When Oracles Go Wrong: Using Preferences as a Means to Explore

Isaac S. Sheidlower
Tufts University
Isaac.Sheidlower@tufts.edu

Elaine Schaertl Short
Tufts University
elaine.g.short@usc.edu

ABSTRACT

When a robot is deployed to learn a new task in a "real-word" environment, there may be multiple teachers and therefore multiple sources of feedback. Furthermore, there may be multiple optimal solutions for a given task and teachers may have preferences among those various solutions. We present an Interactive Reinforcement Learning (I-RL) algorithm, Multi-Teacher Activated Policy Shaping (M-TAPS), which addresses the problem of learning from multiple teachers and leverages differences between them as a means to explore the environment. We show that this algorithm can significantly increase an agent's robustness to the environment and quickly adopt to a teacher's preferences. Finally, we present a formal model for comparing human teachers and constructed oracle teachers and the way that they provide feedback to a robot.

KEYWORDS

human-robot interaction; interactive reinforcement learning

ACM Reference Format:

Isaac S. Sheidlower and Elaine Schaertl Short. 2021. When Oracles Go Wrong: Using Preferences as a Means to Explore. 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, 5 pages. https://doi.org/10.1145/3434074.3447189

1 INTRODUCTION

Current I-RL models will not effectively and quickly learn from multiple teachers, if those teachers have differing preferences that they want to see reflected in the robot's behavior. Furthermore, much of the current research in Interactive Reinforcement Learning (I-RL) adopts two flawed assumptions: first, that an agent that has found only one good solution to a task is satisfactory; and second, that using a "perfect" oracle that guides robot behavior using feedback that exactly aligns with the optimal robot behavior will result in the most effective robot learning. These assumptions are flawed because different users may prefer different solutions to the one the agent is using and, Under current I-RL models, a perfect oracle teacher could cause the agent to get "stuck" in one optimum as a solution. For example, consider a robot deployed in a hospital which may be receiving feedback from many different patients and hospital staff. In that case, the robot should not only quickly respond to the needs of its current user, but to do so even if it means deviating from an already learned behavior. Thus, in situations with multiple teachers, researchers can design agents that are robust to complex, human-populated environments by avoiding these assumptions.

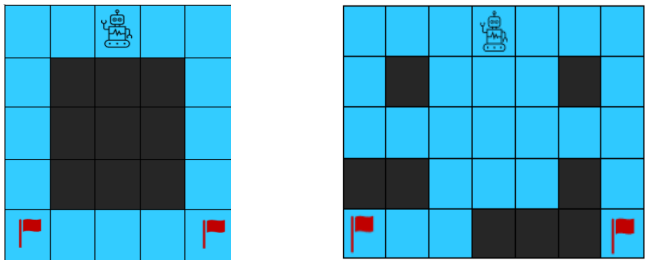
We introduce an algorithm with the two key properties of quick adaptation to teachers and long-term leveraging of feedback from good teachers. The agent will quickly adapt to a new teachers feedback/preferences, but will ultimately disregard that feedback if it's adversarial. In addition, when multiple teachers are providing the agent feedback, the agent will weight such feedback greater from one teacher if it agrees with the environmental reward more so than that of other teachers. M-TAPS (Multi-Teacher Activated Policy Shaping) is an algorithm that uses the differences between human teachers on a given task to gain robustness to its environment. In order to characterize the differences between teachers, we introduce a novel and significantly more expressive and flexible paradigm to model both human teachers and oracles.We show that in simulation, M-TAPS significantly increases the environmental robustness of an I-RL agent when being taught by biased, imperfect, oracles than when being taught by a perfect oracle. This advances the state of I-RL because, by exploiting more human-like feedback, it exposes a limitation of perfect oracles.

2 BACKGROUND

It has been demonstrated that human teachers can outperform oracles in certain tasks [24].Using this high-quality information available from a human teacher, reinforcement Learning (RL) agents can learn faster using input from such human teachers, through a process called interactive reinforcement learning (I-RL). However, for an agent to use feedback from a teacher, the agent must make assumptions both about how to interpret that feedback and how to use the information. Prior works in I-RL, for instance, have treated feedback as a direct reward signal [11] or as advice about a policy [6]. These have greatly increased the learning speed of agents. However, these models generally assume feedback from one fairly consistent teacher and/or one optimal solution to each task and thus may not preform well when that is not the case. Given that research has also shown that human teachers *do* have preferences among different, but still near optimal, solutions [14, 17, 4], there is a need to modify prior I-RL approaches to handle for this phenomenon when there are multiple teachers. If the differences among teachers can be adequately handled by a learning algorithm, the feedback from multiple teachers could both expedite learning and lead to policies that are more robust to a potentially changing real-world environment.

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

ACM ISBN 978-1-4503-8290-8/21/03…$15.00
https://doi.org/10.1145/3434074.3447189



Environment A **Environment B**

Figure 1: Grid worlds with multiple optimal paths to different goal states (flags). The agent receives a reward of 1 for reaching a goal, -1 for falling into a hole (black), and -.5 for every action taken. The agent moves deterministically.

A teacher can help an agent explore if that agent can quickly adapt to the teacher's feedback. However, given that feedback is dependent on the agent's current policy, in the multiple teacher setting the agent needs to quickly adapt to each new teacher. An agent that fails to do so may not only impact that teacher's feedback and preferences but also cause the teacher to become frustrated [15]. Such a failure could adversely impact performance of the agent [3]. We propose that such adaptation can be achieved if the agent can distinguish between multiple teachers and make use of such information at the algorithmic level.

Other RL domains with human teachers have already shown that drawing from a human's sub-optimal policies can be an effective way to learn a task. Inverse Reinforcement Learning and Behavior Cloning use demonstrations rather than feedback to learn a teacher's underlying reward function or clone the demonstrated behavior [1, 20, 16, 9]. Preference Learning, on the other hand, uses an annotating method wherein a teacher ranks potential action trajectories then the agent fits a policy that best matches the teacher's preferences [23, 2]. In Inverse Reinforcement Learning in particular, even sub-optimal demonstrations can lead to decent agent behavior [8, 18]. Research that either combines these two techniques or explicitly makes use of sub-optimal teachers has been successful in outperforming agents with one "optimal" teacher [7, 13, 22]. Also, recent advancements in deep learning have greatly reduced the amount of time needed to teach agents across these domains [21, 3, 10, 25]. Our goal is to translate concepts exploiting sub-optimal teaching into the I-RL domain. I-RL puts the least amount of burden on human teachers, as they neither have to provide demonstrations nor necessarily be constantly attentive [5], yet still naturally allows for the agent to learn from its environment and not just a human teacher.

3 DISCRETE MULTI-TEACHER ACTIVATED POLICY SHAPING

Using Griffith, Subramanian, Scholz, Isbell, and Thomaz's Policy Shaping (PS) algorithm, we investigated the performance and properties of an I-RL agent with both perfect and biased oracle teachers on discrete grid worlds with multiple optimal solutions (i.e. equal environmental reward; Figure 1). Since prior work has shown that human teachers do have preferences among paths even in simple grid world domains [17], we constructed 3 oracles. We implemented a traditional "perfect" oracle which provides the agent with positive feedback with a certain probability if the agent's action was optimal, and negative otherwise. We also implemented two "biased" oracles, one oracle who prefers paths to the right and is adverse to paths to the left, and vice versa for the other oracle. We then compare the perfect oracle against the two biased oracles. In the "naive" baseline and the perfect oracle baseline, the PS algorithm is left unchanged.

Algorithm 1: Discrete Multi-Teacher Activated Policy Shaping (DM-TAPS)

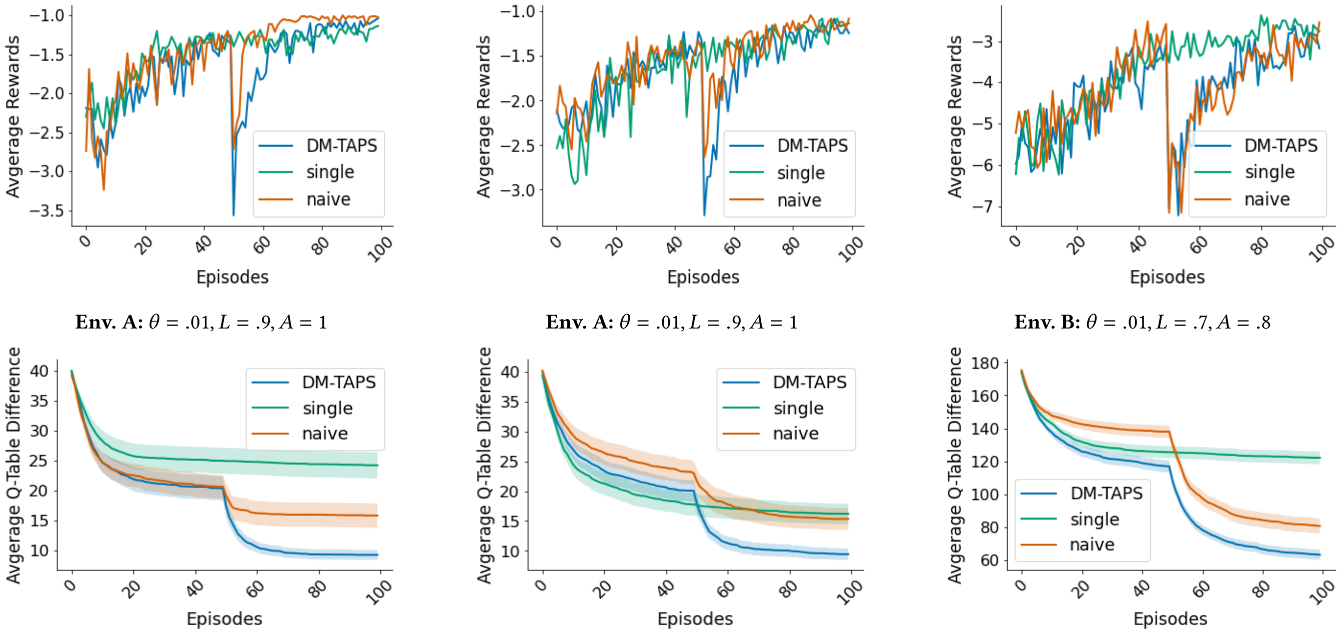
initialize Q_{global}
while *True* **do**
  if *New Teacher* **then**
    initialize Q_n, F_n
    while *Teacher n is present* **do**
      PolicyShaping(Q_n,F_n)
    if $n == 1$ **then**
      $Q_{global} = Q_n$, $F_{global} = F_n$
    else
      UpdateQ(Q_{global},Q_n)
      UpdateF(F_{global},F_n)

Given that the agent *is aware* the teacher has changed, we test our algorithm Discrete Multi-Teacher Activated Policy Shaping (DM-TAPS). The general DM-TAPS algorithm is presented in Algorithm 1. DM-TAPS, introduces two novelties to the PS algorithm. One is the initialization of a new Q-table (Q_n) and a new feedback table (F_n) for each teacher that interacts with the agent. These tables are distinguished from a "global Q-table" and a "global feedback table" which the agent uses while not attended. The other is Update functions that dictate how the individual teacher's Q-table and feedback table should be integrated into the global tables when that teacher leaves. In the version of DM-TAPS used in Figure 2, the Q-table for the nth teacher, Q_n, **is the same as** Q_{global} however their feedback table, F_n, is newly initialized on arrival. In the case where the two update functions are the identity function and Q_n and F_n are initialized to the global versions of the Q table and feedback table, respectively, the algorithm reduces to regular Policy Shaping. Finally, the UpdateF function can also vary. An example is shown in Algorithm 3 where w_n is a weight that depends on how good the teacher was.

Algorithm 2: UpdateQ Function Example

for *s, a in* Q_{global} **do**
  if $Q_{global}(s,a) < 0$ *and* $Q_n(s,a) < 0$ **then**
    $Q_{global}(s,a) = \min(Q_{global}(s,a), Q_n(s,a))$
  else
    $Q_{global}(s,a) = \max(Q_{global}(s,a), Q_n(s,a))$

We ran a series of experiments to determine both how quickly a PS agent's Q-table converges to the true Q-table and how the rewards that agent receives vary under the guidance of different oracles. The environments used are presented in Figure 1. Between conditions, we varied θ, the constant exploration rate parameter in PS, the oracle parameter L which dictates how likely the oracles are to provide feedback, and the oracle parameter A, which is how likely the oracle's feedback is to be accurate/consistent. Furthermore, the biased oracles used in Environment B, only have a



Env. A: $\theta = .01, L = .9, A = 1$ **Env. A:** $\theta = .01, L = .9, A = 1$ **Env. B:** $\theta = .01, L = .7, A = .8$

Env. A: $\theta = .01, L = .9, A = 1$ **Env. A:** $\theta = .1, L = .9, A = 1$ **Env. B:** $\theta = .01, L = .7, A = .8$

Figure 2: Top row: Average reward over 40 trials. Bottom row: Average Q-table difference (see Section 3) over 40 trials with the shaded areas being the standard deviation. θ is the constant exploration rate used in Policy Shaping. L dictates how likely the oracles are to provide feedback while A dictates how often that feedback is accurate. The Q-learning and PS parameters where set as follows: lr=0.7; max moves per episode = 40; C=1, which represents the confidence the agent has in their teacher.

directional preference in the first two rows of the grid-world because, unlike in Environment A, there are multiple valid paths to either direction. During each trial we calculated the "Q-table difference", mean squared error between the agent's Q-table and the true Q-table as a measure of the agent's robustness to the environment, and the cumulative reward (Figure 2).

Algorithm 3: UPDATEF Function Example

for s, a in F_{global} **do**
  $F_{global} = F_{global}(s, a) + w_n * F_n(s, a)$

We present four key results. First, the perfect oracle will do better in terms of reward than the two biased oracles. This is largely due to the agent having to "readjust" to a new teacher half-way through the trial with the biased oracles. Second, as the agent explores more on its own (as θ approaches 1), the differences between teachers decreases (though DM-TAPS will always be more robust and the perfect oracle will always cumulatively lead to more reward). Third, DM-TAPS becomes significantly more robust to the environment than the perfect oracle and does so much more reliably (the lower the exploration rate, the more likely a perfect oracle will have the agent stuck in an optima). Finally, there is a trade off between environmental robustness and a temporary sacrifice in reward, that occurs when teachers change, which becomes greater as environments become more complex. Other notable results are that the naive approach suffers less in terms of reward when teachers switch because the second teacher must provide enough feedback to the agent for it to behave according to the preference, and, besides θ, these results seem very robust to parameter changes.

4 A PARADIGM FOR TEACHER MODELING AND ORACLE CONSTRUCTION

Here we present a general way to model both the human and oracle teachers providing feedback to an agent. With this model we hope I-RL researchers can: construct more human-like oracles for testing their algorithms in simulation; provide a means to compare and contrast different teachers and/or different oracles to test how their algorithm performs with different kinds of teachers; and create models of *how* humans may provide feedback for a given task so that model can be used as a prior when learning from a new human teacher. Currently, most oracles differ only in how often they provide feedback (likelihood) and the probability that they provide feedback to the contrary of the environment's reward function (consistency/accuracy), which does not necessarily capture how humans give feedback [12, 24].

For any given state s, let there be a function AE that represents the favorability or action evaluation, from the teacher's perspective, of taking an action x_i in state s. Let there also be a feedback function F of AE, a threshold function T, and some noise ϵ. Together, these functions determine whether or not the teacher will provide feedback and the value of that feedback. This formulation can be written as a function composition of the form:

$$Feedback = F(AE(s), T(s), \epsilon) \tag{1}$$

In the case of binary feedback, F amounts to a discretizing piecewise function. An example is shown below:

$$Feedback = \begin{cases} X_{i,p_i} \; if AE(s,x_i) - T(s) \geq 0 \\ -W_{i,p_i} \; if AE(s,x_i) - T(s) \leq 0 \end{cases} \quad (2)$$

X_i and $-W_i$ are Bernoulli random variables, with p = p_i, representing the value of feedback after taking action x_i. One may assume that for a non-random teacher, p_i is a function of $AE(s, x_i)$. In this example, the noise ϵ is implicit due to the probabilistic nature of X_i and W_i. The two reasons for distinguishing between so called action evaluations and the probability of providing certain feedback are because teachers may provide feedback more or less often as they become more aware of its impact [19] (while the probability a teacher provides feedback may change, an action may be equally favorable), and because one may want to use scaled feedback in which case the scale can be a function of an action evaluation (a real number) more intuitively than a probability (a number between 0 and 1). These functions may both change over time and may not have any discernible relationship with the environmental reward (e.g. a random teacher). Thus, this model allows for a great degree of flexibility for modeling teachers, without groundlessly assuming how teachers will provide feedback for a given task.

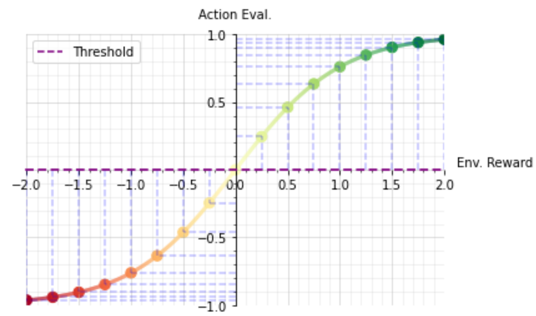
With this model of teachers in mind, we can define three ways a different teacher providing feedback may differ in both the discrete and continuous case: teachers may have a different ranking of actions, i.e. the function AE may differ; they may have a different threshold/T(s) for when they provide positive or negative feedback; they may differ in how likely they are to provide feedback for certain actions (i.e. the relationship of $AE(s, x_i)$ and p_i). Notions of good and bad teachers may be described in terms of these differences and the relationship of these functions to the environmental reward function. Good teachers rank actions that lead to high environmental reward more highly and whose threshold for providing positive feedback is near the boundary between actions that yield positive environmental reward and those that yield negative environmental reward (see Figure 3 b). Note that with this paradigm we can now model certain phenomenon in teachers such as diminishing returns.

5 DISCUSSION AND CONCLUSIONS

We propose an important idea for I-RL and HRI researchers to consider, namely that just the knowledge that an I-RL agent is interacting with a different teacher can be leveraged in various ways to improve agent performance and/or environmental robustness. This aligns with the common sense notion that different people may offer different but valid strategies. In discrete environments, this equates to using different teacher's preferences as a means to guide exploration. For robots deployed in socially interactive contexts, environmental robustness and the ability to adapt to different teachers may be as, if not more, important than the environmental reward for a given task. In particular, in real-world scenarios, human teachers may value a robot that is both personalized and effective. Thus, we propose a general purpose algorithm based on Policy Shaping that in discrete state spaces can gain environmental robustness leading to favorable behavior in the long term, while also readily adapting to new teachers. Importantly, DM-TAPS has a high degree

Good teacher in state x	Bad teacher in state x	Perfect oracle in state x
Action 1	Action 5	Action 2
Action 2	Action 6	Threshold
Action 3	Action 3	Action 1
Threshold	Threshold	Action 3
Action 4	Action 4	Action 4
Action 5	Action 3	Action 5
Action 6	Action 1	Action 6

(a) Discrete Representation



(b) F(AE, T, ϵ) with Binary Feedback

Figure 3: (a) Example of teachers and their feedback models. The more green an action, the more environmentally rewarding. The more red an action, the more environmentally punishing. The threshold for human teachers is yellow meaning that it is a soft threshold. For the oracle, it is black meaning it is a hard threshold. (b) Showing the relationship between AE and the environmental MDP reward for a "good" teacher. The more green an action (represented by dots) is, the more likely positive feedback is given while the more red an action, the more likely negative feedback is given.

of customization, so researchers can tailor it to a certain task or to certain teachers. In the near future, we plan on preforming a study with human participants to investigate the implications of multiple teachers in a continuous state-space using general Multi-Teacher Activated Policy Shaping (M-TAPS), a deep I-RL model.

We also attempted to fill a void in the I-RL literature, namely that there is not a formal method for modeling human teachers mathematically nor a way to construct more varied and unique oracles for simulation. This model opens future areas of research, including: cognitive science/psychology research that investigates how humans provide feedback in various scenarios and then fitting their behavior to the paradigm described (through techniques such as ML); create priors about how teachers may provide feedback in a given task such that they can be more quickly modeled and understood from the learning agents perspective; and defining difference metrics between teachers such that we can better understand the differences between naive and expert teachers in terms of how they provide feedback in a given task. This model, and models like it, can further bridge the gap between computer science and cognitive science/psychology in the field of HRI.

REFERENCES

[1] Pieter Abbeel and Andrew Y. Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Twenty-first international conference on Machine learning - ICML '04*. event-place: Banff, Alberta, Canada. ACM Press, 1. DOI: 10.1145/1015330.1015430. Retrieved 09/13/2020 from http://portal.acm.org/citation.cfm?doid=1015330.1015430.

[2] Riad Akrour, Marc Schoenauer, and Michele Sebag. 2011. Preference-Based Policy Learning. en. In *Machine Learning and Knowledge Discovery in Databases*. Volume 6911. Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors. Series Title: Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, 12–27. ISBN: 978-3-642-23779-9 978-3-642-23780-5. DOI: 10.1007/978-3-642-23780-5_11. Retrieved 12/01/2020 from http://link.springer.com/10.1007/978-3-642-23780-5_11.

[3] Dilip Arumugam, Jun Ki Lee, Sophie Saskin, and Michael L. Littman. 2019. Deep Reinforcement Learning from Policy-Dependent Human Feedback. *arXiv:1902.04257 [cs, stat]*, (February 2019). arXiv: 1902.04257. Retrieved 10/05/2020 from http://arxiv.org/abs/1902.04257.

[4] Andreea Bobu, Dexter R. R. Scobee, Jaime F. Fisac, S. Shankar Sastry, and Anca D. Dragan. 2020. LESS is More: Rethinking Probabilistic Models of Human Behavior. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. event-place: Cambridge United Kingdom. ACM, (March 2020), 429–437. ISBN: 978-1-4503-6746-2. DOI: 10.1145/3319502.3374811. Retrieved 09/13/2020 from https://dl.acm.org/doi/10.1145/3319502.3374811.

[5] Taylor Kessler Faulkner, Reymundo A Gutierrez, Elaine Schaertl Short, Guy Hoffman, and Andrea L Thomaz. 2019. Active Attention-Modified Policy Shaping, 9.

[6] 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 26*. C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors. Curran Associates, Inc., 2625–2633. Retrieved 10/11/2020 from http://papers.nips.cc/paper/5187-policy-shaping-integrating-human-feedback-with-reinforcement-learning.pdf.

[7] Daniel H. Grollman and Aude G. Billard. 2012. Robot Learning from Failed Demonstrations. en. *International Journal of Social Robotics*, 4, 4, (November 2012), 331–342. ISSN: 1875-4805. DOI: 10.1007/s12369-012-0161-z. Retrieved 10/19/2020 from https://doi.org/10.1007/s12369-012-0161-z.

[8] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative Inverse Reinforcement Learning. In *Advances in Neural Information Processing Systems 29*. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors. Curran Associates, Inc., 3909–3917. Retrieved 09/13/2020 from http://papers.nips.cc/paper/6420-cooperative-inverse-reinforcement-learning.pdf.

[9] Jonathan Ho and Stefano Ermon. 2016. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems 29*. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors. Curran Associates, Inc., 4565–4573. Retrieved 10/19/2020 from http://papers.nips.cc/paper/6391-generative-adversarial-imitation-learning.pdf.

[10] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. [n. d.] Reward learning from human preferences and demonstrations in Atari. en, 13.

[11] W. Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: the TAMER framework. en. In *Proceedings of the fifth international conference on Knowledge capture - K-CAP '09*. ACM Press, Redondo Beach, California, USA, 9. ISBN: 978-1-60558-658-8. DOI: 10.1145/1597735.1597738. Retrieved 09/13/2020 from http://portal.acm.org/citation.cfm?doid=1597735.1597738.

[12] Samantha Krening and Karen M. Feigh. 2018. Interaction Algorithm Effect on Human Experience with Reinforcement Learning. en. *ACM Transactions on Human-Robot Interaction*, 7, 2, (October 2018), 1–22. ISSN: 2573-9522, 2573-9522. DOI: 10.1145/3277904. Retrieved 12/04/2020 from https://dl.acm.org/doi/10.1145/3277904.

[13] 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. *arXiv:1909.04121 [cs, stat]*, (December 2019). arXiv: 1909.04121. Retrieved 10/19/2020 from http://arxiv.org/abs/1909.04121.

[14] Minae Kwon, Erdem Biyik, Aditi Talati, Karan Bhasin, Dylan P. Losey, and Dorsa Sadigh. 2020. When Humans Aren't Optimal: Robots that Collaborate with Risk-Aware Humans. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, (March 2020), 43–52. arXiv: 2001.04377. DOI: 10.1145/3319502.3374832. Retrieved 09/13/2020 from http://arxiv.org/abs/2001.04377.

[15] James MacGlashan, Mark K. Ho, Robert Loftin, Bei Peng, David Roberts, Matthew E. Taylor, and Michael L. Littman. 2017. Interactive Learning from Policy-Dependent Human Feedback. *arXiv:1701.06049 [cs]*, (January 2017). arXiv: 1701.06049. Retrieved 10/05/2020 from http://arxiv.org/abs/1701.06049.

[16] Stéphane Ross, Geoffrey J Gordon, and J Andrew Bagnell. [n. d.] A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. en, 9.

[17] Himanshu Sahni, Brent Harrison, Thomas Cederborg, Charles Isbell, Kaushik Subramanian, and Andrea Thomaz. [n. d.] Policy Shaping in Domains with Multiple Optimal Policies (Extended Abstract), 2.

[18] Matthew E Taylor, Halit Bener Suay, and Sonia Chernova. [n. d.] Integrating reinforcement learning with human demonstrations of varying ability. en, 8.

[19] Andrea L Thomaz. [n. d.] Reinforcement Learning with Human Teachers: Evidence of Feedback and Guidance with Implications for Learning Performance, 6.

[20] Faraz Torabi, Garrett Warnell, and Peter Stone. 2018. Behavioral Cloning from Observation. *arXiv:1805.01954 [cs]*, (May 2018). arXiv: 1805.01954. Retrieved 10/19/2020 from http://arxiv.org/abs/1805.01954.

[21] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. 2018. Deep TAMER: Interactive Agent Shaping in High-Dimensional State Spaces. *arXiv:1709.10163 [cs]*, (January 2018). arXiv: 1709.10163. Retrieved 09/29/2020 from http://arxiv.org/abs/1709.10163.

[22] Nils Wilde, Alexandru Blidaru, Stephen L Smith, and Dana Kulić. 2020. Improving user specifications for robot behavior through active preference learning: Framework and evaluation. en. *The International Journal of Robotics Research*, 39, 6, (May 2020), 651–667. ISSN: 0278-3649, 1741-3176. DOI: 10.1177/0278364920910802. Retrieved 12/01/2020 from http://journals.sagepub.com/doi/10.1177/0278364920910802.

[23] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. [n. d.] A Survey of Preference-Based Reinforcement Learning Methods. en, 46.

[24] Chao Yu, Tianpei Yang, Wenxuan Zhu, Dongxu wang, and Guangliang Li. 2018. Learning Shaping Strategies in Human-in-the-loop Interactive Reinforcement Learning. *arXiv:1811.04272 [cs]*, (November 2018). arXiv: 1811.04272. Retrieved 12/04/2020 from http://arxiv.org/abs/1811.04272.

[25] Huixin Zhan, Feng Tao, and Yongcan Cao. 2020. Human-guided Robot Behavior Learning: A GAN-assisted Preference-based Reinforcement Learning Approach. *arXiv:2010.07467 [cs]*, (October 2020). arXiv: 2010.07467. Retrieved 12/01/2020 from http://arxiv.org/abs/2010.07467.