# Stochastically Dominant Preference Optimization: Policy Improvement for All

Extended Abstract

Ali Farajzadeh
University of Illinois Chicago
Chicago, Illinois, United States
afaraj5@uic.edu

Syed M. Abbas
University of Illinois Chicago
Chicago, Illinois, United States
sabbas33@uic.edu

Aadirupa Saha
University of Illinois Chicago
Chicago, Illinois, United States
aadirupa@uic.edu

Brian D. Ziebart
University of Illinois Chicago
Chicago, Illinois, United States
bziebart@uic.edu

## ABSTRACT

Reinforcement learning from human feedback (RLHF) optimizes policies based on users' rankings of output samples rather than using user-provided rewards. These methods typically assume users' underlying utility functions are homogeneous and their rankings differ only due to noise, ultimately optimizing for the average user. Instead, we seek policies that guarantee improvement for all users with respect to their heterogeneous preferences. We introduce stochastic dominance as a stricter guiding criteria for policy optimization that guarantees improvement under any social welfare function. Our approach, stochastically dominant preference optimization (SDPO), avoids explicit reward function estimation while providing individual performance improvement guarantees for users with diverse preferences.

## KEYWORDS

RLHF, policy improvement, stochastic dominance

**ACM Reference Format:**
Ali Farajzadeh, Syed M. Abbas, Aadirupa Saha, and Brian D. Ziebart. 2026. Stochastically Dominant Preference Optimization: Policy Improvement for All: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/

## INTRODUCTION

RLHF [7] improves policies $\pi_0$ using users' rankings $\{\sigma_j\}_{j=1:M}$ of output samples $\{y_i\}_{i=1:N}$ [1, 2, 5, 10, 11, 17, 18, 20]. Many RLHF methods assume a singular reward function $r : \mathcal{Y} \to \mathbb{R}$ motivates users' noisy rankings under the Bradley-Terry model of probabilistic preferences [4] and either estimate the reward function explicitly [7] or implicitly [16]. However, when feedback is from different users with distinct reward functions [14], existing methods provide few formal guarantees and the preferences of the majority can override other users' preferences [6, 9].



*Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus.* © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/

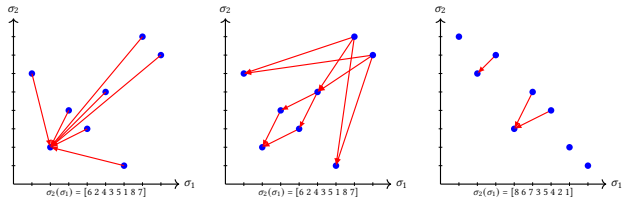


**Figure 1: Given two rankings over outputs, existing methods typically seek a single "compromise" output (left). Stochastic dominance instead allows limited policy probability shifts between outputs when there is moderate (center) or low agreement (right).**

In this paper, we introduce stochastically dominant preference optimization (SDPO), a method that seeks policy improvement for all users without requiring explicit reward estimation. Given an initial policy and a set of user rankings, SDPO learns a new policy that stochastically dominates [8, 13] the initial policy for every user, shifting probability mass from less preferred to more preferred outputs with a consensus in rankings (Figure 1). We integrate this training objective into a REINFORCE-style reinforcement learning process [19] and demonstrate the benefits of SDPO on a common imitation learning environment: `Point Bot`. We find that compared to baseline methods, SDPO succeeds in preserving diverse high-quality behaviors rather than optimizing for average performance.

## APPROACH

Our approach (SDPO) provides **performance improvement guarantees for each user** based only on output rankings. We focus on the broader distributional properties of the policy, rather than moment statistics (i.e., expected rewards), to make the performance improvement guarantees for multiple users feasible.

*Ranking-Based Stochastic Dominance.* We seek performance guarantees for users' underlying reward functions. We broaden our notion of policy improvement to hold given reward function uncertainty. Stochastic dominance (Definition 1) provides this for all ranking-consistent rewards.

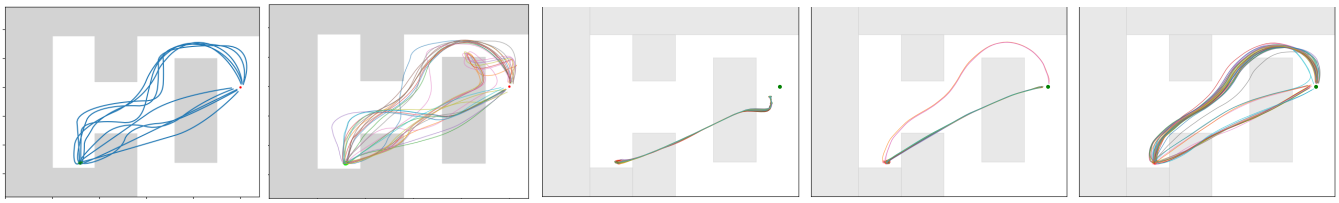

**Figure 2: `Point Bot` demonstrated trajectories (left) originating in the bottom left point and reaching the top right red point. These demonstrations are not available to the learner. Sample trajectories from a behavioral cloned policy (left center) that is provided to the learner as an initial reference policy from which to seek improvement. 50 sample trajectories from the reward model (RM) policy (center), the self-play preference optimization (SPO) policy (right center), and the stochastically dominant preference optimization (SDPO) policy (right).**

Definition 1. *Policy $\pi$* **stochastically dominates** *policy $\pi_0$ given ranking $\sigma_j$ (denoted $\pi \succeq_{\sigma_j} \pi_0$) if the expected reward for all ranking-consistent reward functions $r_{\sigma_j}$ is no smaller: $\mathbb{E}_{y\sim\pi}[r_{\sigma_j}(Y)] \geq \mathbb{E}_{y\sim\pi_0}[r_{\sigma_j}(Y)]$.*

This guarantees policy $\pi$ is at least as desirable to the user as $\pi_0$. Proposition 2 characterizes how the initial policy $\pi_0$ can be modified to $\pi$ to provide potential improvements that are stochastically dominant given multiple user rankings, $\forall j, \pi \succeq_{\sigma_j} \pi_0$.

Proposition 2. *If $\forall j \in [J], \sigma_j^{-1}(y) \geq \sigma_j^{-1}(y')$, then shifting probability from $y$ to $y'$ maintains stochastic dominance.*

*Policy Model Optimization.* We obtain target output distributions over a set of samples $\pi^*_{\{y\}}$ by iterating over random orders of rankers and reassigning probability to each ranker's best-ranked output from outputs worse under all rankings. Using this target distribution, our policy model optimization of $\pi_\theta$ has the form: $\max_\theta \mathbb{E}_{\{y\}\sim\pi_0}\left[\sum_{i=1}^N \pi^*_{\{y\}}(y_i)\log\pi_\theta(y_i)\right]$. We optimize this objective using a REINFORCE-style policy gradient algorithm [19] encouraging trajectories with higher target probability $\pi^*_{\{y\}}$.

## EXPERIMENTS

We evaluate policy fine-tuning given two rankers with distinct preferences using the `Point Bot` environment [12], which requires applying continuous cardinal forces to navigate a robot in a two-dimensional space from start to goal state. Both prefer shorter trajectories, but one is indifferent to gray regions, while the other seeks to avoid them. Our initial policy results from behavioral cloning (BC) [15] demonstrations that reflect these two preferences.

*Baseline methods.* **Reward model (RM)** is an idealized baseline that uses the ground truth rewards of each ranker to train a policy using a weighted combination of the rewards for policy optimization and uses entropy regularization to increase trajectory diversity. **Self-play preference optimization (SPO)** does not assume an underlying reward function and uses preferences directly, making it suitable for heterogeneous preferences. We perform pairwise evaluations using the rankers to guide soft actor-critic policy optimization of the initial BC policy based on self-play evaluation using a replay buffer of size 1000.

***Evaluation measures.*** **User reward improvement** measures the average improvement relative to the initial policy using each ranker's underlying reward, defined in terms of standard deviations of the original policy's reward; **Utilitarian Social Welfare (Util) improvement** [3] measures the average improvement relative to the initial using an equal mixture of each ranker's reward and defined in terms of standard deviations of the original policy's reward; **Stochastic dominance:** for 100 improved policy samples and 100 initial policy samples, how many can be paired (one-to-one) so that the ranker prefers the improved policy samples in each pair; and **Pairwise win rate** measures how frequently an improved policy sample is preferred over an initial policy sample by a ranker.

| | $R_1$ | $R_2$ | Util | Stoch. $R_1$ | Stoch. $R_2$ | Pair $R_1$ | Pair $R_2$ |
|---|---|---|---|---|---|---|---|
| RM | 0.843 | -0.404 | 0.549 | 79.0% | 36.4% | 78.7% | 36.2% |
| SPO | 1.800 | -0.223 | 1.772 | 100.0% | 41.0% | 96.6% | 40.8% |
| SDPO | 0.620 | 0.438 | 1.090 | 100.0% | 98.8% | 69.2% | 72.1% |

**Table 1: Reward improvement (standard deviations), stochastic dominance, and pairwise dominance for each method.**

*Results.* As Table 1 shows, only SDPO consistently improves relative performance across rankers $R_1$, $R_2$, and Util simultaneously. RM and SPO each sacrifice the reward of one ranker in favor of the other. SDPO reliably optimizes stochastic dominance, while SPO and RM are unable to do so consistently. Figure 2 shows that qualitatively, all methods produce more goal-directed trajectories than the BC policy. However, RM completely collapses to a single mode, while SPO produces trajectories of both modes (avoid/ignore obstacles), albeit avoiding obstacles infrequently. In contrast, SDPO trajectories provide coverage of each mode and a range of trade-offs in between.

## CONCLUSIONS

In this paper, we have developed an approach for policy improvement that given distinct user rankings, guarantees improvement for every single ranker and any corresponding non-decreasing social welfare function. We accomplish this by replacing traditional utility maximization as the basis for characterizing policy improvement with stochastic dominance. In contrast to existing approaches that explicitly or implicitly identify a single unifying reward function to optimize, our approach provides improvement guarantees for each individual ranker under any reward function consistent with their provided rankings.

## REFERENCES


[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Riad Akrour, Marc Schoenauer, and Michèle Sebag. 2012. April: Active preference learning-based reinforcement learning. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 116–131.

[3] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems* 35 (2022), 38176–38189.

[4] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345.

[5] Maya Cakmak, Siddhartha S Srinivasa, Min Kyung Lee, Jodi Forlizzi, and Sara Kiesler. 2011. Human preferences for robot-human hand-over configurations. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1986–1993.

[6] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217* (2023).

[7] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[8] Ali Farajzadeh, Danyal Saeed, Syed M Abbas, Rushit N. Shah, Aadirupa Saha, and Brian D Ziebart. 2025. Imitation Beyond Expectation Using Pluralistic Stochastic Dominance. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[9] Michael Feffer, Hoda Heidari, and Zachary C Lipton. 2023. Moral machine or tyranny of the majority?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 5974–5982.

[10] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375* (2022).

[11] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[12] Zaynah Javed, Daniel S Brown, Satvik Sharma, Jerry Zhu, Ashwin Balakrishna, Marek Petrik, Anca Dragan, and Ken Goldberg. 2021. Policy gradient Bayesian robust optimization for imitation learning. In *International Conference on Machine Learning*. PMLR, 4785–4796.

[13] Haim Levy. 1992. Stochastic dominance and expected utility: Survey and analysis. *Management science* 38, 4 (1992), 555–593.

[14] Barbara Plank. 2022. The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. *arXiv preprint arXiv:2211.02570* (2022).

[15] Dean A Pomerleau. 1988. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems* 1 (1988).

[16] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36 (2024).

[17] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33 (2020), 3008–3021.

[18] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[19] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3 (1992), 229–256.

[20] Matt Zucker, Nathan Ratliff, Martin Stolle, Joel Chestnutt, J Andrew Bagnell, Christopher G Atkeson, and James Kuffner. 2011. Optimization and learning for rough terrain legged locomotion. *The International Journal of Robotics Research* 30, 2 (2011), 175–191.