Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

Policy Shaping with Human Teachers

Thomas Cederborg, Ishaan Grover, Charles L Isbell and Andrea L Thomaz

Abstract

In this work we evaluate the performance of a *policy shaping* algorithm using 26 human teachers. We examine if the algorithm is suitable for human-generated data on two different boards in a pac-man domain, comparing performance to an oracle that provides critique based on one known winning policy. Perhaps surprisingly, we show that the data generated by our 26 participants yields even better performance for the agent than data generated by the oracle. This might be because humans do not discourage exploring multiple winning policies. Additionally, we evaluate the impact of different verbal instructions, and different interpretations of silence, finding that the usefulness of data is affected both by what instructions is given to teachers, and how the data is interpreted.

1 Introduction

A long-term goal of *Interactive Machine Learning* is to create systems that can be interactively trained by non-expert end-users. Researchers have shown that there are multiple ways to interpret human feedback, and that some interpretations can lead to better learning algorithms [Thomaz and Breazeal, 2008b]. This paper investigates how human feedback is best interpreted, evaluating the usefulness of an interpretation by the performance of an artificial learner basing policy updates on the interpretation. We extend work on so-called *policy shaping*, where feedback to a reinforcement learning agent is interpreted as an evaluation of an action choice, instead of an evaluation of the resulting state, or an estimate of the sum of future discounted reward.

In particular, in [Griffith *et al.*, 2013], researchers focused on understanding theoretical limits of this approach, only evaluating the algorithm with data from a simulated oracle, and comparing the algorithm's performance to reward shaping-based techniques.The primary contribution of our work in this paper is testing policy shaping with real human teachers, exploring whether actual human teachers are able to achieve similar performance gains.

Further, we show in our experiments that although human teachers are willing to provide a great deal of feedback, they are silent 30% of the time on average. Thus, it becomes important for a policy shaping algorithm to interpret that silence. We compare: (1) the effects of different instructions to people regarding the meaning of silence and (2) the performance impact of different interpretations of silence by the agent.

2 Related Work

Interpreting feedback from humans who are evaluating artificial learners can be problematic for several reasons. Studies have shown that human behavior violates many common assumptions routinely built into machine learning algorithms, and reinforcement learning algorithms in particular [Isbell *et al.*, 2006; Thomaz and Breazeal, 2008b].

People are committed to achieving a joint goal with the learner, and instinctively try to provide rich information about how this goal should be achieved. For example, human teachers will use a mechanism meant as a channel for evaluating actions to try to motivate the learner, or to try to evaluate hypothesized future learner actions. In [Knox *et al.*, 2012] an algorithm is introduced allowing a learner to take advantage of feedback intended for future learner actions. Having a button dedicated to motivational communication was shown to reduce the teacher's tendency to use the evaluation channel for encouragement [Thomaz and Breazeal, 2008b]. As a result, use of the evaluation channel is closer to the assumptions of the learning algorithm, improving performance. The principle holds for other information channels as well. For example, if human teachers see what types of demonstrations will be useful, they produce more useful demonstrations [Cakmak and Lopes, 2012].

Human guidance also results in more focused, but less diverse, skill development [Thomaz and Breazeal, 2008a]. It is possible to explicitly model which parts of the context are visible to a human teacher and allow this more advanced model to influence learning from that teacher [Breazeal *et al.*, 2009], interpreting the human critique as referring to that teacher's flawed model of the world, as opposed to the actual world. Humans might also use positive and negative feedback in different ways [Thomaz and Breazeal, 2007]. Of particular interest for this work is the fact that the way people give feedback is unsuitable for standard optimization because of the human tendency to give more positive than negative reward, and to stop providing positive rewards when an agent appears to have learned the task [Isbell *et al.*, 2006].

These findings are relevant to the policy shaping approach as they motivate the need to interpret human-generated critique in policy space. Of course, the general problem of a learner interpreting a possibly-flawed human teacher is theoretically difficult, but there are a number of successful approaches, either by trying to modify a teacher's behavior by communicating confusion [Breazeal *et al.*, 2005], or by improving the learner's model of the teacher [Lopes *et al.*, 2011; Grizou *et al.*, 2014; Cederborg and Oudeyer, 2014; Loftin *et al.*, 2014]. Our work is closer to the latter.

3 Preliminaries: Reinforcement Learning

Although we are interested in the general problem of interactive machine learning, our work here is focused specifically on policy shaping in Reinforcement Learning (RL). Typically, RL defines a class of algorithms for solving problems modeled as a Markov Decision Process (MDP).

A Markov Decision Process is specified by the tuple (S, A, T, R, γ) for the set of possible world states S, the set of actions A, the transition function $T : S \times A \rightarrow P(S)$, the reward function $R : S \times A \rightarrow \mathbb{R}$, and a discount factor $0 \leq \gamma \leq 1$. We look for policies $\pi : S \times A \rightarrow \mathbb{R}$, mapping state-action pairs to probabilities, which result in high rewards. One way to solve this problem is through Q-learning [Watkins and Dayan, 1992]. A Q-value $Q(s, a)$ is an estimate of the expected future discounted reward for taking action $a \in A$ in state $s \in S$. The Q-value of a state action pair is updated based on the rewards received, and the resulting state. In this paper we use Boltzmann exploration [Watkins, 1989] where the probability of taking an action is $Pr_q(a) = \frac{e^{Q(s,a)/\tau}}{\sum_{a'} e^{Q(s,a')/\tau}}$, where τ is a temperature constant.

In our experiments, parameters were tuned using only Q-learning performance, without teacher critique data, and the values used were $T = 1.5$, $\alpha = 0.05$ and $\gamma = 0.9$.

4 Policy Shaping

As noted in Section 2, building models of how human behavior should be interpreted can be very difficult; however, improving these models can lead to better performance, at least for those learning algorithms that assume these models hold. In this work we use policy shaping—specifically an algorithm introduced in [Griffith *et al.*, 2013]—where we interpret human feedback as evaluating action choices such as "move to the left". Policy shaping is in direct contrast to the reward shaping algorithms where human feedback is interpreted as an evaluation of the state resulting from an action, or an estimate of the sum of future discounted reward.

The model of human psychology underlying the choice of policy shaping is not far fetched: good actions lead to good evaluations, bad actions lead to bad evaluations. By contrast, the assumptions needed to justify treating human feedback as a value to be maximized is contradicted by experimental evidence, and requires that the non-expert human maintain a rather complex model of learning: instead of evaluating the action choice made by a learner, the teacher would need to keep track of and estimate the entire sequence of future rewards, and give a numerical estimate of this sum (so that, for example, an action creating a problem receives more negative reward than the sum of subsequent good action choices limiting the damage).

Finally, let's compare the two models in two different cases. First where the final state is the only thing that matters. And the second where the specific actions matter and the final state is always the same, such as in dancing. If success can be measured in the final state, then the human teacher would need to make sure the total reward is path independent, making it necessary for the teacher to keep track of an enormous amount of information. However, if the goal is to perform the correct actions, i.e. correct dance moves, then the policy shaping interpretation: "evaluations refer to action choice" is favored almost by definition.

In policy shaping, perhaps the most important parameter for learning is the probability that an evaluation of an action choice is correct for a given teacher. This is denoted as C. If an action is good (or bad), then the teacher is assumed to give positive (or negative) critique with probability C ($C = 0$ is a perfectly inverse teacher, $C = 0.5$ is a random non-informative teacher, and $C = 1$ is a flawless teacher). The algorithm makes the simplifying assumption that the accuracy of each critique instance is conditionally independent[1], resulting in the probability $Pr_c(a)$ that an action is good: $Pr_c(a) = \frac{C^{\Delta_{s,a}}}{C^{\Delta_{s,a}} + (1-C)^{\Delta_{s,a}}}$, where $\Delta_{s,a}$ is the number of positive minus the number of negative critique instances in a given data set for state s and action a. This corresponds to an *a priori* belief that, before viewing any critique, an action is good with probability 0.5. In our experiments for this paper we set $C = 0.7$. This value assumes that a human is correct 70% of the time (corresponding to significantly flawed, but still useful, evaluations).

During learning, the probability $Pr(a)$ of taking action $a \in A$ is $Pr(a) = \frac{Pr_q(a)Pr_c(a)}{\sum_{\alpha \in A} Pr_q(\alpha)Pr_c(\alpha)}$, combining the probability $Pr_q(a)$ derived from the the Q-values as specified above with probability $Pr_c(a)$ from the critique data (this product represents the maximum information available from two different, conditionally independent, sources [Griffith *et al.*, 2013]). It is worth noting that the policy shaping algorithm is almost identical to Q-learning in the cases with a very small amount of critique, or with critique that has many state action pairs with the same amount of positive and negative critique, or where C is close to 0.5. In the case with a lot of critique from a consistent and trusted teacher, the data will have a much more dramatic impact on action selection.

5 Experiment

Because the concept of policy shaping has already been shown to work with a simulated teacher [Griffith *et al.*, 2013], our primary research goal in the experiments that we will now describe is to investigate what happens when policy-shaping data is generated by human teachers. Additionally, we are interested in understanding the consequences of interpreting a teacher's silence as an action choice evaluation as well (*e.g.*,

[1] This is an approximation because when a human teacher judges a state-action pair incorrectly once, it may be more probable that this specific state action pair will be judged incorrectly again.

deciding that no news is good news, and so silence is evidence of a good action). Our experiment is designed to inform these two research questions, with our hypotheses being:

H1: Human teachers will provide good data for Policy Shaping but not as good as an oracle.

H2: People have an inherent bias that informs the meaning of silence.

H3: We can manipulate the meaning of people's silence by differing the instructions given.

5.1 Domain

We use the experimental domain of pac-man, because human teachers are easily familiar with it, and it is known to work with policy shaping from prior work. Pac-man consists of a 2-D grid with food, walls, ghosts, and the pac-man avatar. Eating all food pellets ends the game with +500 reward, and being killed by the ghost ends the game with -500 reward. Each food pellet gives +10 reward, and each time step pac-man gets a -1 time penalty. Pac-man's action space is to go up, down, right or left. The state representation includes pac-man's position, the position and orientation of any ghosts and the presence of food pellets. In this version of the game, a ghost moves forwards in a straight line if it is in a corridor. If it has a choice of actions it decides randomly, but does not go back to a square it has just occupied.

5.2 Design

Based on our hypotheses we have four groups to compare in this experiment:

- Oracle: this condition uses a simulated teacher, in a similar manner as prior work.
- Human-unbiased/open ended instructions: a human teacher provides action critiques, with no instruction about the meaning of silence.
- Human-positive bias: a human teacher provides action critiques, with instruction that silence is positive.
- Human-negative bias: a human teacher provides action critiques, with instruction that silence is negative.

In order to facilitate comparisons between these four conditions, each teacher gave critique on the exact same state action pairs, evaluating videos of a pac-man agent playing the game. Each evaluation video lasts 5 minutes and consists of a series of games. The games were generated by the experimenters playing the game until either winning or dying. Because all teachers evaluate the exact same state action pairs, any differences in learning performance can be attributed to the experimental condition. If they had been evaluating an agent online, who was executing a stochastic policy in a stochastic world, they would not be evaluating the same thing and one data set could for example be higher quality due to having been given the chance to evaluate different states.

We solicited participation from the campus community and had 26 volunteers provide data for this experiment. Each participant did one practice evaluation, followed by two evaluations with unbiased instructions, and another two evaluations with either the positive or negative biased instructions (not both). The full experimental protocol is shown visually in Figure 1.

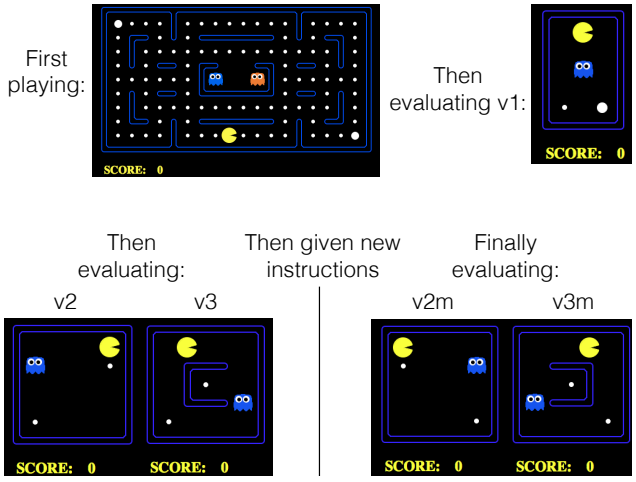

Figure 1: Each teacher first plays the large board to the top left. Then evaluates videos v1, v2 and v3. New instructions are given based on what group the teacher has been assigned to, then v2m and v3m are evaluated.

5.3 Protocol

Practice session:

First, participants were allowed to play the original pac-man board for 5 minutes, to familiarize themselves with the game. After having played a series of games, they were instructed about how to provide critique.

We call this set of instructions the open-ended instructions: "Use the "c" button to say "correct action", and the "b" button to say "bad action". Pac-man will take an action, and then stand still for a little while, which will give you time to evaluate the action just taken." Then they evaluated video v1. This first video allows them to become familiar with the procedure, and to experience the problem of commenting on a learner whose actions are hard to evaluate. The data gathered thus far is not used to test the algorithm.

We then test two different instruction conditions while trying to reduce between-group differences. With a limited sample size, one of the groups will be better at playing the game, give more feedback, or more positive feedback. These three values were recorded, and each new teacher was assigned to the group that would result in the smallest between group difference. Data collection started after group assignment.

Unbiased Evaluation:

Next the participants evaluate the two videos v2 and v3, taking place on the boards 2 and 3 shown in the lower left corner of Figure 1, again with the same open-ended instructions about the 'c' and 'b' buttons. For this step, both groups see exactly the same setup.

Silence-biased Evaluation:

After evaluating videos v2 and v3, participants were then told that the study is actually investigating how pac-man should

handle the cases where there is no feedback. Then each person was given one of the following two instruction sets depending on which group they were assigned.

- **Biased-good** – The teacher was told: Now you will provide evaluations in a case where pac-man will take silence to mean that it did not make a mistake. Press "b" whenever pac-man could have done better. You can be silent when pac-man is acting reasonably good. Press "c" only when you are sure that pac-man took the best action.
- **Biased-bad** – The teacher was told: Now you will provide evaluations in a case where pac-man will take silence to mean that it could have done better. Press "c" when pac-man is acting reasonably good. You can be silent when pac-man is not doing very good. Press "b" only when you are sure that pac-man has made a serious mistake.

Ideally we want to have people give the biased evaluation on the same state-action pairs as the unbiased, but it is hard to predict how a teacher's behavior will be affected by evaluating the exact same video two times in a row. This was solved by a combination of not seeing the two videos of the same board directly after each other, and by mirroring the videos after the biasing instructions were given. Board 2, used to generate video 2, or v2, can be seen in the lower left of figure 1. The mirrored version of board 2 can be seen to the lower right, in figure 1, after the new instructions are given. Pac-man starts to the left instead of the right, and the ghost starts to the right instead of to the left. The states are identical but with right and left reversed. The actions of video v2 was also reversed so that an equivalent game was observed in the video v2m. After receiving the biased instructions, participants evaluated videos v2m and v3m which are identical to v2 and v3, but mirrored. The mirroring of videos will allow us to evaluate human teacher data on the same states action pairs, but under different instruction conditions. They evaluate the same video twice, but it does not look like the same video to them.

5.4 Simulated oracle teacher

The Q-learning algorithm was used to obtain a policy π_o that always wins. This policy was then used to construct an oracle teacher who gave negative critique if a learner did something different from what π_o would have done, and positive critique if the learner acted in accordance with π_o. The oracle was then used to evaluate the videos v2 and v3, giving critique on exactly the same state action pairs as the human teachers. v2m and v3m would have resulted in identical results because they only differed in visual display and verbal instructions, both of which are ignored by the simulated teacher. The simulated teacher gave feedback to a state action pair if that state action pair was encountered during the Q-learning episode that provided the simulated teacher with its winning policy π_o. If the state action pair had not been encountered, it gave no feedback. It is not necessary to explore all possible states to find an optimal policy since some optimal policies reliably avoid large parts of the state space (and therefor does not need to know what to do there). The tendency of a simulated teacher to give feedback as opposed to be silent is thus dependent on which state action pairs it is asked to evaluate.

5.5 Learning

As mentioned above, policy shaping modifies the action selection of a reinforcement learning algorithm, in this case Q-learning with Boltzmann exploration. Policy shaping takes as input a data set and outputs a modification to Q-learning. In our case we compare data sets gathered during evaluations of videos, meaning that all data is available at the start of learning. A Q-learning agent explores the world, playing game after game, and at each step the action selection is modified by policy shaping operating on a fixed data set. While all Q-learning agents will eventually learn these boards, the scores achieved before having converged is dependent on the quality of the data set. Q-value estimates improve during learning, but the evaluation data, and thus the impact of policy shaping, stays the same while the agent learns.

6 Results

6.1 Measuring data quality

In these results we are primarily concerned with measuring differences in the quality of the evaluation data received across the experimental conditions. As a metric for this, we take the average quality during the 1000 games of one learning episode. Q-learning always finds an optimal policy on these boards, making the integration over quality a more informative measure than final performance. Each policy is a distribution over actions for each state. During learning, actions are sampled from this policy to enable exploration. During evaluation, the maximum action is always selected to get a more accurate estimate of how well the learner is able to perform. The quality of a specific policy is defined as the expected reward the learner will achieve when it plays the game and takes the action with the maximum value (as opposed to sampling from a distribution in order to explore the state space and learn). The quality of data is the expected average score during a learning episode of 1000 games.

6.2 Our participants were better teachers

In Table 1 we see the quality of the data for the three groups of our 26 participants. The simulated teacher/oracle is also included, as well as a baseline case with no teacher. The standard deviations are denoted σ, and we indicate the 95% confidence interval with $\pm$. Learning episodes lasting for 1000 games are performed repeatedly with each data set. The average score of the 1000 games of a learning episode constitutes one sample. The sample size was 100 for the "no teacher" and "oracle" case, 780 for the "Open ended case" (30 learning episodes for each of the 26 teachers), and 390 for the biased good and biased bad case (30 learning episodes for each of the 13 teachers in the respective group).

We can see that the simulated teacher produces lower quality data than the human teachers, this is counter to our hypothesis **H1** about the usefulness of human generated data. The oracle knows of one policy that always wins, but does not recognize any other winning policies. When investigating the

Table 1: Teacher Score Comparison

Teacher	Board 2	Board 3
No teacher	$426 \pm 2.29, \sigma = 11.7$	$354 \pm 7.17, \sigma = 36.6$
Oracle	$444 \pm 1.98, \sigma = 10.1$	$372 \pm 5.33, \sigma = 27.2$
Open ended	$462 \pm 1.08, \sigma = 15.4$	$413 \pm 3.99, \sigma = 56.9$
Biased good	$447 \pm 1.76, \sigma = 17.7$	$371 \pm 5.32, \sigma = 53.6$
Biased bad	$459 \pm 1.77, \sigma = 17.8$	$410 \pm 5.56, \sigma = 56$

evaluations in detail, the simulated teacher was observed giving clearly bad feedback in a number of states. Actions were given bad critique even though they represented one way of winning the game, but not the one found by the teacher. The simulated teacher is simply deciding that the action is different from what is recommended by its own winning policy π_o. There are however several high stake situations where there is only one reasonable action, for example when only one action will escape the ghost, and here the simulated teacher gives the right feedback. The simulated teacher will also never advice moving into the ghost, so much of the time the feedback is reasonable, and the information the learner receives is still useful. A human teacher might however approve of several different good strategies, and can therefore beat our simulated teacher.

6.3 Instruction conditions

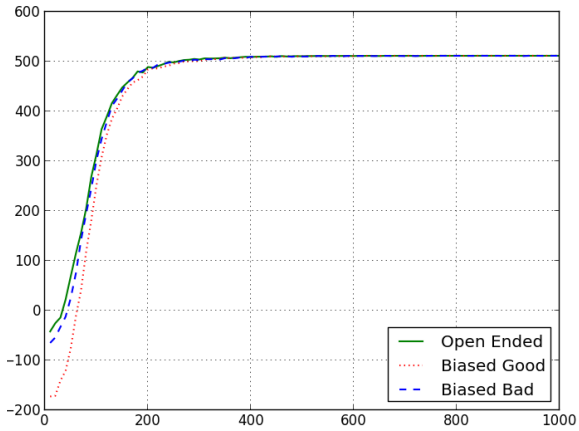

Figure 2: Here we can see the average performance of the learning agent as a function of number of games played on board 2, using evaluations of the state action pairs in the V2 video. The three lines correspond to the three instruction conditions.

In Figures 2 and 3 we can see that trying to bias the teachers toward silence meaning good reduces performance on both boards. The disruption to learning seen in figure 3 is due to the fact that sometimes agents learn to take the food pellet in the center even if it has not taken the other food pellet. Doing this leads to immediate reward, but sometimes leads to the agent being trapped by the ghost with no means of escaping. Due to the stochastic nature of the learning algorithm, this "trick" is sometimes learnt quite late, and convergence sometimes does not happen until around iteration 750.

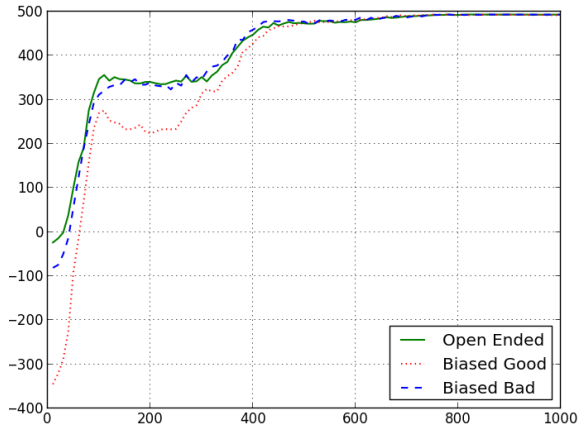

Figure 3: Here we can see the performance on video V3. We can see a plateau, or even a small dip in learning.

6.4 Interpretation of silence

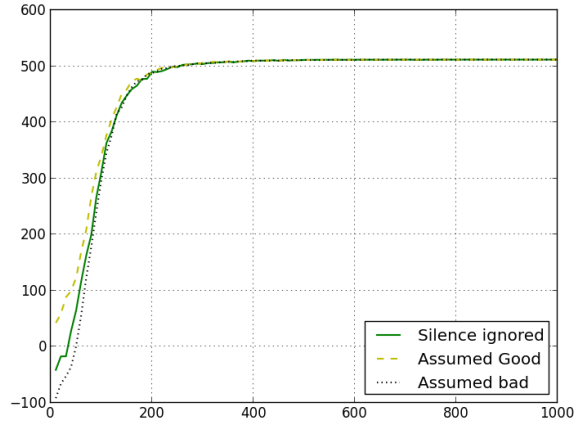

Figure 4: Here we can see the performance of the learning agent as a function of number of training iterations on the V2 video. The three lines correspond to the three different interpretations of silence.

Besides exploring different instruction conditions, we now explore what happens when silence is interpreted as either "good" or "bad". In Figures 4 and 5 we can see that assuming silence to mean good improves performance. The data set from the open ended instructions were used for all lines in these graphs, and the only variation is in how silence is interpreted.

We also ran the experiment with a data set that had a "good" evaluation for every state action pair, and one data set that had a "bad" evaluation for every state action pair. We found that positivity was significantly better than negativity, indicating that actions in the videos are better than random (a random policy almost always dies, while pac-man wins in some of the videos evaluated). The learner has access to all the evaluations, and in the beginning has not explored some of the evaluated states. Because the set of action pairs evaluated are better than random, it makes sense to prefer evalu-

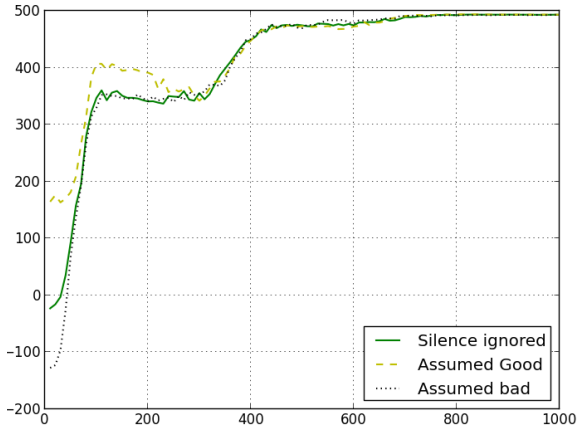

Figure 5: Here we can see the performance on video V3 with the different interpretations of silence. Assuming silence means good leads to significant performance improvements.

ated pairs that have been met with silence over unevaluated pairs. This might explain what happens when we compare the performance of the 4 combinations of biasing instructions for good/bad while assuming silence to mean good/bad in figures 6 and 7. Even when giving instructions biasing silence towards bad, it is still better to assume that silence means good.

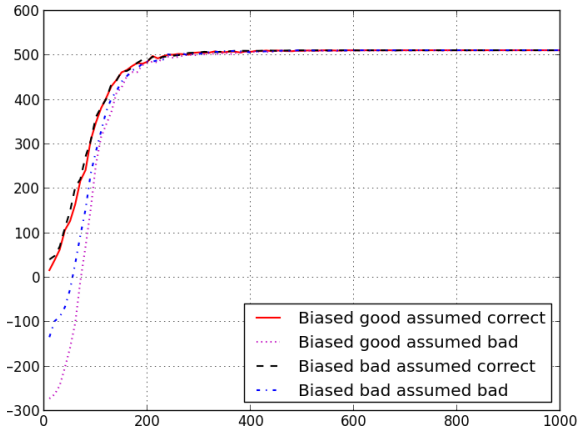

Figure 6: Here we can see what happens when we combine biasing instructions with different interpretations of the V2 video. Note that there is good performance when teachers are biased towards silence meaning bad, but the learner assumes silence to mean good, despite bias and assumptions being miss matched.

These findings about interpreting silence warrant future investigation, for which we have some hypotheses. It could be that people tend to mean silence as good, which would account for the results in Figures 4 and 5, and that this tendency is strong enough that when you try to bias them to give silence a particular meaning it only makes matters worse, as in Figures 2 and 3. So much so that the agent is better off assuming silence is positive regardless, as seen in Figures 6 and 7. However, to fully convince ourselves of this we would need to experiment on a variety of domains with different positive/negative biases. As indicated by our baseline experiment of an agent that assumed good or bad evaluation for every state, the domain of our experiment was such that assuming positive was more successful than negative, which contributes in part to these silence interpretation results.

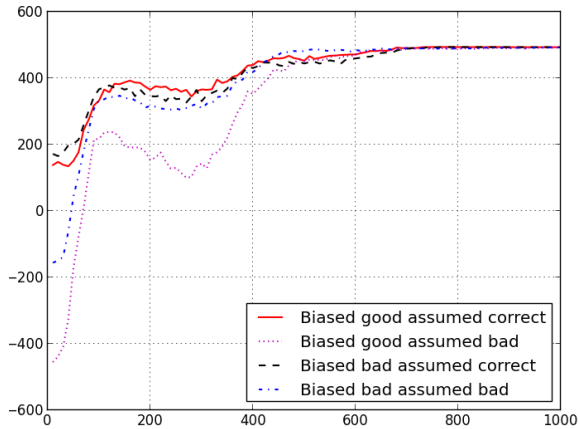

Figure 7: The results for the V3 video reproduce the qualitative findings of the V2 video in figure 6. Even when attempting to bias human teachers towards silence meaning bad, it is still more useful to assume that silence means good.

7 Conclusions

We have experimentally validated the policy shaping algorithm on human teachers, showing that the assumptions about humans that the algorithm is built on top of are reasonable. Not only do the results translate to real human teachers, our participants outperformed the simulated teacher because they were able to recognize many different ways of winning. We also showed that verbal instructions for when to give positive, negative, or no feedback have a significant impact on data quality.

Further, different interpretations of silence can increase or decrease performance. Future work will investigate a learner autonomously re-interpreting silence, and investigate what can be gained from tailoring the interpretations to individual teachers. It is possible that for some teachers silence should be interpreted as an ok action (but maybe not the optimal action), while for others it is better interpreted as a bad action (but perhaps not a critical mistake), and for yet other teachers, silence might not be useful at all. The setup described here was designed to generate data suitable for such a study.

References

[Breazeal *et al.*, 2005] Cynthia Breazeal, Cory D. Kidd, Andrea L. Thomaz, Guy Hoffman, and Matt Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robotics and Systems (IROS)*, pages 708–713, 2005.

[Breazeal *et al.*, 2009] Cynthia Breazeal, Jesse Gray, and Matt Berlin. An embodied cognition approach to mindreading skills for socially intelligent robots. *I. J. Robotic Res*, 28:656–680, 2009.

[Cakmak and Lopes, 2012] Maya Cakmak and Manuel Lopes. Algorithmic and human teaching of sequential decision tasks. In *AAAI Conference on Artificial Intelligence*, 2012.

[Cederborg and Oudeyer, 2014] Thomas Cederborg and Pierre-Yves Oudeyer. A social learning formalism for learners trying to figure out what a teacher wants them to do. *PALADYN Journal of Behavioral Robotics*, pages 64–99, 2014.

[Griffith *et al.*, 2013] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L. Isbell, and Andrea L. Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, 2013.

[Grizou *et al.*, 2014] Jonathan Grizou, Iñaki Iturrate, Luis Montesano, Pierre-Yves Oudeyer, and Manuel Lopes. Interactive learning from unlabeled instructions. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, 2014.

[Isbell *et al.*, 2006] Charles L. Isbell, Michael Kearns, Satinder Singh, Christian Shelton, Peter Stone, and Dave Kormann. Cobot in LambdaMOO: An adaptive social statistics agent. *Autonomous Agents and Multiagent Systems*, 13(3), November 2006.

[Knox *et al.*, 2012] Bradley .W Knox, Cynthia Breazeal, and Peter Stone. Learning from feedback on actions past and intended. In *In Proceedings of 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2012.

[Loftin *et al.*, 2014] Robert Loftin, Bei Peng, James MacGlashan, Michael L. Littman, Matthew E. Taylor, Jeff Huang, and David L. Roberts. Learning something from nothing: Leveraging implicit human feedback strategies. In *Proceedings of the Twenty-Third IEEE International Symposium on Robot and Human Communication (RO-MAN)*, 2014.

[Lopes *et al.*, 2011] Manuel Lopes, Thomas Cederborg, and Pierre-Yves Oudeyer. Simultaneous acquisition of task and feedback models. In *International Conference on Development and Learning (ICDL)*, 2011.

[Thomaz and Breazeal, 2007] Andrea L. Thomaz and Cynthia Breazeal. Asymmetric interpretations of positive and negative human feedback for a social learning agent. In *Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2007.

[Thomaz and Breazeal, 2008a] Andrea L. Thomaz and Cynthia Breazeal. Experiments in socially guided exploration: Lessons learned in building robots that learn with and without human teachers. *Connection Science, Special Issue on Social Learning in Embodied Agents*, pages 91–110, 2008.

[Thomaz and Breazeal, 2008b] Andrea L. Thomaz and Cynthia Breazeal. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence Journal*, pages 716–737, 2008.

[Watkins and Dayan, 1992] Christopher J. Watkins and Peter Dayan. Q learning: Technical note. *Machine Learning*, 8:279–292, 1992.

[Watkins, 1989] Christopher J. Watkins. Models of delayed reinforcement learning. In *PhD thesis, Psychology Department, Cambridge University*, 1989.