A Comprehensive Survey on Reinforcement Learning from Human Feedback (RLHF)

AI 533 Intelligent Agents and Decision Making

Winter 2025

Hyuntaek Oh
Oregon State University
ohhyun@oregonstate.edu

Introduction

Reinforcement learning (RL) has emerged as a powerful framework for solving decision-making problems, achieving remarkable success in fields such as robotics, natural language processing (NLP), and computer vision. However, traditional RL is often hindered by its limitations like sample inefficiency, challenges in defining effective reward functions, and safety issues in simulations and real-world applications. These issues have triggered great interest in reinforcement learning from human feedback (RLHF), a paradigm that integrates human knowledge directly into the learning process.

Integration of human feedback into reinforcement learning (RLHF) addresses these challenges, providing an efficient mechanism to guide agents to behaviors aligned with human intentions and expectations. Human evaluative feedback - ranging from scalar evaluations, linguistic feedback, to binary corrections - can make robots and autonomous systems obtain an implied understanding of tasks, overcoming the limitations such as sparse and ambiguous reward signals commonly encountered in traditional RL frameworks.

The importance of RLHF is derived from its ability to bridge the gap between pure algorithmic decision-making and human-based performance criteria. Human feedback is not only an additional data source, but it also includes rich contextual and evaluative insights that are typically beyond the expression of predefined reward structures. By using human feedback, the agent in RL frameworks can learn efficiently and improve its performance significantly.

In this review, we analyze recent advances in reinforcement learning from human feedback, exploring how various forms of human inputs have been effectively used to improve autonomous systems. We discuss the strengths and weaknesses of these approaches, examine the review papers' novel approaches, evaluate experiment designs, and describe open challenges and future directions.

1 Primitive Skill-based Robot Learning from Human Evaluative Feedback

1.1 Paper Information

The paper "Primitive Skill-based Robot Learning from Human Evaluative Feedback" was written by Ayano Hiranaka, Minjune Hwang, Sharon Lee, Chen Wang, Li Fei-Fei, jiajun Wu, and Ruohan Zhang, and published in the International Conference on Intelligent Robots and Systems (IROS) in 2023.

1.2 Key Problem

Reinforcement Learning (RL) for long-horizon robot manipulation in real-world environments has been hindered by sample inefficiency and safety issues since robot manipulation tasks often involve continuous state and action spaces with complex dynamics, and a reward function alone is not sufficient to control a robot's unintended behaviors. Similarly, in Reinforcement Learning from Human Feedback (RLHF), minimizing the associated human effort and cost is crucial for practicality. Especially in long-horizon robot manipulation tasks, the number of required human feedback grows significantly, making it impractical to rely heavily on human feedback.

1.3 Summary

The authors introduce Skill-based Evaluative Feedback (SEED), a framework that leverages RLHF and primitive skill-based motion control. There are several contributions of using SEED to RLHF research field.

- Combining human evaluative feedback with primitive skills, which has not previously explored, reduces human effort.
- Evaluation without execution allows human to provide feedback on skill selections before robot actions are executed, enhancing safety.
- By conducting extensive experiments on five manipulation tasks of varying complexities in simulation and real-world, SEED significantly outperforms alternative approaches with respect to sample efficiency, safety, and human effort.

To address key problems, SEED breaks down long-horizon tasks into sequences of primitive skills, parameterizing the predefined primitive skills such as picking and placing instead of knowing underlying low-level motor control mechanisms. Similar to supervised learning, skill-based evaluative feedback is used to train a critic network optimizing skill selection and enhancing sample efficiency. Feedback is in the form of scalar values (good, neutral, bad) to guide the agent's learning process. Before the robot acts, humans evaluate each action on the visual representation of the robot's skills and parameters in advance. After checking whether its future action is safe, humans allow the robot to execute actions, reducing safety risks in real-world applications. During off-policy, SEED balances the replay buffer by ensuring an equal mix of good and bad feedback samples, improving convergence speed.

SEED introduces simplified affordance scores, which are inspired by MAPLE [6], to penalize actions that are inappropriate for the situation. This auxiliary reward helps the agent learn useful, proper skill parameters more efficiently. SEED achieved higher sample efficiency, outperforming SAC [7], TAMER [8], and MAPLE [6] in terms of learning speed and success rate. It also improved safety by utilizing the evaluation without execution technique, preventing dangerous actions from being executed. Through three different long-horizon tasks, SEED significantly reduces the number of human feedback by learning effectively, while TAMER [8] failed to achieve any successful task completion.

1.4 The Strengths and Weaknesses of the paper

This paper presents an empirical advance rather than a new theoretical insight. SEED solves standing open problems such as sample inefficiency and safety in real-world robot learning.

The problem formulation is aligned with real-world robot challenges. However, SEED requires predefined skills that have difficulty generalizing to new tasks. If the problem is large-scale real-world, it may still require extensive human effort even with the reduced number of feedback.

SEED provides a faster and safer solution compared to existing RLHF and skill-based RL methods.

The paper evaluated SEED with other algorithms in different categories such as learning and feedback efficiency independently and did not consider imitation learning with human feedback, which can be a potential candidate with similar or higher performance.

SEED's originality is sufficient to contribute to RLHF fields since the integration of primitive skills and human feedback has not been explored before.

1.5 Experiment Setup, Results, and Validation

SEED was evaluated on five robotic manipulation tasks in both simulation (Robosuite) and real-world settings using a Franka Emika Panda robot arm. The results were compared to Soft Actor-Critic [7], TAMER [8](RLHF-based model), MAPLE [6](Primitive skill-based RL), and MAPLE-aff [6](MAPLE with affordance rewards) by using the metrics that evaluate the performance with task success rate.

Given the complex and challenging stacking task that was provided, SEED learns to solve the task in only 0.8 million steps, meaning that it is sample efficient compared to 4 million steps of MAPLE-aff [6]. It significantly reduces safety violations by allowing human evaluation before execution. It also requires less human feedback to achieve learning, making it practical.

The experiments in the paper validate the main claims that improve sample efficiency, enhance safety, and lower human effort.

1.6 Key Takeaways, Limitations, and Open Challenges

SEED demonstrates that integrating RLHF with primitive skills and evaluation without execution can significantly improve RL performance in real-world robotic tasks. It relies on predefined primitive skills, which may limit its generalization to new tasks. There is a need to extend SEED to more diverse and unstructured robotic tasks without predefined skills.

2 Reinforcement Learning with Human Feedback for Realistic Traffic Simulation

2.1 Paper Information

"Reinforcement Learning with Human Feedback for Realistic Traffic Simulation" by Yulong Cao, Boris Ivanovic, Chaowei Xiao, and Marco Pavone, was submitted to the IEEE International Conference on Robotics and Automation (ICRA) 2024.

2.2 Key Problem

catching human preference on realism and unifying diverse traffic simulation models are main challenges to develop effective, reliable autonomous vehicle systems. Existing simulation models depend heavily on their target setting, making it difficult to integrate human preferences across different models.

2.3 Summary

The paper introduces a reinforcement learning with human feedback (RLHF) framework called "TrafficRLHF" by incorporating human feedback into the simulation process, encouraging more human-like traffic behavior in diverse environments. The contributions of the paper are:

- Providing the first dataset of realism alignment for traffic modeling
- Formulating a reward model that quantifies realism according to human preferences
- Offering a versatile RLHF-based framework that enhances the realism of a wide range of existing traffic models

In the framework, there are three stages to tackle the key problems: human feedback collection in stage 1, a reward model is trained using the human feedback to predict how closely a given traffic scenario aligns with human driving behavior in stage 2, and the trained reward model is used for fine-tuning of the traffic models, thereby improving the realism of the generated traffic scenarios. Due to its model-agnostic characteristic, TrafficRLHF can be used to improve different traffic models without major architectural modifications.

TrafficRLHF with fine-tuning reward model significantly reduced unrealistic driving actions such as collisions, off-road conditions, and unpredicted stops. In TABLE 3, for Conditional Traffic Generation

(CTG), it reduced realism scores to 0.38 (from 0.57) and failure rate to 0.05 (from 0.27), and cost to 3.70 (from 13.21).

2.4 The Strengths and Weaknesses of the paper

This paper is much closer to empirical advance than a new theoretical insight. To solve the failures of traditional traffic simulation models to align generated traffic with real-world human driving behavior, TrafficRLHF leverages a combination of human preferences and traffic simulations.

The problem formulation of the paper was well-designed since it addressed not only global semantic maps, but it also defined decision-relevant context consisting of local semantic maps for all agents.

However, there is a potential edge case such as the ambiguity of human preferences since labeling maps and vehicles may be slightly different, depending on how the annotator processes it. Unlike traditional generative traffic models, TrafficRLHF provides more reliable realism, a lower failure rate, and a relatively lower cost.

The paper did not compare TrafficRLHF with other reinforcement learning methods such as Deep Q-learning (DQN), which can be a competitive candidate.

Even though applying RLHF alone is not a new idea, its application to traffic simulation and realism alignment based on human feedback is novel.

2.5 Experiment Setup, Results, and Validation

The experimental evaluation of TrafficRLHF on the nuScenes dataset and various traffic simulation models such as Conditional Traffic Generation (CTG), Bi-Level Imitation Learning System (BITS), and TrafficGen, with or without a fine-tuning reward model. The evaluation metrics include realism score, failure rate, and cost.

TrafficRLHF successfully improves the realism of traffic simulation by aligning the simulation-based behaviors with human preferences. It significantly reduces failure rates, and it can be applied to multiple traffic simulation models by utilizing a reward model with fine-tuning.

The experiments support the main claims of the paper since TrafficRLHF improves realism and safety in traffic simulations.

2.6 Key Takeaways, Limitations, and Open Challenges

One of the key takeaways is TrafficRLHF successfully integrates human feedback into traffic simulations, improving realism and reducing failure rates. However, the author does not explicitly account for rare, high-risk driving behaviors such as aggressive drivers or emergency braking that may be crucial for operating autonomous vehicles safely. Moreover, it is necessary for future works to reflect the drivers in the real world since the agents (the vehicles) in the experiment may have similar driving behaviors.

3 Reinforcing an Image Caption Generator Using Off-Line Human Feedback

3.1 Paper Information

Paul Hongsuck Seo, Piyush Sharma, Tomer Levinboim, Bohyung Han, and Radu Soricut authored the paper "Reinforcing an Image Caption Generator Using Off-Line Human Feedback" in the Association for the Advancement of Artificial Intelligence (AAAI) 2020.

3.2 Key Problem

The key problem addressed in the paper is the mismatch between the training objective of image captioning models and the actual human evaluation criteria used to assess caption quality, and collecting direct ratings from human feedback is expensive and limited, creating a sparse training signal.

3.3 Summary

This paper is about proposing a method based on off-policy policy gradient with an alternative sampling distribution, learning to generalize the human raters' judgments to a previously unseen set of images to tackle the key problem above. The proposed method successfully deals with the sparsity of information about the rating function by leveraging the signal from instance-level human caption ratings to improve captioning models. The paper's contributions are:

- The authors propose to train captioning models using human ratings produced during evaluations of previous models
- The authors propose an off-policy policy gradient method to cope with the sparsity in available caption ratings
- The authors present a set of experiments using human evaluations that demonstrates the effectiveness of our approach

The main approach employs an off-policy policy gradient method, which updates the caption generation model by taking gradient steps towards captions that received higher human ratings. The expected ratings of the output captions:

$$J(\theta) = E_{I \sim D(I), c \sim p_\theta(c|I)}[r(c|I)],$$

where $p_D(I)$ is the dataset distribution for I and $p_\theta(c|I)$ is the conditional caption distribution estimated by a model parameterized by θ, and $r(c|I)$ is possibly an aggregate of multiple ratings from different raters. Based on the expected ratings, alternative distribution $q(c|I)$ is used to adopt an off-policy policy gradient technique, instead of the true policy distribution $p_\theta(c|I)$ for sampling. The approximation of the policy gradient is:

$$\nabla_\theta J_{PG}(\theta) \approx \frac{1}{S} \sum_{s=1}^{S} \frac{p_{\theta(c|I)}}{q(c|I)} (r(C_s|I_s) - b) \nabla_\theta ln p_\theta(c_s|I_s)$$

, where $q(c|I) = (1 - \epsilon) p_D(c|I) + \epsilon U(c)$

, $\frac{p_{\theta(c|I)}}{q(c|I)}$ represents the importance weight for sample caption c_s and image I_s, and S is the number of samples. With these equations, the authors integrate curriculum learning by initially training the model using MLE on a large-scale captioning dataset, and then fine-tuning with reinforcement learning using human ratings as rewards.

According to Table 2 in the paper, the proposed off-policy gradient method (offPG) improves single-caption based on human evaluation, achieving a "Goodness" score of 68.42% compared to 66.23% for the baseline model. Table 3 (side-by-side human evaluations) shows that offPG outperforms the baseline by 7.45% in Informativeness, by 5.90% in Correctness, and by 1.69% in Fluency.

3.4 The Strengths and Weaknesses of the paper

The paper provides an empirical advance by applying existing reinforcement learning techniques. It may not solve a standing open problem but addresses a practical challenge in terms of bridging the gap between training objectives and human evaluations in image captioning.

The authors formulated why using human ratings as rewards in the RL framework is reasonable, and they pointed out what the problems of existing methods in image captioning are in a sound manner. Especially, the use of importance-weighted policy gradients properly addresses the sparsity and distributional mismatch between rated captions and random samples.

Due to the simplification of their RL problem, there may be an edge case regarding the different dimensions of caption quality feedback. Their model may not represent multiple dimensions of caption quality since the human rating in their evaluation does not encode these individual aspects. Their solution for an existing problem achieves significant improvements over baseline methods on important quality dimensions.

There may be a feasible, practical supervised baseline such as self-critical training, which can be a candidate to compare.

Their method, off-policy RL with sparse feedback, is a novel methodological contribution to image captioning.

3.5 Experiment Setup, Results, and Validation

For experiments, the authors used the Conceptual Captions dataset [9] for image captioning, and the training and validation splits have 3.3M and 16K samples, respectively. For human feedback, they utilized the Caption Rating dataset where each caption was rated by multiple raters. An encoder-decoder Transformer Network architecture was adopted as their captioning model, whereas the Baseline model is trained with MLE only.

The authors evaluated on a separate T2 dataset with two forms of human evaluations performed: single-caption evaluation and side-by-side evaluation. Off-policy gradient (offPG) showed significant improvements over all baselines in various metrics.

The experiment setup validates the claim that using sparse, human feedback collected previously in an off-policy RL framework significantly improves caption quality.

3.6 Key Takeaways, Limitations, and Open Challenges

The method in the paper showed that using human feedback is a more effective way to solve the key problem compared to the baseline method. Simplification to single numerical scores for training human judgments can be a limitation of the paper since it can ignore explicit distinctions among different aspects of caption quality. Modeling multiple quality dimensions is an open challenge, considering how to efficiently incorporate multi-dimensional feedback.

4 Learning Rewards From Linguistic Feedback

4.1 Paper Information

The paper "Learning Rewards from Linguistic Feedback" was published by Theodore R. Sumers, Mark K.Ho, Robert D. Hawkins, Karthik Narasimhan, and Thomas L. Griffiths at the thirty-fifth Association for the Advancement of Artificial Intelligence (AAAI) conference in 2021.

4.2 Key Problem

Reinforcement Learning (RL) agents usually learn reward functions from numeric rewards, binary evaluations, or commands. However, these reward functions are not aligned with the feedback from the real world since one of the forms of human feedback is unconstrained, naturalistic language. The key problem in the paper is how to make agents learn reward functions from human language feedback.

4.3 Summary

The authors propose a reinforcement learning framework allowing agents to infer reward functions from human natural language feedback. They utilize two methods: aspect-based sentiment analysis to decompose human linguistic comments into sentiment scores and learning an end-to-end mapping from utterances and context to rewards with linear regression. Experiments involve human-human and human-agent interactions by showing that the approach can make agents interpret diverse, naturalistic human feedback. The key contributions of the paper are:

- Human linguistic feedback is decomposed into sentiment and target features
- The pairs of sentiment and MDP features based on the probability distribution capture human preferences
- The agents can infer the underlying reward function efficiently from linguistic inputs

The method first decomposes human feedback using aspect-based sentiment analysis into two components: sentiment that is a positive or negative evaluation and features that represent states or actions within a Markov Decision Process (MDP). After decomposing the feedback, Bayesian linear regression is applied to these features and the corresponding sentiment scores. Then, they compare two agent models - ’literal’ and ’pragmatic’ - to a neural inference network trained end-to-end.

Three different types of models - a "literal" sentiment analysis model, a "pragmatic" sentiment model, and an end-to-end neural inference network - were evaluated. In human-agent interaction experiments, the "pragmatic" model (Live Mean Normalized Score 43) outperformed both the "literal"(Live Mean Normalized Score 34) and inference network models (Live Mean Normalized Score 35). This implies the value of leveraging linguistic pragmatics in interpreting naturalistic human feedback.

4.4 The Strengths and Weaknesses of the paper

The paper offers an empirical advance rather than a new theoretical insight. The proposed method addresses practical challenges in interactive learning by successfully demonstrating inference of latent rewards from human language.

The paper's formulation of the problem is sound since it addresses naturalistic, varied human feedback rather than constrained or predefined linguistic forms. However, there are potential edge cases. For example, rare, ambiguous, or even contradictory linguistic expressions may not clearly map to features or sentiment.

The solution the paper provides to an existing problem is faster and more effective by learning rewards from human feedback. Unlike traditional reinforcement learning from human feedback, which requires structured or constrained input, the proposed method presents an interpretation of human language through sentiment and MDP feature regression, closely approaching human-level performance.

One potential gap is the comparison of their method with inverse reinforcement learning approaches that also utilize linguistic feedback, but in more constrained or structured forms.

Previous works typically relied on predefined or structured commands or binary rewards, whereas this paper introduces a naturalistic approach that can include diverse forms of human language, meaning that it is a sufficiently novel contribution.

4.5 Experiment Setup, Results, and Validation

The experiment setup in the paper includes a collaborative task where human teachers provided linguistic feedback to human and learner agents playing a cooperative game. In each episode, the learner controlled a robot to collect objects with hidden, different rewards, while the teacher observed the reward and provided unconstrained language feedback such as "Good job". Three types of artificial learners (literal, pragmatic, and inference network) use the feedback to infer the hidden reward function.

The experiment results indicate that all three artificial learners successfully inferred rewards from unconstrained linguistic feedback provided by humans. Among the three, the "pragmatic" model outperformed the others, nearly reaching human-level performance, showing the benefits of integrating pragmatic biases.

The experiments validate the paper's main claims and show that the proposed methods successfully interpret unconstrained human linguistic feedback to infer rewards by decomposing natural language into sentiment and MDP features.

4.6 Key Takeaways, Limitations, and Open Challenges

Aspect-based sentiment analysis and pragmatic reasoning the paper proposed are key takeaways to demonstrate that natural language feedback can be used in a practical way. However, the performance depends on ground truth from human language, meaning that the misinterpretation of rare, vague, or contradictory statements can be problematic. Developing robust language understanding methods would be needed to clearly handle the nuances of the human feedback.

5 Interactive Reinforcement Learning with Inaccurate Feedback

5.1 Paper Information

The paper "Interactive Reinforcement Learning with Inaccurate Feedback" by Talor A. Kessler Faulkner, Elaine Schaertl Short, and Andrea L. Thomaz, was published at the International Conference on Robotics and Automation (ICRA) in 2020.

5.2 Key Problem

The key problem in the paper is how to make Reinforcement Learning (RL) agents learn effectively from imperfect human feedback. Traditional interactive RL assumes that human feedback is always reliable, but human teachers can provide inaccurate or inconsistent feedback.

5.3 Summary

The paper introduces a framework to handle such inaccuracies, and the authors propose an algorithm, Revision Estimation from Partially Incorrect Resources (REPaIR), which can estimate and correct potentially incorrect feedback over time to improve learning efficiency.

- The authors introduce a framework for interactive RL with inaccurate feedback
- The authors propose REPaIR, an algorithm that detects and corrects inaccurate human feedback by estimating its reliability based on cumulative rewards

The authors propose a framework called Imperfect Feedback Markov Decision Process (IFMDP), which extends standard RL models by incorporating both correct feedback and potentially incorrect human feedback. The REPaIR algorithm can track feedback reliability by comparing received feedback against cumulative rewards, assign trust scores to human feedback, and filter or revise feedback. The learning process is similar to standard RL with human feedback, but REPaIR is used to estimate and adjust the feedback dynamically and update its policy based on corrected feedback.

Applying REPaIR into other baselines such as TAMER-P, TAMER-W, and Policy Shaping (PS) improved learning performance. In 83.33% of test cases, REPaIR matched or exceeded the performance of baselines in a variety of simulation settings.

5.4 The Strengths and Weaknesses of the paper

The paper provides both a new theoretical contribution and an empirical advance to interactive reinforcement learning. It offers a mathematical foundation for understanding how RL agents can learn when human-provided rewards are partially incorrect. The authors propose REPaIR, a novel algorithm that estimates and corrects inaccurate human feedback during learning. It does not explicitly handle a standing open problem, but it tackles a major limitation in interactive RL, which is based on the assumption that human feedback is always reliable.

The sound formulation the authors presented articulates that human feedback can be unreliable and proposes REPaIR that dynamically adjusts trust in feedback based on cumulative rewards. There may be an edge case such as highly misleading or adversarial feedback. If a human intentionally provides misleading or systematically biased feedback, the algorithm may struggle to handle it.

REPaIR improves unreliable feedback detections by correlating feedback with long-term rewards, adjusting feedback dynamically. This means that the paper provides a better solution.

Deep learning using multiple neural networks may possibly be applied to baselines for comparison since it can be a competitive candidate.

The REPaIR algorithm is a sufficient novelty in its contributions, but the imperfect feedback system is not enough since there is already a similar approach such as corrupt reward models.

5.5 Experiment Setup, Results, and Validation

The experiment setup evaluates the REPaIR algorithm in both a simulated environment and real-world robotics tasks to test its ability to handle imperfect human feedback. In the simulated environment, the agent learns to place objects into two bins under pre-defined rules. In the real-world robotics task, a Kinova Jaco robot arm learns to grasp a cup using feedback from a noisy object detection system.

For simulation results, REPaIR outperformed all baselines in 83.33% of test cases where the feedback correctness is unknown. While baseline methods showed lower performance when feedback accuracy drops, REPaIR maintains relatively stable performance. At low feedback quality, REPaIR still learns better policies. For real-world robot results, REPaIR can grasp objects more reliably and matches or exceeds baseline performance in handling sensor errors.

The experiments validate the main claims in the paper since REPaIR improves learning when feedback is inaccurate, estimates and corrects unreliable feedback, and generalizes to real-world robotics tasks in a practical way.

5.6 Key Takeaways, Limitations, and Open Challenges

REPaIR improves learning from inaccurate human feedback, maintains decent performance even with noisy feedback, and generalizes to real-world robotics. However, there is no mention of speed optimization, meaning how long the robot takes to do its work. REPaIR assumes feedback errors are random, not systematic, meaning that it will be able to show poor performance if human feedback is biased or adversarial. Future work could explore how REPaIR adapts when human feedback is intentionally misleading or biased in specific states.

Conclusion

The papers reviewed demonstrate that incorporating human feedback significantly improves performance in terms of efficiency, safety, and alignment of the agents across various domains such as image caption generation, robotic manipulation, linguistic reward learning, realistic traffic simulation, and inaccurate feedback scenarios.

There are some key takeaways from using human evaluative feedback in reinforcement learning frameworks, based on the review papers above. One of the takeaways is improvement in performance and efficiency. For example, the use of off-line human evaluation in image captioning provides meaningful signals, effectively guiding models to outputs more closely aligned with human judgments. Primitive skill-based robot learning with evaluative feedback also shows notable improvements in safety and faster skill learning, making robots perform complex real-world manipulation tasks more efficiently.

Another is the flexibility and expressiveness of human feedback. Linguistic and evaluative feedback, for instance, encourages agents to infer human preferences with meanings that would be challenging to interpret through traditional reward functions alone. Natural language processing offers an expressive and flexible measure for guiding autonomous systems, showing that linguistic understanding can enhance reinforcement learning outcomes.

Addressing feedback quality and reliability is also one of the key takeaways. While leveraging human evaluations promotes learning performance, handling imperfect or noisy feedback needs to be considered like the REPaIR algorithm, showing that even when human feedback is inconsistent or inaccurate, agents' learning should be able to be maintained.

Due to the applicatory property of human feedback in various domains, there exist several open challenges. Many existing RLHF frameworks require domain-specific modifications for generalization and scalability, requiring extensive customization. While RLHF effectively reduces the complexity of reward design, providing evaluative or linguistic feedback still brings considerable burdens on human trainers. In the long-term deployments, it is necessarily guaranteed that agent behaviors remain continuously aligned with human intentions for trustworthy and dependable systems.

References

[1] A. Hiranaka et al., “Primitive Skill-Based Robot Learning from Human Evaluative Feedback,” in 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Oct. 2023, pp. 7817–7824. doi: 10.1109/IROS55552.2023.10341912.

[2] Y. Cao, B. Ivanovic, C. Xiao, and M. Pavone, “Reinforcement Learning with Human Feedback for Realistic Traffic Simulation,” in 2024 IEEE International Conference on Robotics and Automation (ICRA), May 2024, pp. 14428–14434. doi: 10.1109/ICRA57147.2024.10610878.

[3] P. H. Seo, P. Sharma, T. Levinboim, B. Han, and R. Soricut, “Reinforcing an Image Caption Generator Using Off-Line Human Feedback,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 03, pp. 2693–2700, Apr. 2020, doi: https://doi.org/10.1609/aaai.v34i03.5655.

[4] T. R. Sumers, M. K. Ho, R. D. Hawkins, K. Narasimhan, and T. L. Griffiths, “Learning Rewards From Linguistic Feedback,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 7, pp. 6002–6010, May 2021, doi: 10.1609/aaai.v35i7.16749.

[5] T. A. Kessler Faulkner, E. Schaertl Short and A. L. Thomaz, "Interactive Reinforcement Learning with Inaccurate Feedback," 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 2020, pp. 7498-7504, doi: 10.1109/ICRA40945.2020.9197219.

[6] Soroush Nasiriany, Huihan Liu, and Yuke Zhu. Augmenting reinforcement learning with behavior primitives for diverse manipulation tasks. In 2022 International Conference on Robotics and Automation (ICRA), pages 7477–7484. IEEE, 2022.

[7] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In International conference on machine learning, pages 1861–1870. PMLR, 2018.

[8] W Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The tamer framework. In Proceedings of the fifth international conference on Knowledge capture, pages 9–16. ACM, 2009.

[9] Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In ACL, 2556–2565.