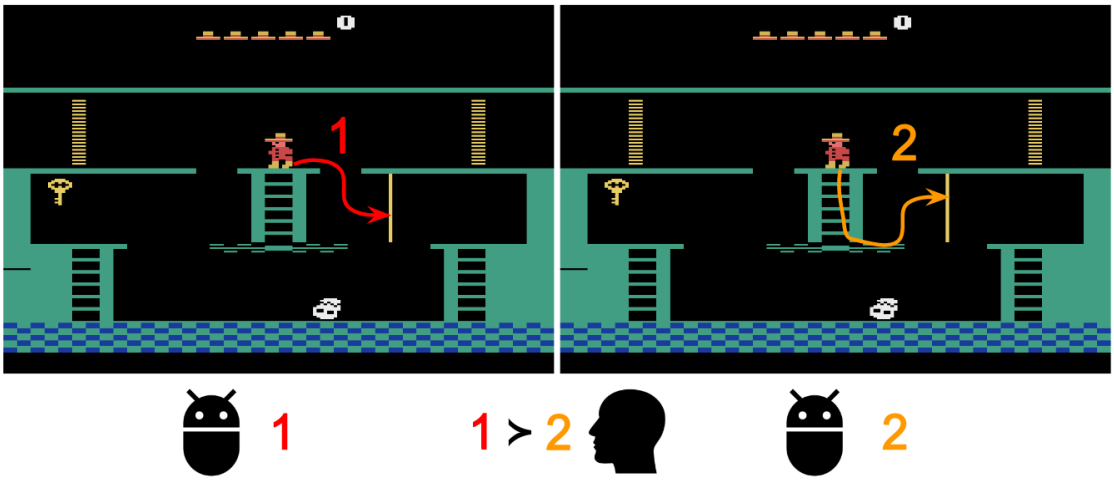

Figure 5: In learning from human preference, the learning agent presents two learned behavior trajectories to the human trainer, and the human tells the agent which trajectory is preferable. Here the human trainer prefers trajectory 1.

imitation learning at a high level and let a reinforcement learning agent learn to perform low-level actions to achieve high-level goals. The motivation for this combination is that humans are good at specifying high-level abstract goals while the agents are good at executing low-level fine-grained controls.

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. On the contrary, in tasks like Tetris, providing low-level demonstration is easier, since high-level goals are not easy to specify. The second concern is safety. Only providing high-level goals requires the agents to learn low-level policies by themselves through trial-and-error, which is suitable for simulated agents but not for physical robots. Therefore in robotic tasks, low-level action demonstrations are often required.

Training AIs to attend to the right things

We are surrounded by a complex world full of information. Humans have developed strategies for selecting information, known as selective attention. Attention was not a main research focus in the pre-deep RL era since in general we do not include irrelevant features in the hand-crafted feature set. As deep RL agents migrate from a simple digital world to the complex real world, the same challenge awaits AI agents: How do they select important information from a world full of information to ensure that resources are devoted to the key components?

Human selective attention is developed through evolution and is refined in a lifelong learning process. Given the amount of training data required during this process, it may be easier for AI agents to learn attention directly from humans [7]. Learning to attend can help select important state features in high-dimensional state space, and help the agent infer the target or goal of an observed action by the human teachers.

The first step is to use eye-tracking datasets to train an agent to imitate human gaze behaviors, i.e., learning to attend to certain features of a given image. The problem is formalized as a visual saliency prediction problem in computer vision research and is well-studied. Next, knowing where humans would look provides useful information on what action they will take. Therefore it is intuitive to leverage learned attention models to guide the learning process. The challenge here is how to use human attention in decision learning. 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 highlight the important visual features. Experimental results so far have shown that including gaze information leads to higher accuracy in recognizing or predicting human actions, manipulating objects, driving, cooking, and playing video games [19, 22, 18].

Attention data can often be collected in parallel with demonstration data, and has the potential to be combined with other types of learning as well (e.g., combining with evaluative feedback or preference [12, 6]). In the future, when decision-making AIs are ready to work alongside and collaborate with humans, understanding and utilizing human

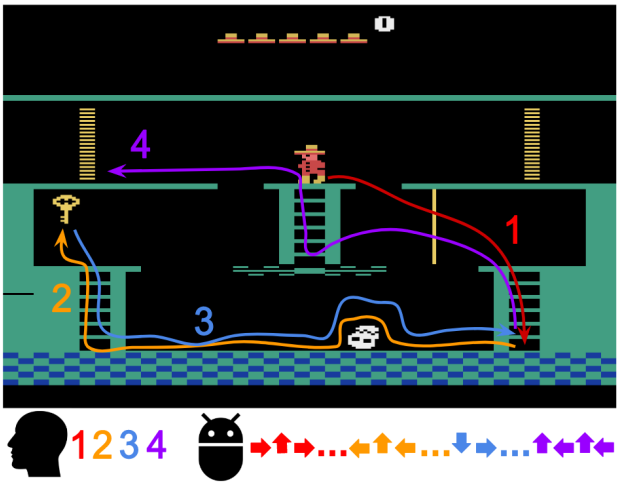

Figure 6: In hierarchical imitation, the basic idea is to have the human trainer specify high-level goals. For example, the red goal is to reach the bottom of the ladder. An agent will learn to accomplish each high-level goal by performing a sequence of low-level actions potentially learned through reinforcement learning itself.

attention would be even more important.

Training AIs by without action labels

The last type of learning paradigm, imitation from observation, is different from the previous four. The setting is very much like standard imitation learning, except the agent does not have access to labels for the actions demonstrated by the human trainer. While this makes imitation from observation a very challenging problem, it enables us to utilize a large amount of human demonstration data that do not have action labels, such as Youtube videos.

In order to make imitation from observation possible, the first step is action label inference. A straightforward solution is to interact with the environment, collect state action data, and then learn an inverse dynamics model. Applying this learned model on two consecutive demonstrated states (for example, the first image and the second image) would output the missing action (go left) that had resulted in that state transition. After retrieving the actions, the learning problem can be treated as a conventional imitation learning problem [15, 16].

However, in practice, there are other challenges as well. For example, embodiment mismatch may arise when the teacher has different physical characteristics than that of the learning agent. Furthermore, viewpoint mismatch arises when there may be a difference in the point of view present in the demonstration video and that with which the agent sees itself. As an example, both of these challenges are present when a robot watches human videos to learn how to cook.

Why do we need human guidance?

The most efficient method to teach a learning agent is often agent-dependent and task-dependent, since each of these five learning paradigms offers their unique advantages compared to standard imitation and reinforcement learning methods. For example, teaching a robot with many joints to sweep the floor through evaluative feedback might require a lot less human effort than through demonstration. Including attention information might be most useful in a visually rich environment.

In addition to their unique advantages, there are at least two good reasons that they are important. The first one is to build human-centered AIs. Decision-leaning agents can be trained in the factory, but they need to adapt to individual human needs when they are brought home by the customers. Evaluative feedback, preference, goals, attention, and observation are natural signals for ordinary customers to communicate with the agents about human needs.

As an example, lets look at household robots that are being developed by the Stanford Vision and Learning Lab [14]. Although these robots may acquire general household skills like food preparation and cleaning, it is up to individual humans to decide how much dressing they want in their salad, and how should the bedroom be rearranged. These

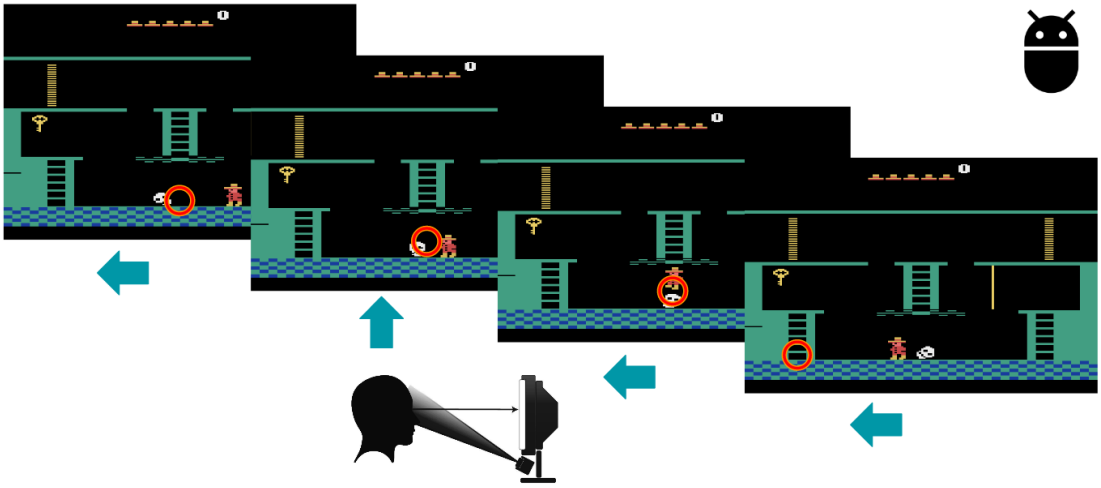

Figure 7: In learning attention from humans, the agent has access to human attention information in addition to the action demonstrations. The eye movement data indicated by the red circles here can be recorded using an eye tracker. This data reveals the current behavioral goal (such as the object of interest, e.g., the skull and the ladder) when taking an action.

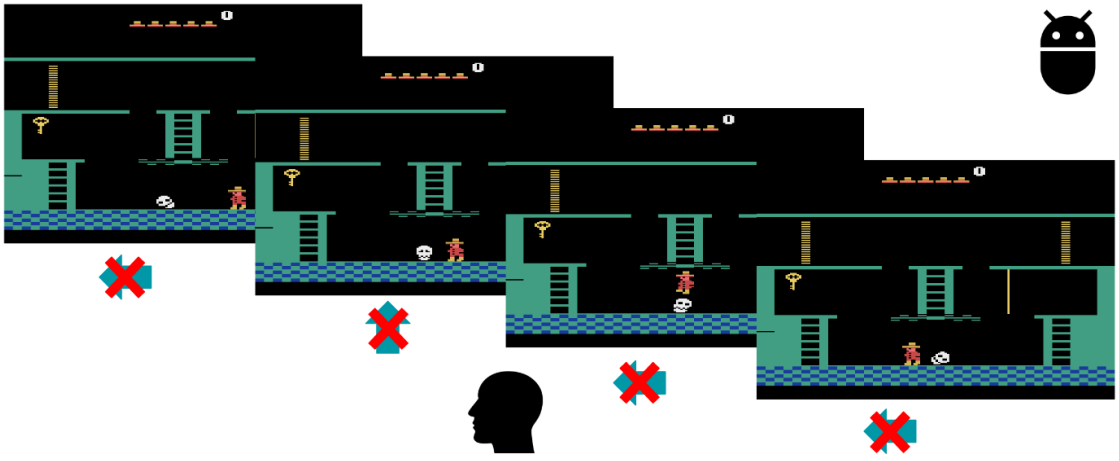

Figure 8: Fig 8. In imitation from observation, the setting is very much like standard imitation learning (Fig 2), except that the agent does not have access to the action labels demonstrated by the human trainer.

needs are difficult to demonstrate or provide reward functions for. For other human-centered AI applications, such as autonomous driving, cooking, and customer service, using human guidance to widen the communication channel between humans and AIs is also likely to be important.

The second reason that we need these guidance signals is that they allow us to build large-scale human-AI training systems, such as crowd-sourcing platforms. Providing demonstrations to AIs often requires humans to be experts in the demonstrated tasks, and the necessary hardware, which could be expensive in some cases (e.g., teleoperation devices like VR). In contrast, providing guidance may require less expertise with minimum hardware requirements. For example, the trainer can monitor the learning process of an agent from home, and use a mouse and keyboard to provide feedback, indicate preferences, select goals, and direct attention.

Future challenge: a lifelong learning paradigm

The learning frameworks discussed here 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 at their early stage of learning. Then the children gradually learn to develop joint attention with adults through gaze following. Later children begin to adjust their behaviors based on the evaluative feedback and preference received when interacting with other people. Once they develop the ability to reason abstractly about task structure, hierarchical imitation becomes feasible. At the same time, learning through trial and error (in other words, reinforcement) is always one of the most common types of learning. Our ability to learn from all types of resources continues 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. For humans, our ability to learn from all types of resources continues to develop through a lifetime. In the long term, perhaps human guidance can play a role in allowing AI to do so as well.

References

[1] Marc G Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C Machado, Subhodeep Moitra, Sameera S Ponda, and Ziyu Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

[2] Erdem Biyik and Dorsa Sadigh. Batch active preference-based learning of reward functions. In *Conference on robot learning*, pages 519–528. PMLR, 2018.

[3] Richard W Byrne and Anne E Russon. Learning by imitation: A hierarchical approach. *Behavioral and brain sciences*, 21(5):667–684, 1998.

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

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

[6] Lin Guan, Mudit Verma, Sihang Guo, Ruohan Zhang, and Subbarao Kambhampati. Widening the pipeline in human-guided reinforcement learning with explanation and context-aware data augmentation. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Sihang Guo, Ruohan Zhang, Bo Liu, Yifeng Zhu, Dana Ballard, Mary Hayhoe, and Peter Stone. Machine versus human attention in deep reinforcement learning tasks. *Advances in Neural Information Processing Systems*, 34, 2021.

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

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

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

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

[12] Akanksha Saran, Ruohan Zhang, Elaine S Short, and Scott Niekum. Efficiently guiding imitation learning agents with human gaze. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1109–1117, 2021.

[13] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[14] Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on Robot Learning*, pages 477–490. PMLR, 2022.

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

[16] Faraz Torabi, Garrett Warnell, and Peter Stone. Recent advances in imitation learning from observation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6325–6331. AAAI Press, 2019.

[17] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[18] R Zhang, A Saran, B Liu, Y Zhu, S Guo, S Niekum, D Ballard, and M Hayhoe. Human gaze assisted artificial intelligence: A review. In *International Joint Conference on Artificial Intelligence*, 2020.

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

[20] Ruohan Zhang, Faraz Torabi, Lin Guan, Dana H Ballard, and Peter Stone. Leveraging human guidance for deep reinforcement learning tasks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6339–6346. AAAI Press, 2019.

[21] Ruohan Zhang, Faraz Torabi, Garrett Warnell, and Peter Stone. Recent advances in leveraging human guidance for sequential decision-making tasks. *Autonomous Agents and Multi-Agent Systems*, 35(2):1–39, 2021.

[22] Ruohan Zhang, Calen Walshe, Zhuode Liu, Lin Guan, Karl S Muller, Jake A Whritner, Luxin Zhang, Mary M Hayhoe, and Dana H Ballard. Atari-head: Atari human eye-tracking and demonstration dataset. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*. AAAI Press, 2020.