Learning Reward Functions from a Combination of Demonstration and Evaluative Feedback

Eric Hsiung
Brown University
Providence, RI
eric_hsiung@brown.edu

Eric Rosen
Brown University
Providence, RI
eric_rosen@brown.edu

Vivienne Bihe Chi
Brown University
Providence, RI
vivienne_chi@brown.edu

Bertram F. Malle
Brown University
Providence, RI
bfmalle@brown.edu

***Abstract*—As robots become more prevalent in society, they will need to learn to act appropriately under diverse human teaching styles. We present a human-centered approach for teaching robots reward functions by using a mixture of teaching strategies when communicating action appropriateness and goal success. Our method incorporates two teaching strategies for learning: explicit action instruction and evaluative, scalar-based feedback. We demonstrate that a robot instantiating our method can learn from humans who use both kinds of strategies to train the robot in a complex navigation task that includes norm-like constraints.**

***Index Terms*—Human-robot interaction, interactive reinforcement learning, feedback strategies, learning systems, norms**

I. Introduction

As robots take on more socially significant roles—as caretakers, assistants, and collaborators—we must develop algorithms allowing robots to learn appropriate behavior in society. People teach robots in ways most intuitive to them, which may not align with current learning algorithms' expected inputs. Thus, robots must be able to learn from diverse forms of human feedback, such as instruction, demonstration, praise, and criticism, and they must be able to respond to teachers with different styles and preferences of teaching.

Recent work on learning reward functions (Jeon et al. [1], Biyik et al. [2]) has emphasized the importance of learning from diverse forms of feedback. These authors model human feedback choices as a Boltzmann rational process, and they employ Bayesian inverse reinforcement learning (Ramachandran and Amir [3]) to incorporate both the *choice* of feedback and the feedback *value* to align robot behavior with what the human desires. However, teachers have different personalities, teaching styles, and goals when teaching students (Clegg et al. [4], Kline [5], Grasha [6]), and they may not always act in a Boltzmann rational manner.

We are taking first steps toward algorithms that specifically integrate two frequent teaching strategies in social settings: instructing a learner on the correct actions in a given context and providing evaluative feedback when observing the learner's actions. Our preliminary findings indicate that an intelligent agent can successfully learn from combinations of instruction and evaluative feedback and that it can also learn from human feedback that considers both the long-term and immediate consequences of the agent's current behavior.

II. Related Work

Previous work on human-robot teaching often assumes that human teaching takes one of two forms: either action-based instruction or scalar-based evaluation. Examples of action-based teaching are human demonstrations, which—in reinforcement-learning (RL) frameworks—are example sequences of state-action pairs. Scalar-based teaching assigns some scalar value to the observed quality of the agent's action. Several methods exist for learning reward functions or policies from these two forms of feedback.

Inverse reinforcement learning (IRL) algorithms aim to learn a reward model inducing an optimal policy consistent with a set of expert (typically human) demonstrations. Abbeel and Ng [7] aimed to find a policy closely matching the policy of human experts by using linear programming to learn a reward model that explains the expert policy. Ambiguity in potential reward models was observed by Ng and Russell [8], indicating that the reward model inducing a given policy is not unique. To determine if any two inferred reward functions are equivalent, Gleave et al. [9] proposed the Equivalent-Policy Invariant Comparison (EPIC) that canonicalizes the reward functions and computes a pseudometric. If two reward functions are equivalent under EPIC, then they induce the same policy and therefore the same behavior.

The aforementioned methods require no interaction between the learning agent and the human teacher because the agent is trained on a pre-collected dataset of human demonstrations. However, collecting demonstration data can be expensive. Brown and Niekum [10] investigated the minimal number of maximally informative demonstrations to use as the demonstration dataset for IRL and aimed to leverage this information in a Bayesian Information-Optimal IRL approach [10]. Other work has focused on *interactive* IRL to learn from incrementally obtained demonstrations. Arora et al. [11] proposed an online, incremental IRL framework. Kamalaruban et al. [12] considered teachers providing demonstrations conditioned on observing the agent's behavior and teaching under full knowledge of the agent's policy and parameters, or only given noisy samples of the agent's policy.

Additionally, previous work has shown that humans are able to teach an agent desired behaviors by giving scalar-based evaluative feedback. In *interactive RL*, a human teacher

978-1-6654-0731-1/22/$31.00 ©2022 IEEE

provides feedback dynamically as part of the ongoing learning process. Knox proposed TAMER [13], which assumes humans provide time-delayed *reward-based* feedback that can be treated as scalar values and can be used to estimate the human's internal reward model. MacGlashan et al. proposed COACH [14, 15], offering empirical evidence that human participants can provide *policy-dependent*, *advantage-based* scalar feedback to train agents. Policy-dependent feedback is based on an evaluation of the perceived current policy of the agent. Advantage-based feedback intuitively represents how much more valuable taking an action in a certain state is compared to the average value of all actions that could be taken from that state, under the agent's current policy. Policy-dependent feedback enables effective teaching strategies like policy shaping (e.g., giving breadcrumbs for suboptimal behavior); COACH demonstrated evaluative feedback can directly improve an agent's policy via policy gradients.

Other work has considered how agents can learn from mixtures of human feedback. Li et al. proposed IRL-TAMER [16], in which an agent initially learns from demonstration and then is fine-tuned by scalar-based feedback via TAMER. Mourad et al. [17] also considered initially learning from demonstration under supervision before switching to binary evaluative feedback for fine-tuning. The binary feedback was myopic and policy-independent, with agents learning appropriate policies under their framework.

Bayesian IRL (Ramachandran and Amir [3]) approaches have been applied to *active learning* problems using preference-based feedback (Sadigh et al. [18]). The human teacher chooses the preferable trajectory from a trajectory pair that is determined and proposed by the agent. Jeon et al. [1] offers a framework incorporating *choice* of human feedback, from a diverse set of feedback types, as additional information regarding reward, assuming that a human's feedback choice is Boltzmann rational. Finally, Palan et al. [19], Biyik et al. [2] integrated learning from demonstrations with preference-based feedback in active learning settings.

Our work differs from prior work in that we aim to incorporate both action instruction and scalar-based human feedback, in light of empirical evidence that some mixture of these strategies is natural for human teachers to give. Chi and Malle [20] found in initial studies that most human participants who train an agent to act appropriately in a medical setting chose a combination of the two kinds of teaching; and they were guided both by the robot's performance and by their own accumulated impressions of the robot. Here we develop a human-centered learning algorithm allowing an agent to learn reward models from any combination of action- and scalar-based evaluative feedback.

III. Problem Statement

Teaching Objective. We aim to interactively teach an agent to successfully reach a goal while acting *appropriately* in line with norm-like environmental constraints (i.e., prohibited rooms). In this first stage, we do not use human participants as teachers but instead construct teacher models that guide agent behavior with different types of teaching. Here we compare teacher models that provide either scalar-based evaluative feedback with varying properties (e.g., advantage based, policy-dependent) or a mixture of action-based and evaluative feedback.

Learning Agent Model. In our approach, the agent maintains an internal estimate of a feature-based reward model $R(\phi(s); \mathbf{w})$, where ϕ is a state-based feature extractor; s is an environment state with features, such as position, or room type; and $\mathbf{w}$ are learned weights. The agent must learn a reward representation inducing a stochastic policy $\pi(a|s)$ that the teacher deems appropriate behavior, where $\pi(a|s)$ is the probability of taking action a in state s.

Reward-Induced Policy. To induce its policy, the agent solves a Markov Decision Process (MDP) by using its current internal reward function estimate as the reward function. Specifically, the agent models its world as a sequence of MDPs represented by (S, A, T, γ, R), where S is the set of environment states, A is the set of actions the agent can take, T is a transition model capturing the environment dynamics, γ is the agent's internal discount factor, and $R = R(\phi(s); \mathbf{w})$ is the agent's current internal reward estimate, which updates whenever feedback is received. Whenever the reward estimate is updated, the MDP is solved with dynamic programming using the softmax Bellman operator (Song et al. [21]) to obtain the $Q(s, a)$ state-action values and the corresponding stochastic policy. Using a computational graph allows the agent's stochastic policy to remain differentiable with respect to $\mathbf{w}$. The stochastic policy is simply the softmax of the Q values. The agent uses this policy to act in the world and elicit feedback from the teacher model.

Learning Reward. The agent learns its reward estimate by using a likelihood objective to update the reward function such that the induced policy better accounts for the teacher's feedback. Whenever feedback is provided, the teacher model has a choice of providing action-feedback $a_f \in A$ or evaluative scalar-feedback f, which is used to update the agent's reward model. To accomplish this independently of the teacher's underlying feedback strategy, hence flexibly for teachers with different strategies, the agent uses a likelihood objective—adjusted based on feedback received—to maximize the likelihood its induced policy results in desired behavior.

IV. Technical Approach

The agent utilizes different loss functions conditioned on the type of feedback received in order to adjust the internal reward function it uses to induce its policy.

Action Instruction. The agent seeks to maximize the likelihood $L(\pi; \mathbf{w}) = \prod_{(a_f, s) \in \mathcal{B}} \pi(a_f|s; \mathbf{w})$ that its policy takes appropriate actions a_f (i.e., actions that the teacher has indicated to be appropriate in state s), which have been stored as a result of interactions in a feedback buffer $\mathcal{B}$.

Evaluative Scalar Feedback. The agent seeks to adjust the likelihood of taking the specific action a from state s, based on the scalar feedback f it has received from the teacher. The pseudo-loss $L(\pi; \mathbf{w}) = -\log(\pi(a|s; \mathbf{w}))f$ creates an

appropriate gradient update with respect to $\mathbf{w}$. The evaluative feedback f that the teacher provides is a real number from $\mathbb{R}$ instead of being limited to binary feedback (good vs. bad).

Modeled Teacher Feedback Strategies

We model teacher feedback choices (instruct vs. evaluate) as a function of different strategies, characterized by two components: (1) policy-dependence vs. independence, which captures how much of the learner's current behavior the teacher takes into account when providing feedback; (2) myopic vs. forward-looking, which refers to variations in the teacher's discount factor, or the relative weights given to near-term and long-term consequences of the agent's current policy.

Policy-Independent Strategies. We created four *policy-independent* strategies: (1) Pure *action-based feedback*, where the teacher always responds with the optimal action a_f the agent should have taken from state s. (2) *Myopic raw advantage*, where the teacher uses their own optimal policy advantage values as feedback. (3) *Myopic ranked advantage*, which is the same as (2), except that the advantage values are mapped onto integers in the interval $[-2, 2]$ for better differentiation. (4) *Myopic ranked path-cost*, where the teacher ranks how optimal the agent's immediate action is, assuming it follows an optimal policy forever afterwards, but additionally penalizes immediate violations.

Policy-Dependent Strategies. We created two *policy-dependent* strategies: (1) *Forward-looking raw advantage feedback*, in which the teacher evaluates the agent's current policy using a high-discount factor, emphasizing the long-term consequences of the agent's current behavior. (2) *Forward-looking ranked advantage feedback*, which is the same as (1), except that the advantage values are mapped onto integers in $[-2, 2]$.

V. Evaluation

We evaluated how well the agent learns from the various teacher models described above, and we also considered the impact of the agent's own tendency to be myopic vs. forward-looking by varying its γ parameter.

Environment. We conducted our experiments in a 5x10 GridWorld, where each cell emits the agent's internal reward-estimate value based on the cell's color upon entry. The teacher provides feedback to the agent so that it learns to reach a goal-colored cell while learning to avoid prohibited-colored cells. The color of a cell is the feature used in the agent's reward function; there are 5 colors for which the agent needs to learn reward values. Code is available at https://github.com/hsbwncpodoet/hrilbr.

Evaluation Metrics

We evaluate the agent's learned behavior from feedback using two key metrics: (1) deterministic *norm violations*, and (2) deterministic *goal success*.

Deterministic Violations. The agent follows a greedy, deterministic version of its policy from each possible initial state, until either (1) all states are visited, or (2) a previously visited state is reached. The metric is the number of violations averaged over all initial states (50 in the present case).

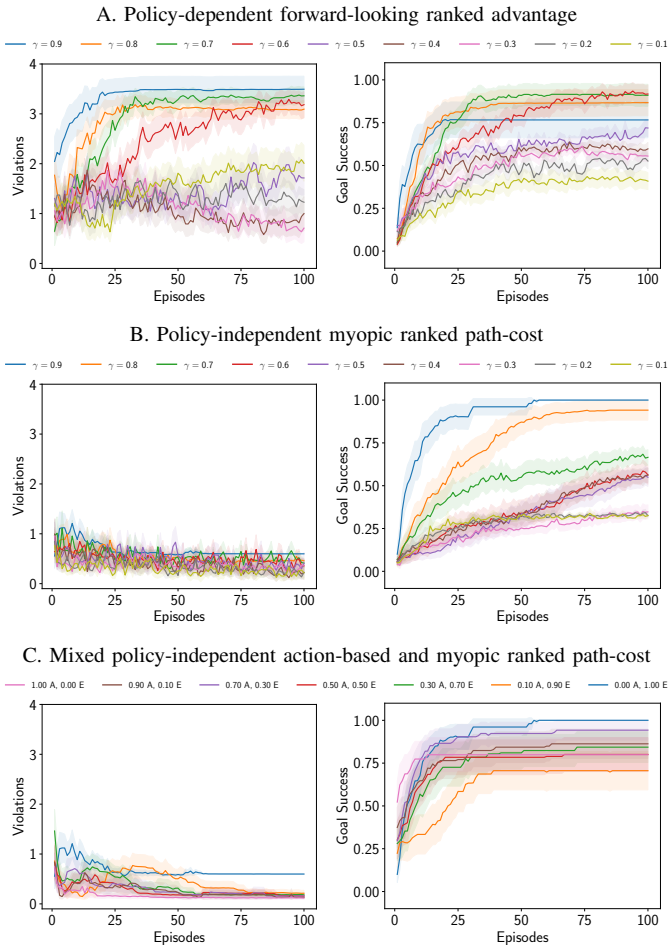

Fig. 1. **Deterministic goal success and policy violations.** The mean over 50 seeds is shown; shaded areas correspond to ± 0.25 standard deviation. Max episode length is 100 timesteps. **(a)** Policy-dependent forward-looking ranked advantage. **(b)** Policy-independent myopic ranked path-cost. **(c)** Mixed policy-independent action-based and policy-independent myopic ranked path-cost. **Legend:** For (a) and (b), the agent's internal discount factor γ. For (c), probability of giving A=action-based, E=evaluative at each timestep. Pure-action and pure-evaluative feedback is included.

Deterministic Goal Success. This metric is the percentage of initial states from which the agent is able to reach the goal state while following the deterministic, greedy policy for N steps, where $N = |S|$, the size of the state space.

VI. Results

We compare how the violation and goal success metrics are affected by (a) various teacher models (e.g., action-only, policy-dependent or policy-independent evaluative, and action-evaluative mixed) and (b) the agent's own discount factor γ. In mixed feedback strategies, the pure-action and pure-evaluative feedback strategies represent boundary cases.

Fig. 1 summarizes our simulations. The teaching strategy of *policy-dependent forward-looking ranked advantage* feedback (panel A.) leads to inconsistent learned behavior. Under this strategy, a teacher trades off keeping the agent's

norm violations low and propelling the agent to the goal. A forward-looking agent (high γ) taught with this strategy shows goal success but commits many norm-violating behaviors. A myopic agent (low γ) taught with this strategy commits few norm violations but hardly ever reaches the goal.

In contrast, giving *policy-independent myopic ranked path-cost* feedback (Fig. 1, panel B.) is more effective, consistently leading to appropriate behavior by forward-looking agents. Taught with this strategy, forward-looking agents (high γ) learn to reach the goal, whereas myopic agents (low γ) have lower goal success; the agents' low number of norm violations is not affected by γ.

In Fig. 1, panel C., teachers giving varying proportions of *action-based* and *evaluative feedback* (of the more effective policy-independent myopic ranked path-cost kind) elicit considerable goal success in their learners and help them keep their norm violations very low. The results suggest that the specific proportions have modest impact on goal success and norm violations: greater proportions of action-based feedback tend to keep norm violations minimal and greater proportions of evaluative feedback tend to increase goal success. A 70% action + 30% evaluation strategy appears to offer the best compromise.

VII. Discussion

Our evaluation results indicate that certain teaching and learning combinations influence how well agents learn appropriate behavior: a forward-looking teacher model providing policy-dependent ranked advantage feedback trades off goal success and violations, depending on how myopic the learning agent's discount factor γ is. Interestingly, while this teacher has access to the agent's current policy, the teacher is unaware of the agent's true γ, and assumes a *long-term* outlook when evaluating the agent's policy. This indicates a misalignment between how the teacher and agent value future consequences; when learning, the agent interprets the teacher's feedback as the *immediate consequence* of the most recent action.

In contrast, a myopic teacher model providing policy-independent ranked path-cost feedback appears to successfully teach an agent appropriate behavior: the agent learns to reach the goal while minimizing norm violations. Here, the teacher evaluates the appropriateness only of the agent's most recent action, by comparing against exemplar behavior, and then assumes the agent will subsequently be an exemplar for the foreseeable future. We observe that this type of feedback is consistent with how the agent interprets feedback during learning, which explains why variations in the agent's γ alter goal success but do not affect violation minimization.

Teacher models that provided a combination of *action-based* and *evaluative feedback* trained the agent to both successfully reach the goal and to keep norm violations very low. As a policy-independent teaching method, the feedback corresponds well with the agent's interpretation of the immediate consequences of its behavior. If this kind of teacher is representative of human teachers in natural environments—as preliminary behavioral experiments suggest (Chi and Malle [20])—then our agent should display effective learning when encountering real-world teachers. Notably, the mixed-strategy teacher models we implemented were successful despite *randomly* mixing the two feedback types, whereas humans are likely to display systematic choice patterns (i.e., when to use which strategy), and they may teach our agent even better or faster.

Our approach assumes that agent and teacher have single, fixed, preset discount factors that decay future rewards geometrically. It may be beneficial to consider how diverse discount factors and diverse time horizons influence the learning and teaching process, as represented by hyperbolic discounting (Kurth-Nelson and Redish [22], Fedus et al. [23]). Using eligibility traces (Klopf [24], Sutton [25], MacGlashan et al. [15], van Hasselt et al. [26]) for learning to dynamically adjust between immediate consequences and long-term positive outcomes would likely improve the learning process.

VIII. Conclusion and Future Work

As robots become more socially integrated into society, human-centered learning algorithms will need to be flexible and adaptable to accommodate diverse human teaching styles. Our investigation was inspired by evidence that humans prefer to use combinations of teaching strategies. The results of this investigation suggest that learning algorithms can indeed integrate different teaching strategies with considerable success.

We focused on two teaching strategies prevalent in social settings that have been widely studied in HRI work: instructing a learner to perform correct actions (action-based feedback) and evaluating the learner's actions (evaluative feedback). We integrated the two types of feedback under a single likelihood-based algorithm to estimate a reward function that induces policies maximally likely to satisfy the teacher's defined goal and norm constraints. We evaluated this integrative algorithm in simulated training sessions using goal success and norm violation metrics. Importantly, we compared policy-dependent and policy-independent teacher models—implemented with ranked advantage and ranked path-cost, and myopic and forward-looking strategies—to assess how well the agent learned under their tutelage. Our results indicate that agents benefit the most from combined demonstration and evaluative feedback strategies. We envision this method can be used to teach robots in a variety of environments, but especially in complex social environments, such as hotels or hospitals.

In future work, we will investigate training agents with more realistic teacher models, including a range of human participants to observe the broadest array of learning outcomes in natural environments. We also aim to address generalization of the agent's feature-based reward representation and aim to evaluate our method in complex environments with large numbers of context-specific norms. Finally, we hope to explore how learning agents can address a fundamental challenge of norm-based social action: while norms generally require community agreement on what the appropriate behavior is, people often disagree. A robot learner must be able to respond to such disagreements—perhaps by asking humans themselves to clarify or resolve the disagreement.

REFERENCES

[1] H. J. Jeon, S. Milli, and A. D. Dragan, "Reward-rational (implicit) choice: A unifying formalism for reward learning," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/2f10c1578a0706e06b6d7db6f0b4a6af-Abstract.html

[2] E. Biyik, D. P. Losey, M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh, "Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences," *CoRR*, vol. abs/2006.14091, 2020. [Online]. Available: https://arxiv.org/abs/2006.14091

[3] D. Ramachandran and E. Amir, "Bayesian inverse reinforcement learning," in *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, M. M. Veloso, Ed., 2007, pp. 2586–2591. [Online]. Available: http://ijcai.org/Proceedings/07/Papers/416.pdf

[4] J. M. Clegg, N. J. Wen, P. H. DeBaylo, A. Alcott, E. C. Keltner, and C. H. Legare, "Teaching Through Collaboration: Flexibility and Diversity in Caregiver–Child Interaction Across Cultures," *Child Development*, vol. 92, no. 1, Jan. 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/cdev.13443

[5] M. A. Kline, "How to learn about teaching: An evolutionary framework for the study of teaching behavior in humans and other animals," *Behavioral and Brain Sciences*, vol. 38, 2015, publisher: Cambridge University Press. [Online]. Available: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/how-to-learn-about-teaching-an-evolutionary-framework-for-the-study-of-teaching-behavior-in-humans-and-other-animals/017C2C246E9C206562CAE3DB590B01EC

[6] A. F. Grasha, "A Matter of Style: The Teacher as Expert, Formal Authority, Personal Model, Facilitator, and Delegator," *College Teaching*, vol. 42, no. 4, pp. 142–149, 1994, publisher: Taylor & Francis, Ltd. [Online]. Available: https://www.jstor.org/stable/27558675

[7] P. Abbeel and A. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the twenty-first international conference on Machine learning*, 2004.

[8] A. Y. Ng and S. J. Russell, "Algorithms for inverse reinforcement learning," in *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, P. Langley, Ed. Morgan Kaufmann, 2000, pp. 663–670.

[9] A. Gleave, M. Dennis, S. Legg, S. Russell, and J. Leike, "Quantifying differences in reward functions," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: https://openreview.net/forum?id=LwEQnp6CYev

[10] D. S. Brown and S. Niekum, "Machine teaching for inverse reinforcement learning: Algorithms and applications," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 7749–7758. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33017749

[11] S. Arora, P. Doshi, and B. Banerjee, "I2RL: online inverse reinforcement learning under occlusion," *Auton. Agents Multi Agent Syst.*, vol. 35, no. 1, p. 4, 2021. [Online]. Available: https://doi.org/10.1007/s10458-020-09485-4

[12] P. Kamalaruban, R. Devidze, V. Cevher, and A. Singla, "Interactive teaching algorithms for inverse reinforcement learning," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, S. Kraus, Ed. ijcai.org, 2019, pp. 2692–2700. [Online]. Available: https://doi.org/10.24963/ijcai.2019/374

[13] W. B. Knox, "Learning from human-generated reward," PhD dissertation, The University of Texas at Austin, 2012. [Online]. Available: https://www.bradknox.net/wp-content/uploads/2013/06/thesis-knox.pdf

[14] J. MacGlashan, M. Littman, D. Roberts, R. Loftin, B. Peng, and M. E. Taylor, "Convergent actor critic by humans," in *International Conference on Intelligent Robots and Systems*, 2016.

[15] J. MacGlashan, M. K. Ho, R. Loftin, B. Peng, G. Wang, D. L. Roberts, M. E. Taylor, and M. L. Littman, "Interactive learning from policy-dependent human feedback," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 2285–2294. [Online]. Available: https://proceedings.mlr.press/v70/macglashan17a.html

[16] G. Li, B. He, R. Gomez, and K. Nakamura, "Interactive reinforcement learning from demonstration and human evaluative feedback," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2018, pp. 1156–1162.

[17] N. Mourad, A. Ezzeddine, B. N. Araabi, and M. N. Ahmadabadi, "Learning from demonstrations and human evaluative feedbacks: Handling sparsity and imperfection using inverse reinforcement learning approach," *Journal of Robotics*, vol. 2020, 2020. [Online]. Available: https://doi.org/10.1155/2020/3849309

[18] D. Sadigh, A. D. Dragan, S. Sastry, and S. A. Seshia, "Active preference-based learning of reward functions," in *Robotics: Science and Systems XIII, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, July 12-16, 2017*, N. M. Amato, S. S. Srinivasa, N. Ayanian, and S. Kuindersma, Eds., 2017. [Online]. Available: http://www.roboticsproceedings.org/rss13/p53.html

[19] M. Palan, G. Shevchuk, N. C. Landolfi, and D. Sadigh, "Learning reward functions by integrating human demonstrations and preferences," in *Robotics: Science and Systems XV, University of Freiburg, Freiburg im Breisgau, Germany, June 22-26, 2019*, A. Bicchi, H. Kress-Gazit, and S. Hutchinson, Eds., 2019. [Online]. Available: https://doi.org/10.15607/RSS.2019.XV.023

[20] V. B. Chi and B. F. Malle, "Instruct or evaluate: how people choose to teach norms to social robots," in *Companion of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. New York, NY, USA: Association for Computing Machinery, 2022.

[21] Z. Song, R. Parr, and L. Carin, "Revisiting the softmax bellman operator: New benefits and new perspective," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 2019, pp. 5916–5925. [Online]. Available: http://proceedings.mlr.press/v97/song19c.html

[22] Z. Kurth-Nelson and A. Redish, "Temporal-difference reinforcement learning with distributed representations," *PLoS ONE*, vol. 4(10), 2009. [Online]. Available: https://doi.org/10.1371/journal.pone.0007362

[23] W. Fedus, C. Gelada, Y. Bengio, M. G. Bellemare, and H. Larochelle, "Hyperbolic discounting and learning over multiple horizons," *CoRR*, vol. abs/1902.06865, 2019. [Online]. Available: http://arxiv.org/abs/1902.06865

[24] A. H. Klopf, "Brain function and adaptive systems: A heterostatic theory," 1972.

[25] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Mach. Learn.*, vol. 3, pp. 9–44, 1988. [Online]. Available: https://doi.org/10.1007/BF00115009

[26] H. van Hasselt, S. Madjiheurem, M. Hessel, D. Silver, A. Barreto, and D. Borsa, "Expected eligibility traces," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 9997–10 005. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17200