Policy Shaping with Human Feedback

Through my research in Professor Charles Isbell's lab, I have made contributions to the study of Interactive Machine Learning. Interactive Machine Learning poses the simple but powerful research question: "how can machines take better advantage of a person's input in order to learn?" My research aims to optimize the information obtained from humans while building infrastructure through which humans can interact with a machine learning agent in an intuitive and natural way. The end goal of this work is to create human-centered machine learning so that end users can interact with artificial intelligence (AI) and effectively guide its learning process without having to become machine learning experts themselves.

I am contributing to a doctoral student's project that tackles this research question, specifically in the context of reinforcement learning. Reinforcement learning consists of a set of techniques that approach the task of learning to navigate environments modeled by Markov Decision Processes (MDPs).[1] An MDP consists of a set of states that an AI agent can visit, actions available in each state, and a reward that the agent receives in each state. The agent seeks to find a *policy* that maps each state to a specific action that it should take to be successful. The highlight of my work was building infrastructure for experiments that test the performance of *policy shaping*, an algorithmic framework developed by Griffith *et al*, which integrates human feedback into an AI agent's internal learning algorithm.[2] Griffith *et al* introduces an algorithm called **Advise**, which treats the human feedback as direct policy labels. The algorithm then combines its estimate of the human's policy with the agent's policy learned using the Bayesian Q-learning algorithm. Bayesian Q-learning is an algorithm through which an agent interacts with an environment, maintains and updates estimates of the long-term expected discounted reward it can accumulate via a particular policy, and uses these estimates to learn the optimal policy.[3]

While Griffith *et al* runs experiments using a simulated oracle to test the performance of the **Advise** algorithm, I measured how well the **Advise** algorithm performs when people use it to train an agent whose goal is to navigate a maze environment successfully. In my experiments, I varied two parameters: consistency and likelihood. Consistency, denoted by C, is the probability that the feedback received by the agent is consistent with the feedback that the human intends to provide. Likelihood, denoted by L, represents the probability that feedback is received by the agent after performing an action, where this feedback indicates whether this action was good or bad with respect to the agent's goal.

In order to vary L and C, I developed two modes of interaction between the human and the AI agent. The first involves two buttons that the human can press, a plus button and a minus button. These buttons indicate the positive or negative feedback the human may wish to provide in response to a particular action made by the agent. With this approach, $C = 1$, since the button clearly communicates the intended feedback. The second mechanism is a pipeline through which a human speaks to the agent as it is navigating the environment. I used Carnegie Mellon's Sphinx speech recognition tool[4] to transcribe the speech to text and Stanford's Sentiment Analysis tool[5] to extract either positive, neutral, or negative sentiment from this text, which I inputted into the **Advise** algorithm as the human's feedback. Due to several factors with this process discussed below, $C < 1$ with this approach. With both approaches, L was varied by providing feedback after each action in one set of trials ($L = 1$) and providing feedback whenever the human interacting with the agent deemed it necessary ($L < 1$) in another set of trials.

Through building the speech-to-sentiment pipeline, I discovered that even if algorithms have theoretical guarantees, they may have implementation challenges in practice. Specifically, I

found the modification of policy shaping to incorporate human feedback to be significantly more difficult than when using a simulated oracle. First, the nature of the speech-to-sentiment pipeline causes *C* to be less than 1. The Sphinx speech recognition system and Stanford sentiment analysis tool are not 100% accurate, producing noise in the conversion of the human's speech to sentiment. Additionally, it takes longer for a human to speak than to press a button, causing a delay in processing, which may occur several moves later. Because the algorithm attributes the feedback the human provides to the most recent action the agent took, human input is incorrectly communicated to the agent. The feedback should be attributed to the state and action which the human was speaking about, which, in this case, was several moves in the past.

To address these issues, I made a novel modification to the policy shaping algorithm. I implemented a function that maintained a table storing the history of states the agent visited, the corresponding actions the agent made in those states, and the corresponding times that the agent visited those states. I also measured the average speech-to-sentiment conversion time in 100 trials to be seven seconds and estimated three seconds for the length of human feedback for a particular action via speech. As a result, when feedback was outputted by the speech-to-sentiment pipeline, I updated the **Advise** algorithm's state-action pair that occurred ten seconds ago by identifying the correct state and action in the history tables I had constructed. Through this process, I learned that in order to achieve the theoretical guarantees of an AI algorithm, it is necessary to design architecture that is robust to the variability of its components, such as the unpredictability of human speech and feedback in the case of the policy shaping framework.

Another key contribution of my work was the set of experiments I designed to test how policy shaping works in practice. In my experiments, a human interacts with an agent learning to navigate a maze environment in one of the two modes described above. Each trial consists of a series of policy shaping episodes and offline episodes. In a policy shaping episode, the human can see the agent moving in the maze environment and can therefore provide feedback in real time. In an offline episode, the human's feedback from the policy shaping episodes guides the agent's exploration of the environment while no further feedback is provided. Instead, the agent uses its internal Bayesian Q-learning algorithm to update its estimates of the best actions to take. In these episodes, the agent's moves are not visible to the human. In each trial, which comprised a series of policy shaping episodes and offline episodes, I recorded the total reward accumulated by the agent during each episode in addition to the distribution of positive vs. negative feedback provided by the human in each policy shaping episode. For each mode of interaction and setting of the *L* parameter, I averaged the results of five trials. In addition to these experiments, I also recorded the results of Bayesian Q-learning without policy shaping as a baseline.

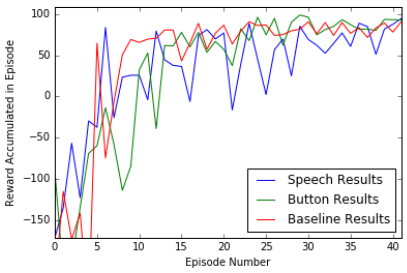

Figure 1. Episode vs Total Reward graph for the three experiments conducted with L < 1.

The graphs displayed illustrate the results of the experiment where $L < 1$. The Episode vs Total Reward graph (see Figure 1) displays the rate at which the agent learns optimal path through the maze environment. While the results for when the human uses the button to interact with the agent were slightly better than the baseline results as shown in this graph, the results for when a human speaks to the agent were surprisingly worse. One potential reason for the discrepancy between the performance of the two modes of interaction is that I could not set *L* to the same value for both options; even if the human provided feedback in the same manner in both cases, the likelihood of feedback was inherently different when talking than when pressing a

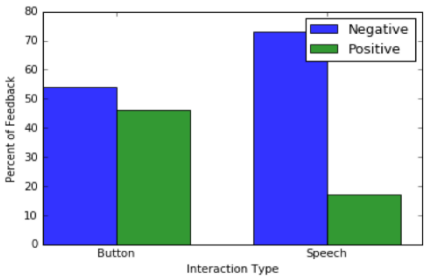

Figure 2. Distribution of positive and negative feedback for the two modes of human interaction with the AI agent.

button. Another potential cause was that the feedback the human provided was still not being attributed to the correct state, even after my algorithmic innovations improved the consistency of the feedback framework. In future work, a more robust way to measure the length of human speech needs to be developed so that the agent correctly receives the human's intended feedback.

An important qualitative result from these experiments was that the agent didn't react to the feedback the human provided immediately. For example, if the agent was traveling in a particular direction and the person said, "No, don't do that!" the agent did not change direction right away, which is what a human might expect it to do. This phenomenon occurred because human feedback may not fully affect the agent's decision making until several episodes later after it has explored the state space many times. Consequently, the majority of the human feedback provided was negative for both modes of interaction (see Figure 2), especially for interaction via speech.

While the agent's delayed response to feedback is not an issue in the case of a simulated oracle, this problem has negative implications when an actual person uses it, since the human teacher may become frustrated or confused. The ultimate goal of policy shaping is to apply the **Advise** algorithm to more sophisticated human-robot interaction, so in future work, the problems I identified may need to be addressed through algorithmic changes to the **Advise** algorithm's update function. Additionally, my graduate student mentor and I are currently designing a human subject experiment to test the algorithm with many people and concretely measure whether people are satisfied with the AI's response to their advice.

In summary, I made novel contributions to the **Advise** algorithm and tested how well the algorithm serves its ultimate purpose: to allow humans to teach AI. The infrastructure I designed provides a natural mechanism for human communication with an AI agent, while the results from experiments I designed suggest that modifications to the **Advise** algorithm must occur prior to using the policy shaping framework in human-robot interaction.

References:

1. Russell, S. J., & Norvig, P. (2014). *Artificial Intelligence: A Modern Approach.* Harlow: Pearson.
2. Griffith, S.; Subramanian, K.; Scholz, J.; Isbell, C.; and Thomaz, A. L. 2013. Policy shaping: Integrating human feedback with reinforcement learning. In Burges, C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K., eds., Advances in Neural Information Processing Systems 26. Curran Associates, Inc. 2625–2633.
3. Dearden, R., Friedman, N. & Russell, S. (1998), Bayesian Qleaming, in 'Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)'.
4. Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., Warmuth, M., & Wolf, P. (2003). The CMU SPHINX-4 speech recognition system. In Proceedings of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing. Hong Kong.
5. C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014, pp. 55–60.