Interactive Reinforcement Learning from Imperfect Teachers

Taylor A. Kessler Faulkner
taylor.k.f@utexas.edu
Department of Computer Science
The University of Texas at Austin
Austin, Texas, USA

Andrea Thomaz
athomaz@ece.utexas.edu
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, Texas, USA

ABSTRACT

Robots can use information from people to improve learning speed or quality. However, people can have short attention spans and misunderstand tasks. Our work addresses these issues with algorithms for learning from *inattentive* teachers that take advantage of feedback when people are present, and an algorithm for learning from *inaccurate* teachers that estimates which state-action pairs receive incorrect feedback. These advances will enhance robots' ability to take advantage of imperfect feedback from human teachers.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; • **Computing methodologies** → **Reinforcement learning**; **Learning from critiques**.

KEYWORDS

Reinforcement Learning; Robot Learning; Human-Robot Interaction

ACM Reference Format:
Taylor A. Kessler Faulkner and Andrea Thomaz. 2021. Interactive Reinforcement Learning from Imperfect Teachers. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3434074.3446361

1 INTRODUCTION

Enabling people to teach robots in the wild will allow more robots to be deployed without expert supervision, potentially learning from *inattentive* or *inaccurate* human teachers. Interactive Reinforcement Learning (Interactive RL) has the ability to give robots two sources of information: an environmental reward function and feedback from human teachers. Robots can use both of these sources, balancing how much they learn from each one. Common methods in interactive RL have effectively incorporated feedback into an RL framework, but often assume that teachers are constantly available, give correct feedback, or randomly give bad feedback in any state [17]. However, people often need breaks [23, 25], and can have inaccurate task models leading to structured errors [19], both of which can lead to decreased performance if the robot expects full attention or correct feedback (Figure 1).

HRI '21 Companion, March 8–11, 2021, Boulder, CO, USA

ACM ISBN 978-1-4503-8290-8/21/03.
https://doi.org/10.1145/3434074.3446361

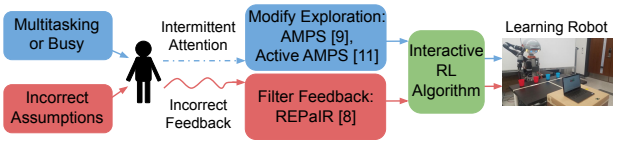

Figure 1: The approach for this research.

First, for *inattentive* teachers, we developed two algorithms, Attention-Modified Policy Shaping (AMPS) [9] and Active AMPS [11]. AMPS and Active AMPS capitalize on human attention by increasing exploration when the teacher is available and decreasing exploration otherwise, allowing the robot to learn quickly with less time from teachers. Second, our work on learning from *inaccurate* teachers enables robots to decide what teacher-provided information to trust, using additional sources of information such as the environmental reward function in interactive RL. We developed an algorithm, Revision Estimation from Partially Incorrect Resources (REPaIR), that translates incorrect feedback to usable feedback for the robot [8]. REPaIR takes advantage of feedback patterns, assuming that people will give incorrect feedback when confused about correct actions in specific areas of the state space. For these algorithms, we ran simulation experiments and human studies with a robot, using pushing, sorting, and picking tasks.

2 RELATED WORK

Interactive RL allows a Markov Decision Process (MDP) to take input from a human teacher [17]. This input can take many forms, such as binary or scalar values [10, 13, 24, 26], advice on future actions [14, 15, 18, 21], or action intervention [20]. Human feedback can also replace the environmental reward function [12]. Interactive RL algorithms often assume that the teacher is continuously paying attention, or giving consistently correct or incorrect feedback.

There has been prior research in active RL without present teachers [7]. However, this work is not based on feedback, but rather a potentially incomplete specification of an MDP by a researcher. There has also been prior work in active RL that uses human feedback, but does not enable teachers to take breaks [1, 2, 4–6].

Incorrect feedback has been addressed in interactive RL [10, 16, 22]. Some works assume feedback is randomly incorrect, with a static probability of incorrect feedback over the state space [10], or slowly decreased reliance on feedback over time [13]. Instead, our work assumes that there are patterns of incorrectness that appear given misunderstandings of tasks or robot capabilities, which gives the robot the ability to learn to predict when teachers are more likely to be incorrect. This enables a robot to take more advantage of correct feedback when such a pattern occurs, instead of distrusting all feedback equally. Other works assume that there are

multiple teachers [16]. Sridharan stores multiple policies from the environmental reward function and one policy from feedback [22], using comparisons to weight human feedback. Lin et al. moderate trust by keeping track of the current trust metric in a deep RL network, comparing teacher advice and learned Q-values [18]. These methods may discount good feedback at the beginning of learning, when the Q-values and initial policies are likely incorrect.

3 COMPLETED WORK

We developed two algorithms, Attention-Modified Policy Shaping (AMPS) [9] and Active Attention-Modified Policy Shaping (Active AMPS) [11] as new Interactive RL methods for learning from *inattentive* teachers, both built off of Policy Shaping (PS) [3, 10], an interactive RL algorithm that integrates binary positive and negative feedback from a teacher. AMPS changes exploration methods based on a teacher's presence. The detection of attention is not our focus, so the robot is told when attention is present. When a teacher is watching, the agent explores more to gather feedback on a wide variety of states, choosing between previously unseen state-action pairs and previously teacher-approved state-action pairs with equal probability (these probabilities could be experimented with in future work). When no teacher is watching, the agent explores less to increase the predictability of its actions, choosing previously approved state-action pairs when possible. We ran studies to compare AMPS and PS. The robot's task is pushing a cup through a grid to reach a goal location. In a simulated study with simulated teachers, we found that AMPS significantly outperforms PS with teachers available for 20 out of 100 learning episodes, achieving a 44% higher area under the learning curve. A human study with ten participants did not achieve significant results, potentially due to limited experiment length, but continued learning in simulation post-study showed that AMPS performed significantly better.

While AMPS allows teachers to take breaks, the burden is placed on the user to decide when to pay attention. Active AMPS enables a robot to ask for attention when it is unsure of any positive actions to take in a state, and spaces the requests for attention by at least t actions to allow teaching breaks. After these breaks, if the robot has not received any positive feedback on any action from its current state, the robot may ask for attention. This active criteria could be modified for future work. Active AMPS uses the same exploration criteria as AMPS during periods of attention and inattention. We test Active AMPS against PS and AMPS, both of which receive attention and feedback exactly every t actions. A human study with twelve participants did not show any significant difference between the performance of the three algorithms, potentially due to limited participants and experiment length, but participants gave significantly less feedback and had significantly more free time using Active AMPS than both AMPS and PS. In a simulated experiment with simulated teachers, we tested performance on a task sorting cups by color, and found that Active AMPS significantly outperforms PS by 27.1% and the AMPS algorithm by 11.0%.

We developed the REPaIR algorithm to address learning from *inaccurate* feedback [8]. This algorithm can translate incorrect information, from either human teachers or sensors, to usable information for the robot. REPaIR assumes that the robot has two sources of feedback: F, feedback that may be incorrect, and R, an environmental reward function which is assumed to be correct. REPaIR acts as a filter to Interactive RL algorithms to correct for incorrect feedback. Cumulative rewards ($R_e = \sum_{i=0}^{n} r_i$ for each reward r on action i in episode e) collected at the end of each episode are used as ground-truth information to update the filter. As the agent learns, it saves the highest achieved cumulative reward by each state-action pair (s_i, a_i). A trust t_i is assigned in the range $[0, 1]$ to the feedback f_i on (s_i, a_i). REPaIR determines whether to invert, keep, or discard feedback as follows, where t_{min} and t_{max} are threshold parameters. If $t_{(s_i,a_i,f_i)} \geq t_{max}$, REPaIR keeps the feedback: $f_i = f_i$. If $t_{(s_i,a_i,f_i)} \leq t_{min}$, REPaIR inverts the feedback: $f_i = -f_i$. Otherwise, REPaIR discards the feedback: $f_i = 0$. In experiments, we tested interactive RL algorithms [10, 13] both with and without the REPaIR filter. A robot experiment using sensor feedback and a gridworld cup picking task showed a slight average performance increase using REPaIR with Policy Shaping, although not a significant one. However, it did show that the robot was able to learn the task using REPaIR. In simulation, the agent learned in a grid to place a specific number of objects into two distinct bins. These experiments show that adding the REPaIR filter to an interactive RL algorithm enables expected robot performance to match or exceed expected performance of baseline interactive RL algorithms when robots have no prior knowledge of feedback correctness.

4 FUTURE WORK

Future work will focus on moving these algorithms, particularly REPaIR, to larger and potentially continuous state spaces. These algorithms have all been tested in smaller grid-world domains, and REPaIR specifically was built for state spaces where every visited state-action pair can easily be stored in memory with the corresponding maximum cumulative reward. For REPaIR, future work will focus on moving from recording observed performance to predicting future performance, by using machine learning algorithms to avoid requiring each state-action-feedback tuple to be stored. We also plan to test how much feedback these algorithms require, as the realistic limits on feedback for human teachers may limit how complex these tasks can be. Our Active AMPS algorithm works to decrease the amount of feedback necessary from a teacher, using feedback on 18.7 actions on average in simulation versus 203.1 for PS, so it is possible that combining Active AMPS with REPaIR may make learning from humans more feasible in a larger state space.

5 CONCLUSION

We propose that Interactive RL agents should change the way they learn based on human attention and errors, in order to take better advantage of human feedback. We present our completed work towards this goal, with the AMPS, Active AMPS, and REPaIR algorithms. Together, these algorithms enable a wider range of skilled human teachers to successfully teach robots skills using Interactive RL with less required attention.

6 ACKNOWLEDGEMENTS

This material is based upon work supported by the Office of Naval Research award numbers N000141612835 and N000141612785, National Science Foundation award numbers 1564080 and 1724157, and the NSF-GRFP under Grant No. DGE-1610403.

REFERENCES

[1] Riad Akrour, Marc Schoenauer, and Michèle Sebag. 2012. April: Active preference learning-based reinforcement learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 116–131.

[2] Maya Cakmak, Crystal Chao, and Andrea L Thomaz. 2010. Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development* 2, 2 (2010), 108–118.

[3] Thomas Cederborg, Ishaan Grover, Charles L Isbell, and Andrea L Thomaz. 2015. Policy shaping with human teachers. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[4] Sonia Chernova and Manuela Veloso. 2009. Interactive policy learning through confidence-based autonomy. *Journal of Artificial Intelligence Research* 34 (2009), 1–25.

[5] Jeffery Allen Clouse. 1996. *On integrating apprentice learning and reinforcement learning*. University of Massachusetts Amherst.

[6] Finale Doshi-Velez, Joelle Pineau, and Nicholas Roy. 2012. Reinforcement learning with limited reinforcement: Using Bayes risk for active learning in POMDPs. *Artificial Intelligence* 187 (2012), 115–132.

[7] Arkady Epshteyn, Adam Vogel, and Gerald DeJong. 2008. Active reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*. ACM, 296–303.

[8] Taylor A Kessler Faulkner, Elaine Schaertl Short, and Andrea L Thomaz. 2020. Interactive Reinforcement Learning with Inaccurate Feedback. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 7498–7504.

[9] Taylor Kessler Faulkner, Elaine Schaertl Short, and Andrea Lockerd Thomaz. 2018. Policy Shaping with Supervisory Attention Driven Exploration. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 842–847.

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

[11] Taylor Kessler Faulkner, Reymundo A Gutierrez, Elaine Schaertl Short, Guy Hoffman, and Andrea L Thomaz. 2019. Active attention-modified policy shaping: socially interactive agents track. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 728–736.

[12] W Bradley Knox and Peter Stone. 2008. Tamer: Training an agent manually via evaluative reinforcement. In *Development and Learning, 2008. ICDL 2008. 7th IEEE International Conference on*. IEEE, 292–297.

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

[14] Samantha Krening and Karen M Feigh. 2018. Interaction algorithm effect on human experience with reinforcement learning. *ACM Transactions on Human-Robot Interaction (THRI)* 7, 2 (2018), 16.

[15] Samantha Krening and Karen M Feigh. 2019. Newtonian Action Advice: Integrating Human Verbal Instruction with Reinforcement Learning. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 720–727.

[16] Andrey Kurenkov, Ajay Mandlekar, Roberto Martin-Martin, Silvio Savarese, and Animesh Garg. 2019. AC-Teach: A Bayesian Actor-Critic Method for Policy Learning with an Ensemble of Suboptimal Teachers. In *Conference on Robot Learning (CoRL)*.

[17] Guangliang Li, Randy Gomez, Keisuke Nakamura, and Bo He. 2019. Human-Centered Reinforcement Learning: A Survey. *IEEE Transactions on Human-Machine Systems* (2019).

[18] Zhiyu Lin, Brent Harrison, Aaron Keech, and Mark O Riedl. 2017. Explore, exploit or listen: Combining human feedback and policy model to speed up deep reinforcement learning in 3D worlds. *arXiv preprint arXiv:1709.03969* (2017).

[19] Jens Rasmussen. 1982. Human errors. A taxonomy for describing human malfunction in industrial installations. *Journal of occupational accidents* 4, 2-4 (1982), 311–333.

[20] William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. 2018. Trial without error: Towards safe reinforcement learning via human intervention. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2067–2069.

[21] Emmanuel Senft, Séverin Lemaignan, Paul E Baxter, Tony Belpaeme, et al. 2016. Sparc: an efficient way to combine reinforcement learning and supervised autonomy. In *Future of Interactive Learning Machines Workshop at NIPS'16* (Barcelona, Spain).

[22] Mohan Sridharan. 2011. Augmented reinforcement learning for interaction with non-expert humans in agent domains. In *2011 10th International Conference on Machine Learning and Applications and Workshops*, Vol. 1. IEEE, 424–429.

[23] Michael B Steinborn and Lynn Huestegge. 2016. A walk down the lane gives wings to your brain. Restorative benefits of rest breaks on cognition and self-control. *Applied Cognitive Psychology* 30, 5 (2016), 795–805.

[24] Kaushik Subramanian, Charles L Isbell Jr, and Andrea L Thomaz. 2016. Exploration from demonstration for interactive reinforcement learning. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 447–456.

[25] Philip Tucker. 2003. The impact of rest breaks upon accident risk, fatigue and performance: a review. *Work & Stress* 17, 2 (2003), 123–137.

[26] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. 2018. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Thirty-Second AAAI Conference on Artificial Intelligence*.