Interactive Teaching Strategies for Agent Training

Ofra Amir
Harvard University
oamir@seas.harvard.edu

Ece Kamar
Microsoft Research
eckamar@microsoft.com

Andrey Kolobov
Microsoft Research
akolobov@microsoft.com

Barbara J. Grosz
Harvard University
grosz@eecs.harvard.edu

Abstract

Agents learning how to act in new environments can benefit from input from more experienced agents or humans. This paper studies interactive teaching strategies for identifying when a student can benefit from teacher-advice in a reinforcement learning framework. In student-teacher learning, a teacher agent can advise the student on which action to take. Prior work has considered heuristics for the teacher to choose advising opportunities. While these approaches effectively accelerate agent training, they assume that the teacher constantly monitors the student. This assumption may not be satisfied with human teachers, as people incur cognitive costs of monitoring and might not always pay attention. We propose strategies for a teacher and a student to jointly identify advising opportunities so that the teacher is not required to constantly monitor the student. Experimental results show that these approaches reduce the amount of attention required from the teacher compared to teacher-initiated strategies, while maintaining similar learning gains. The empirical evaluation also investigates the effect of the information communicated to the teacher and the quality of the student's initial policy on teaching outcomes.

1 Introduction

When agents are face the task of learning how to act in a new environment, they can benefit from the input of more experienced agents and humans. Multiple lines of work have focused on incorporating different types of input to agent learning. *Learning from demonstration* approaches aim at inferring a policy from expert demonstrations [Argall *et al.*, 2009; Abbeel and Ng, 2004; Chernova and Veloso, 2007]. In *reward shaping*, an agent learns from positive or negative signals provided by an expert [Knox and Stone, 2010; Loftin *et al.*, 2015]. In *action critiquing*, an agent practices by interacting with the environment and an expert evaluates its actions at the end of each session [Judah *et al.*, 2010].

We focus on a paradigm for giving feedback to an agent in real time, called *student-teacher training*. In this framework, an experienced "teacher" agent helps accelerate the "student" agent's learning by providing advice on which action to take next [Clouse, 1996; Torrey and Taylor, 2013]. The student updates its policy based on reward signals from the environment as in typical reinforcement learning, but its exploration is guided by the teacher's advice.

Prior work has considered two modes of advice-giving in this framework: student-initiated [Clouse, 1996] and teacher-initiated [Torrey and Taylor, 2013]. Torrey and Taylor [2013] considered a setting with a limited advice budget and developed heuristics that guide the teacher's choice of advising opportunities. They demonstrated significant learning gains when using these heuristics in empirical studies. While the amount of advice that the teacher can provide was limited, their formulation assumed that the student's current state is always communicated to the teacher, and that the teacher continuously monitors the student's decisions until the advice budget runs out. These assumptions have significant drawbacks. For human teachers, constantly paying attention diminishes the value of automation, imposes cognitive costs [Miller *et al.*, 2015] and can be simply unrealistic. Even if the teacher is a computer agent, transmitting the student's every state to the teacher can have a prohibitive communication cost.

In this paper, we propose interactive student-teacher training, in which the student *and* the teacher jointly decide when advice should be given. In these *jointly-initiated* teaching strategies, the student determines whether to ask for the teacher's attention, and the teacher, if asked to pay attention to the student's state, decides whether to use this opportunity to give advice, given a limited advice budget. We begin by comparing the teacher-initiated and student-initiated approaches experimentally, showing that heuristics for teacher-initiated training are more effective at improving the student agent's policy than student-initiated ones, but they require more teacher attention. Then we demonstrate that the jointly-initiated teaching strategies can reduce the amount of attention required of the teacher compared to teacher-initiated strategies, while maintaining similar learning gains. Thus, our work integrates the teacher-initiated and student-initiated approaches, alleviating their disadvantages.

Collaborative approaches for assisting agents are particularly important for semi-autonomous agents [Zilberstein, 2015] (e.g., self-driving cars), as such agents will have long-term interactions with people and will have opportunities to

continuously improve their policies based on these interactions. Therefore, in addition to comparing the effectiveness of different interactive training strategies, we investigate the effect on learning performance of factors that may vary across agent-human settings. In particular, our empirical evaluations analyze the effect of the information communicated to the teacher and the quality of the initial policy of the student on teaching outcomes.

The contributions of the paper are threefold. First, it extends prior work on the teacher-student reinforcement learning framework by considering both communication and attention requirements, motivated by settings in which a person will assist agents' learning. Second, it proposes jointly-initiated advising approaches to reduce attention demands for the teacher. Third, it empirically evaluates the proposed approaches in the Pac-Man domain, and explores the effect of various aspects of teaching strategies on student learning gains and teacher attention requirements.

2 Related Work

Our paper builds and expands on prior work studying the student-teacher reinforcement learning framework [Clouse, 1996; Torrey and Taylor, 2013], discussed in Section 3.

Chernova & Veloso [2007] proposed a confidence-based approach in which a learning agent asks for demonstrations when it is uncertain of its actions. In their approach, in contrast to the teacher-student framework, the agent only learns from the expert demonstration without receiving a signal from the environment. Judah et al. [2014] proposed a framework for active imitation learning, in which an agent can query an expert for its policy for a given state. They also assumed that the learning agent does not receive a reward signal from the environment. Furthermore, they assumed that the agent can simulate trajectories and does not query for demonstration during execution. Rosenstein & Barto [2004] proposed a supervised actor-critic RL framework in which a supervisor's action is integrated with a learning agent's action. In contrast to our work, they assume a continuous action space. Griffith et al. [2013] proposed a Bayesian approach for integrating human feedback on the correctness of actions to shape an agent's policy. Rosman and Ramamoorthy [2014] developed methods for deciding when to advise an agent, but assume teachers have access to a knowledge base of common agent trajectories.

Our motivation for aiming to reduce the attention required from the teacher is rooted in prior research on human-agent interaction and human attention. These works have developed methods for detecting human attention and for incorporating it into agent decision making, taking into consideration the limited attention resources available to people as well as the costs of interruptions [Horvitz *et al.*, 1999; 2003]. *Adjustable autonomy* approaches take into account the user's focus of attention when deciding whether to act autonomously or transfer autonomy to the user [Tambe *et al.*, 2002; Goodrich *et al.*, 2001]. Models of human attention are also key in developing approaches for supporting humans supervising autonomous systems [Cummings and Mitchell, 2008; Fong *et al.*, 2002].

3 Student-Teacher Reinforcement Learning

The student-teacher framework [Clouse, 1996] includes two agents: a student and a teacher. We assume that the teacher has already established a fixed policy for acting in the environment, denoted $\pi_{teacher}$, whereas the student uses a reinforcement learning algorithm to learn its policy, denoted $\pi_{student}$. At any state s, the teacher can give advice to the student by sharing $\pi_{teacher}(s)$. This formulation requires that the teacher and the student share the same action space but does not assume they share the same state representation. When the student receives advice from the teacher, it takes the suggested action. That action is then treated as any other action chosen by the student during the learning period, and Q-values are updated using the same learning algorithm.

Similarly to Torrey & Taylor [2013], we specify a limited advice budget for the teacher, but in contrast, we do not assume constant monitoring of the student by the teacher. Rather than specifying an attention budget, we consider the attention required of the teacher as an additional metric by which we evaluate the different teaching strategies. We consider two different metrics for attention: (1) the number of states in which the teacher had to assess the student's state (to decide whether to give advice), and (2) the overall duration of the teaching period (i.e., the last time step in which the teacher had to assess the student's state).

Instead of fixing an attention budget, we choose to analyze the amount and duration required by training strategies, because considerations about attention vary among different settings. For example, if a person is helping an autonomous car to improve its policy, the overall duration of the teaching period does not matter because the person is always present when the car drives. However, the person might not pay attention to the road at all times and therefore there is a cost associated with monitoring the car's actions to decide whether to intervene. Moreover, if teaching in this setting requires the human to take control over the car, then providing advice incurs an additional cost beyond monitoring the agent's behavior (i.e., deciding whether to intervene requires less effort than actually intervening). Thus, in this setting, we would like to minimize the number of states in which we require the teacher's attention as well as the number of times the teacher is required to give advice. In contrast, if an expert is brought to a lab to help train a robot, teaching is done during a dedicated time period in which the teacher watches a robot student. Here, minimizing the overall duration of the teaching period will be more important than minimizing the number of states in which attention is required.

3.1 Teacher-Initiated Advising

Torrey & Taylor [2013] proposed several heuristics for the teacher to decide when to give advice, using the notion of state *importance*. Intuitively, a state is considered important if taking a wrong action in that state can lead to a significant decrease in future rewards, as determined by the teacher's Q-values. Formally, the importance of a state, denoted $I(s)$, is defined as:

$$I(s) = \max_a Q^{teacher}_{(s,a)} - \min_a Q^{teacher}_{(s,a)} \quad (1)$$

Three variations of this heuristic were suggested: (1) ***Advise Important***: giving advice when $I(s) > t_{ti}$ (where t_{ti} is a predetermined threshold); (2) ***Correct Important***: giving advice if $I(s) > t_{ti}$ and $\pi_{student}(s) \neq \pi_{teacher}(s)$. This heuristic assumes that the teacher has access to the student's chosen action for the current state; (3) ***Predictive Advising***: giving advice if $I(s) > t_{ti}$ and the teacher predicts that the student will take a sub-optimal action. This approach assumes that the teacher does not know the student's intended action and instead develops a predictive model of the students' actions over time. We do not include *Predictive Advising* in our study to avoid the assumption that a person would develop a predictive model of the students' actions. Moreover, even with agent teachers, the ability to predict an action will greatly depend on the size of the action space.

We also evaluate the ***Early Advising*** baseline heuristics used by Torrey & Taylor. Using the *Early Advising* heuristic, the teacher gives advice in all states until the entire advice budget is spent. Similarly, with ***Early Correcting***, the teacher advises the student in any state in which $\pi_{student}(s) \neq \pi_{teacher}(s)$ until exhausting the advice budget (as *Correct Important*, this heuristic assumes that the student's intended action is communicated to the teacher).

3.2 Student-Initiated Advising

We consider several heuristics for the student agent to determine when to ask the teacher for advice. Similarly to *correct important*, the ***Ask Important*** heuristic uses the notion of state-importance to decide whether the student should ask for advice. It uses the *student's* Q-values when computing Equation 1 and asks for advice when $I(s) > t_{si}$, where t_{si} is a threshold set for the student agent.

The ***Ask Uncertain*** heuristic [Clouse, 1996] also considers Q-values differences (Equation 1) to decide whether to ask for advice, but differs from *Ask Important* in that it asks for advice when the difference is *smaller* than a given threshold t_{unc} (Equation 2). Intuitively, low Q-value difference signals that the student is uncertain about which action to take. This heuristic asks for advice when:

$$\max_a Q^{student}_{(s,a)} - \min_a Q^{student}_{(s,a)} < t_{unc}, \quad (2)$$

where t_{unc} is the given student's threshold for uncertainty.

Chernova & Veloso [Chernova and Veloso, 2007] used the distance from a state to its previously visited nearest-neighbor state as one measure of confidence that is based on the agent's familiarity with the state. We implement this approach in the ***Ask Unfamiliar*** heuristic. In our settings states are described by a feature vector and we use Euclidean distance between feature vectors to determine the nearest neighbor. The student then asks for advice when:

$$distance(s, NN(s)) > t_{unf}, \quad (3)$$

where $NN(s)$ is nearest neighbor of state s.

3.3 Jointly-Initiated Advising

The student-initiated and teacher-initiated advising approaches both have shortcomings. The teacher-initiated approach requires the teacher to always pay attention. On the other hand, the advising decisions of the student are likely to be weak since they are guided by the student's noisy Q-value estimates. We design *jointly-initiated* advising approaches to address these shortcomings by having the right division of tasks between the teacher and the student. These approaches do not require the teacher to pay continuous attention while still utilizing the more informed signal of the teacher about whether advice is beneficial in a given state.

In *jointly-initiated* advising, the student decides whether to ask for the teacher's *attention* based on one of the student-initiated approaches for asking for advice. Then, the teacher decides whether to provide advice based on one of the teacher-initiated advising approaches. We denote a jointly-initiated heuristic by [X–Y], where X is a student heuristic for asking the teacher's attention and Y is a teacher heuristic for determining whether to give advice in the current state. For instance, ***[Ask Important–Correct Important]*** means that the *student* asks for the teacher's attention when $I(s)_{student} > t_{si}$. The teacher will then assess the state and will give advice if $I(s)_{teacher} > t_{ti}$.

Once the teacher decides to give advice, it will continue monitoring the student's actions until advice is no longer needed, and will later resume monitoring only when the student asks for the teacher's attention next[1]. The motivation for this approach is that once the teacher is already paying attention, it will be better able to judge whether additional advice is required in consequent states, until the student takes the right course of action. In addition, it requires less context-switching of the teacher.

4 Empirical Evaluation

Our experiments have four objectives: (1) **Comparing student-initiated and teacher-initiated strategies**: we assess the relative strengths and weaknesses of the existing approaches; (2) **Evaluating the proposed jointly-initiated approaches**: we compare the performance of the joint heuristics to that of the best-performing prior heuristics, as determined by the first experiment; (3) **Exploring the effect of the student's initial policy quality on performance**: in real-world settings, autonomous agents will likely start with some pre-programmed basic policy rather than learn "from scratch". Therefore, we evaluate the benefits of the teaching sessions when varying the quality of the student's initial policy. The quality of the initial policy is varied in two ways: by varying the length of the student's independent training (without access to teacher advice) prior to the teaching session, and by pre-training the student in limited settings that do not include some important features of the game, so that the student cannot learn certain skills; (4) **Exploring the effect of sharing the student's intended action with the teacher**: while sharing the student's intended action can reduce the use of the teacher's advice budget, sharing the student's action might be infeasible in some domains, and also incurs additional communication costs. Therefore, we explore the extent to which sharing the intended student action benefits learning.

[1] We evaluated the student-initiated approaches using this continued monitoring, but it did not lead to significant differences.

Figure 1: The Pac-Man Game.

4.1 Experimental Setup

We used the Pac-Man vs. Ghosts League competition [Rohlfshagen and Lucas, 2011] as our experimental domain. Figure 1 shows the game maze used in our experiments. This game configuration includes two types of food pellets: regular pellets (small dots) are worth 10 points each and power pellets (larger dots) are worth 50 points each. In addition, after eating a power pellet, ghosts become edible for a limited time period. Pac-Man receives 200 points for each eaten ghost. A game episode ends when Pac-Man is eaten by a ghost, or after 2000 time steps. Ghosts chase Pac-Man with 80% probability and otherwise move randomly. In each state, Pac-Man has at most four moves (right, left, up or down).

Due to the large size of the state space, we use a high-level feature representation for state-action pairs. Specifically, we use the 7-feature representation from Torrey & Taylor's [2013] implementation. Q-values are defined as a weighted function of the feature values $f_i(s,a)$:

$$Q(s,a) = \omega_0 + \sum_i \omega_i \cdot f_i(s,a) \quad (4)$$

The student agent employed the Sarsa(λ) algorithm to learn the weights in Equation 4. We used the same parameter configuration as Torrey & Taylor [2013]: $\epsilon = 0.05, \alpha = 0.001, \gamma = 0.999, \lambda = 0.9$. The teacher agent was trained with the same learning configuration until its performance converged. Table 1 summarizes the heuristics for teacher-initiated and student-initiated advising strategies studied in our experiments, together with the thresholds we used for each of them. The thresholds were determined empirically.

4.2 Evaluation Metrics

We evaluate the student's *learning rate* by assessing the student's performance at different time points during training. Specifically, in each trial, we paused training after every 100 game episodes and evaluated the student's policy at that time point by averaging 30 evaluation episodes (in which the student uses its current policy without exploring or updating its Q-values). Because trials have high variance depending on the student's eploration, we generate a learning curve by aggregating 30 separate trials. For example,Figure 2 (left) shows the learning curves comparing the performance of teacher-initiated and student-initiated approaches. The x-axis represents training episodes, and y-axis values show the average episode reward at that point in the training session.

For the teacher, *cumulative attention* is evaluated by averaging the number of states in which the teacher was asked to monitor the student. We assume that the teacher completely stops monitoring once the advice budget is fully used. Cumulative attention curves are generated by averaging the total number of states in which the teacher's attention was requried after every 100 game episodes, averaging these values over 30 trials. The left plot of Figure 2 shows a cumulative attention curve. As in the learning curve, the x-axis corresponds to training episodes. The y-axis values show the number of states in which the teacher's attention was required up to a given time point.

The overall *attention duration* required from the teacher is the average number of states it takes to use the entire advice budget. This metric can also be assessed by looking at the x-value of the point in which the cumulative attention (y-value) flattens in the cumulative attention curves. For example, in Figure 2 (right), when using the *correct important* heuristic, the overall duration of required attention is 90 episodes, while *early correcting* only requires an attention duration of 10 episodes as the advice budget gets used quickly.

To assess the statistical significance of differences in average rewards and cumulative attention, we ran paired t-tests comparing the averages after each 100 training episodes.

Heuristic	**Initiator**	**Threshold**	**Shared Action**
Early Correcting	teacher	None	Yes
Early Advising	teacher	None	No
Correct Important	teacher	200	Yes
Advise Important	teacher	200	No
Ask Important	student	50	Yes
Ask Uncertain	student	30	Yes
Ask Unfamiliar	student	Avg. distance to nearest neighbor	Yes

Table 1: Student-initiated and teacher-initiated heurisitcs.

4.3 Teacher Vs. Student Heuristics

Figure 2 (left) shows the student's learning rate when using heuristics for teacher-initiated and student-initiated advising. When using the teacher-initiated approaches, the teacher constantly monitors the state of the student until the advice budget runs out. When student-initiated advising approaches are used, the teacher only monitors the student when it is asked to advise. In all cases, the student's intended action is available to the teacher when giving advice.

Substantially and signficantly higher learning gains were obtained when using the teacher-initiated *Correct Important* heuristic compared to all other heuristics ($p < 10^{-16}$). This can be seen in Figure 2 (left). For example, after 200 training episodes with the teacher, an average reward of 3055.64 is obtained when using *Correct Important*, compared to an average reward of 2688.03 when using the next best heuristic. All other heuristics led to higher learning rates compared to the *no advice* (green dashed line) condition ($p < 10^{-10}$). There were no statistically significant differences between any other pairs of advising strategies.

The higher learning gains obtained when using teacher-initiated advising are expected; the teacher has more knowl-

edge than the student about the domain and has a good policy for making decisions in it, which allows it to choose effective teaching opportunities. Consider the game state shown in Figure 1 as an illustrative example. Intuitively, this is an important state: if Pac-Man (the student) makes a wrong move, it might be eaten by a ghost; if, however, it proceeds towards the power pellet, it will have an opportunity to earn a high reward for eating a ghost. The teacher, which already knows the environment, can identify that this state is important based on its Q-values, while the student might not yet have enough information to come to this conclusion.

While the *Correct Important* heuristic results in the highest learning gains, it requires significantly more teacher attention than the other approaches. This can be seen in Figure 2 (right), which shows the average cumulative attention required of the teacher; i.e., the total number of states in which teacher's attention was required up to a given episode. Teacher attention is required in significantly more states when using *Correct Important* compared to all other heuristics (72382.06 states compared to only 6358.86 states for *Ask Unfamiliar*, which is the most attention-demanding student-initiated heuristic, $p < 0.0001$). In addition, the overall duration of teacher's attention (i.e., number of episodes until the advice budget is fully used) is larger for *Correct Important* (90 episodes compared to less than 40 episodes required when using any of the other heuristics).

4.4 Jointly-Initiated Teaching Strategies

The results reported so far show that the teacher-initiated advising strategies outperform the student-initiated ones in terms of learning gains, but require more attention. In this subsection, we present results from an evaluation of the jointly-initiated teaching strategies, which aim to reduce the attention required from the teacher while maintaining the benefits of student-initiated advising. We thus compare their performance with that of the top performing teacher-initiated teaching strategy, *Correct Important*.

As Figure 3 (left) shows, when using the heuristic *[Ask Important–Correct Important]*, the student obtains similar rewards to those obtained when using *Correct Important*. The rewards at any given time point were on average slightly higher when using *[Ask Important–Correct Important]*, but while this difference was statistically significant ($p = 0.008$), it was not substantial (average difference of 18.5 points). Figure 3 (right) shows that the *[Ask Important–Correct Important]* heuristic required the teacher's attention in fewer states (64,711.47 states compared to 72,382.07 states). This difference was statistically significant ($p < 10^{-5}$), and substantial. However, the overall duration of required teacher's attention when using *[Ask Important–Correct Important]* is 140 episodes (indicated by the x value corresponding to the maximal total attention), compared to 50 episodes when using the *Correct Important* heuristic. That is, while the jointly-initiated teaching strategy requires the teacher's attention in fewer states, the duration of the training session, and thus teacher's needed attention span, is longer.

To ensure that the performance is achieved as a result of the student's choice of states in which to ask for advice, we also evaluate a random baseline where the student asks for the teacher's attention with 0.5 probability (the average rate of asking for advice by the *Ask Important* heuristic until the advice budget runs out). As shown in Figure 3, this random baseline (*Ask Random–Correct Important]*, dashed purple) does not perform as well as *[Ask Important–Correct Important]*. Moreover, it requires significantly more cumulative teacher attention, as well as a longer teaching period (both attention and learning gains differences were statistically significant, $p < 10^{-5}$). This shows that while the student's perception of importance is not as accurate as that of the teacher, it is still useful for identifying advising opportunities.

The strength of the *[Ask Important–Correct Important]* heuristic is its recall for important states. While the student heuristic *Ask Important* has many false positives when trying to identify important states due to the students' inaccurate Q-values, combining it with the teacher's *Correct Important* heuristic, which assesses whether the state is truly important, mitigates this weakness.

The other jointly-initiated teaching strategies, *[Ask Uncertain–Correct Important]* and *[Ask Unfamiliar–Correct Important]*, lead to some improvement in learning rate compared to the *No Advice* baseline, but perform significantly worse than *Ask Random* and require more cumulative attention, because they rarely capture important states. That is, they suffer from a high false negative rate when trying to identify important states, and therefore when the teacher uses *Correct Important* in combination with these approaches, it typically decides not to give advice (as the state is not important). This is evident by the long duration it takes until the advice budget runs out when using these heuristics (Figure 3).

[Ask Uncertain–Correct Important] suffers from a higher false negative rate because in its essence, *Ask Uncertain* captures states with a small Q-value range rather than those with a high one. While in some of these states the student might be uncertain of its actions, it might also mean that none of the actions will lead to significantly decreased performance. *[Ask Unfamiliar–Correct Important]* also suffers from possibly missing important states and using the advice when its impact is smaller, because unfamiliar states might not be important ones. In addition, appropriately identifying unfamiliar states likely requires more sophisticated domain-dependent similarity methods.

4.5 The Effect of Student's Initial Policy

The quality of the students initial policy may affect the effectiveness of different advising strategies. To gain insights into this relationship, we experimented with student agents that differ in the quality of their initial policies.

We observe similar trends and relative performance of the different teaching strategies when varying the length of the student's independent training prior to the teaching session. As the quality of the initial policy of the student improves (i.e., the student's initial policy is based on more independent learning episodes, in the same game settings), the performance of the jointly-initiated and student-initiated teaching strategies that are based on state importance can better identify important states in which the teacher's attention is required. However, in general, the overall benefit of advising the student decreases, as there is less room for improvement

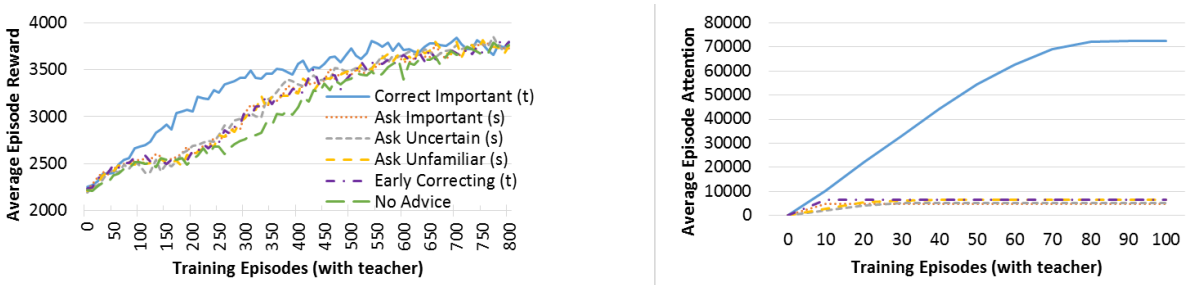

Figure 2: Average reward (left) and average attention (right) obtained by student-initiated and teacher-initiated approaches. The student was trained for 100 episodes prior to teaching, and actions were shared with teacher.

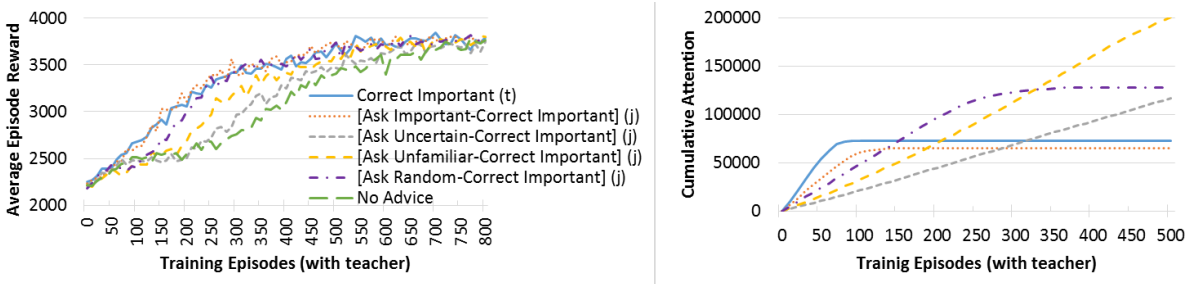

Figure 3: Average reward (left) and cumulative attention (right) obtained by jointly-initiated and teacher-initiated advising. The student was trained for 100 episodes prior to teaching, and actions were shared with teacher.

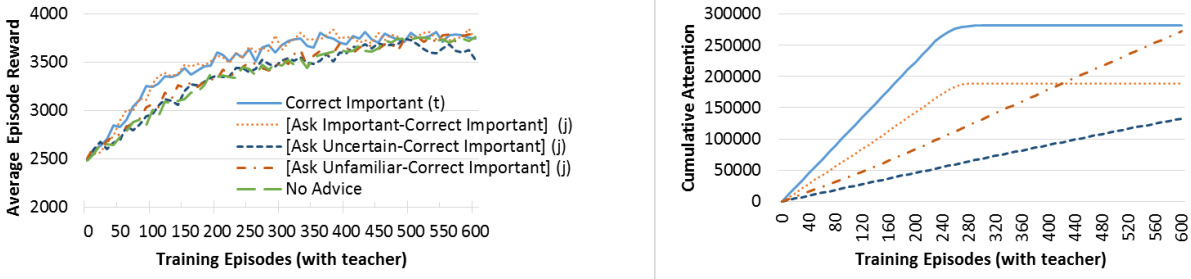

Figure 4: Average reward (left) and cumulative attention (right) obtained by jointly-initiated and teacher-initiated advising. The student was trained for 300 episodes prior to teaching.

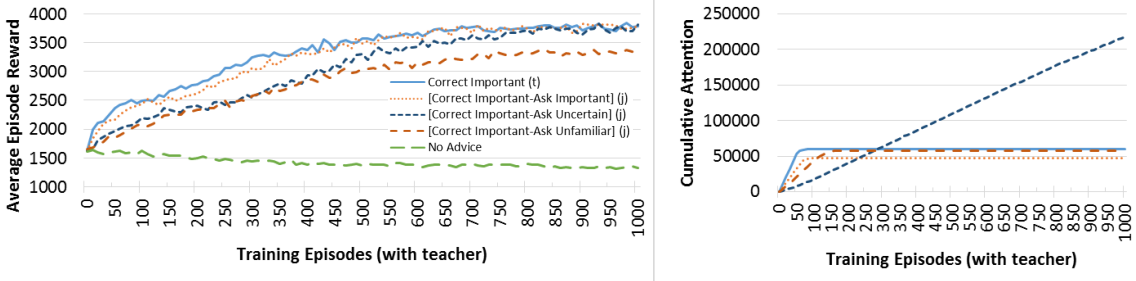

Figure 5: Average reward (left) and cumulative attention (right) obtained by jointly-initiated and teacher-initiated advising. The student trained for 150 episodes prior to teaching in game without power pellets, and actions were shared with teacher.

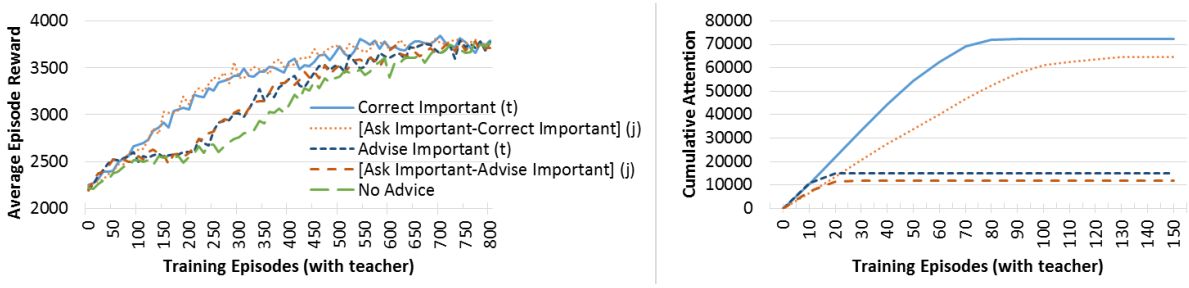

Figure 6: Average reward (left) and cumulative attention (right) obtained by jointly-initiated and teacher-initiated advising when the action is shared vs. when it is not shared. The student was trained for 100 episodes prior to teaching.

of higher-quality initial policies.

Figure 4 shows the performance of the jointly-initiated approaches and the *Correct Important* heuristic when the student's initial policy was established after 300 episodes of individual training. While the overall trends are similar to those obtained when the initial policy was determined after only 100 episodes (Figure 3), the reduction in the required teacher attention when using *[Ask Important–Correct Important]* compared to that required when using *Correct Important* increases when the student starts with a better initial policy. For example, when the student was independently trained for 100 episodes, advising based on *[Ask Important–Correct Important]* required attention in 7670.6 fewer states than advising based on *Correct Important*; the difference in required attention increased to 92162.6 states if the student was trained for 300 episodes independently. In addition, when the student had longer independent training, the overall attention duration when using *[Ask Important–Correct Important]*) was only 10 episodes longer than when using *Correct Important* (compared to 50 episode gap when the independent training only lasts 100 episodes).

Teacher advice is especially beneficial when the student learns its initial policy in a limited setting that does not allow the student to explore the complete state space. Figure 5 shows the performance of the different teaching strategies when applied to a student that developed an initial policy in settings without power pellets. Without any advising (green bottom line), the student's policy does not improve, as it does not manage to learn about the positive rewards of eating power pellets and consequently does not learn to eat ghosts. Since the student has already established low weights for features that correspond to the possibility of eating power pellets, without guidance it is not able to learn this new skill. However, with teaching, it quickly improves its policy.

While *[Ask Important–Correct Important]* is still the best performing heuristic for jointly-initiated advising, the *[Ask Unfamiliar–Correct Important]* heuristic does relatively better compared to settings in which the student was trained on the same game instance prior to teaching. Although it is outperformed by the *[Ask Uncertain, Correct Important]* heuristic in terms of student performance (after 400 episodes), it requires significantly less teacher attention. The relative performance improvement for the *Ask Unfamiliar* approach for getting the teacher's attention can be explained by the fact that states that involve high proximity to power pellets may appear less familiar (as they were not included in the initial independent training), and also correspond to important states.

4.6 Sharing the Student's Intended Action

Correct Important and *[Ask Important–Correct Important]* both assume that the intended action is shared with the teacher so that the teacher can correct the student if the state is considered important *and* the student's intended action is suboptimal. In contrast, *Advise Important* and *[Ask Important–Advise Important]* give advice when the state is considered important, regardless of the student's intended action. As Figure 6 (left) shows, sharing the intended action significantly improves performance for all heuristics as it allows the teacher to save the teaching budget to prevent student mistakes. However, it also requires more attention from the teacher (right figure). The effect of sharing the action is similar for the teacher-initiated and jointly-initiated approaches.

5 Conclusion and Future Work

This paper investigates interactive teaching strategies in a student-teacher learning framework. We show that using a joint decision-making approach, the amount of attention required from the teacher can be reduced compared to teacher-based approaches, while providing similar learning benefits.

Our empirical evaluation used the Pac-Man game, which is an example of a dynamic domain where making mistakes in some states is particularly harmful (e.g., being eaten by a ghost). Real-world dynamic settings such as semi-autonomous driving will also likely have these types of important states (e.g., causing an accident). However, the importance heuristic will likely not generalize well to domains where "unrecoverable" mistakes are rare (e.g., mapping a new territory, where each action can be immediately reversed). In such settings, unfamiliarity- and uncertainty-based heuristics are likely to be more useful.

There are several directions for future work. Methods for generalizing the teacher's advice beyond a specific state could further accelerate learning. For instance, the student knows the teacher decided not could try to learn a model of important states based on the teacher's responses to advice requests. From the human-agent interaction perspective, we plan to study people's attention while interacting with semi-autonomous systems and incorporate attention models into

the student's strategies for asking for advice. As AI systems continue seeking input from humans in critical domains such as driving, we see value in developing approaches that reason about considerations specific to humans and benefit from the complementary abilities of humans and machines.

Acknowledgments

We thank Eric Horvitz for helpful comments and suggestions.

References

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

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

[Chernova and Veloso, 2007] Sonia Chernova and Manuela Veloso. Confidence-based policy learning from demonstration using gaussian mixture models. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, page 233. ACM, 2007.

[Clouse, 1996] Jeffery Allen Clouse. On integrating apprentice learning and reinforcement learning. 1996.

[Cummings and Mitchell, 2008] Mary L Cummings and Paul J Mitchell. Predicting controller capacity in supervisory control of multiple uavs. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 38(2):451–460, 2008.

[Fong *et al.*, 2002] Terrence Fong, Charles Thorpe, and Charles Baur. Robot as partner: Vehicle teleoperation with collaborative control. In *Multi-robot systems: From swarms to intelligent automata*, pages 195–202. Springer, 2002.

[Goodrich *et al.*, 2001] Michael A Goodrich, Dan R Olsen, Jacob Crandall, and Thomas J Palmer. Experiments in adjustable autonomy. In *Proceedings of IJCAI Workshop on Autonomy, Delegation and Control: Interacting with Intelligent Agents*, pages 1624–1629, 2001.

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

[Horvitz *et al.*, 1999] Eric Horvitz, Andy Jacobs, and David Hovel. Attention-sensitive alerting. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 305–313. Morgan Kaufmann Publishers Inc., 1999.

[Horvitz *et al.*, 2003] Eric Horvitz, Carl Kadie, Tim Paek, and David Hovel. Models of attention in computing and communication: from principles to applications. *Communications of the ACM*, 46(3):52–59, 2003.

[Judah *et al.*, 2010] Kshitij Judah, Saikat Roy, Alan Fern, and Thomas G Dietterich. Reinforcement learning via practice and critique advice. In *AAAI*, 2010.

[Judah *et al.*, 2014] Kshitij Judah, Alan P Fern, Thomas G Dietterich, et al. Active lmitation learning: formal and practical reductions to iid learning. *The Journal of Machine Learning Research*, 15(1):3925–3963, 2014.

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

[Loftin *et al.*, 2015] Robert Loftin, Bei Peng, James MacGlashan, Michael L Littman, Matthew E Taylor, Jeff Huang, and David L Roberts. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Autonomous Agents and Multi-Agent Systems*, pages 1–30, 2015.

[Miller *et al.*, 2015] David Miller, Annabel Sun, Mishel Johns, Hillary Ive, David Sirkin, Sudipto Aich, and Wendy Ju. Distraction becomes engagement in automated driving. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 59, pages 1676–1680. SAGE Publications, 2015.

[Rohlfshagen and Lucas, 2011] Philipp Rohlfshagen and Simon M Lucas. Ms pac-man versus ghost team cec 2011 competition. In *Evolutionary Computation (CEC), 2011 IEEE Congress on*, pages 70–77. IEEE, 2011.

[Rosenstein *et al.*, 2004] Michael T Rosenstein, Andrew G Barto, Jennie Si, Andy Barto, Warren Powell, and Donald Wunsch. Supervised actor-critic reinforcement learning. *Handbook of Learning and Approximate Dynamic Programming*, pages 359–380, 2004.

[Rosman and Ramamoorthy, 2014] Benjamin Rosman and Subramanian Ramamoorthy. Giving advice to agents with hidden goals. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 1959–1964. IEEE, 2014.

[Tambe *et al.*, 2002] Milind Tambe, Paul Scerri, and David V Pynadath. Adjustable autonomy for the real world. *Journal of Artificial Intelligence Research*, 17(1):171–228, 2002.

[Torrey and Taylor, 2013] Lisa Torrey and Matthew Taylor. Teaching on a budget: Agents advising agents in reinforcement learning. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*, pages 1053–1060. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

[Zilberstein, 2015] Shlomo Zilberstein. Building strong semi-autonomous systems. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 4088–4092, 2015.