Offline Safe Policy Optimization From Heterogeneous Feedback

Extended Abstract

Ze Gong
Shenzhen Institute of Advanced Technology (SIAT), CAS
Shenzhen, China
ze.gong@siat.ac.cn

Pradeep Varakantham*
Singapore Management University
Singapore, Singapore
pradeepv@smu.edu.sg

Akshat Kumar*
Singapore Management University
Singpaore, Singapore
akshatkumar@smu.edu.sg

ABSTRACT

Offline Preference-based Reinforcement Learning (PbRL) enables reward and policy learning from human preferences without extensive reward engineering and direct interaction with human annotators. However, ensuring safety remains a critical challenge, especially in long-horizon continuous control tasks. Previous works on safe RL from human feedback (RLHF) typically learn reward and cost models before applying constrained RL, which can lead to compounding errors and suboptimal performance. To address these challenges, (a) instead of indirectly learning policies (from rewards and costs), we introduce a framework that learns a policy directly based on pairwise preferences regarding the agent's behavior in terms of rewards, as well as binary labels indicating the safety of trajectory segments; (b) we propose PreSa (Preference and Safety Alignment), a method that combines preference learning module with safety alignment into a single objective, optimized via a Lagrangian paradigm that directly learns reward-maximizing safe policy *without explicitly learning reward and cost models*, avoiding the need for constrained RL. Experiments on continuous control tasks with both synthetic and real human feedback demonstrate that PreSa consistently learns safe policies with high rewards, outperforming state-of-the-art baselines, and offline safe RL approaches with ground-truth reward and cost. A full version of this paper is available at: https://arxiv.org/abs/2512.20173.

KEYWORDS

Offline Safe RL; Preference-based RL

ACM Reference Format:
Ze Gong, Pradeep Varakantham, and Akshat Kumar. 2026. Offline Safe Policy Optimization From Heterogeneous Feedback: 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/XGRR7655

1 INTRODUCTION

To align the intelligent agents with human values, preference-based reinforcement learning (PbRL) [2, 11, 14, 16, 21] (also known as Reinforcement Learning from Human Feedback (RLHF) [1, 4, 15, 17, 20, 22, 23]) has emerged as a powerful learning paradigm by training the agent's policy from human pairwise preference over agent behaviors, without the need for explicit reward engineering. Offline PbRL [6, 7, 9, 19] further improves feedback efficiency by avoiding costly online interactions with human annotators. However, preference alignment alone is insufficient in safety-critical tasks. Agents must also be explicitly aligned with human safety considerations. Recently, safe RLHF [3] addresses this issue by learning both reward and cost models from human feedback and then applying constrained reinforcement learning. While effective for contextual bandit problems such as language model finetuning, this paradigm is less suitable for long-horizon continuous control due to compounding model errors and the instability of constrained RL.

To overcome these limitations, we introduce the framework of *Offline Safe Policy Optimization from Heterogeneous Feedback (Offline Safe POHF)*, in which a policy is learned directly from offline datasets that include two types of feedback: (a) *pairwise preferences* over agent behaviors with respect to rewards, and (b) *binary safety labels* indicating whether each trajectory segment is safe or not. Notably, unlike prior work [3], our approach does not require pairwise preferences over the agent behaviors on cost, as such annotations are often scarce and costly to collect in practice [4]. To address the aforementioned challenges, we propose PreSa (Preference and Safety Alignment), a novel algorithm that learns policies directly from heterogeneous feedback without explicit reward or cost modeling, and integrates preference and safety alignment into a single objective. Our key insight is that safety alignment can be formulated as a feasibility constraint over policies, while preference alignment provides the optimization objective. The resulting problem can be effectively solved using the Lagrangian method directly on the offline dataset. Consequently, we derive a fully supervised learning objective that avoids explicit reward and cost model learning, as well as the conventional constrained RL phase.

The contributions of this paper are threefold. First, we introduce the *Offline Safe POHF* framework for continuous control tasks, enabling safe policy learning from offline preference and safety feedback. Second, we propose PreSa, a unified optimization formulation that integrates preference and safety alignment through a constrained objective, without reward and cost modeling, as well as additional constrained RL stage. Third, we conduct comprehensive evaluations on both synthetic and real human feedback datasets, demonstrating that PreSa achieves effective alignment and robust safety performance in continuous control tasks.

2 PROBLEM DEFINITION

Consider a Constrained Markov Decision Process (CMDP) with unknown reward and cost functions, the *Offline Safe POHF* problem

*Equal advising.

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/XGRR7655

is defined by an offline dataset, $D = \{(\sigma^+, y^+, \sigma^-, y^-)\}$, where each σ is a trajectory segment. The dataset contains:

(a) *Pairwise preferences*: $\sigma^+ \succ \sigma^-$, indicating that σ^+ is preferred to σ^- with respect to reward; and
(b) *Binary safety labels*: $y \in \{-1, +1\}$, denoting whether the segment is safe or unsafe. y^+ and y^- are the safety labels associated with σ^+ and σ^-, respectively.

The goal is to learn a policy that maximizes expected reward while satisfying safety constraints, relying solely on the feedback information contained in D.

3 METHOD (PRESA)

PreSa consists of two components: preference alignment and safety alignment, which are integrated into a single objective.

3.1 Preference Alignment

We adopt a contrastive preference learning objective [6] that directly learns a policy from pairwise trajectory preferences. Given a preferred segment σ^+ and a less preferred segment σ^-, the probability of preference $P_{\pi_\theta}[\sigma^+ \succ \sigma^-]$ is modeled using a softmax over cumulative log-likelihood ratios under the policy [6, 10]. The policy is optimized to maximize the likelihood of observed preferences: $L_{\text{pref}}(\pi_\theta, D) = \mathbb{E}_{(\sigma^+,\sigma^-)\sim D}\left[-\log P_{\pi_\theta}[\sigma^+ \succ \sigma^-]\right]$. This loss function presents a closed-form formulation for directly learning a policy that aligns with human preferences.

3.2 Safety Alignment

Safety alignment treats safety labels as supervision signals over trajectory segments. Each segment is assigned a utility score $u_\theta(\sigma)$ based on its log-probability under the policy relative to a reference policy. The safety objective encourages higher scores for safe segments and lower scores for unsafe ones:

$$L_{\text{safety}}(\pi_\theta, D) = \mathbb{E}_{(\sigma, y_\sigma)\sim D}\left[1 - \texttt{sigmoid}\left(y_\sigma \cdot u_\theta(\sigma)\right)\right], \quad (1)$$

where y_σ denotes the safety label associated with segment σ. We observe that, in the context of a typical binary classification problem, minimizing the loss function in Equation (1) is equivalent to maximizing the probability of correctly classifying each segment with respect to safety. This classification-like objective determines which segments are safe, and which are unsafe. Consequently, this objective induces a feasible policy set: $\Pi = \{\pi | p(Y = y_\sigma | \sigma; \pi) \geq \delta, \forall \sigma\}$, where δ is a predefined parameter that controls the stringency with which we accept a segment as being correctly classified according to the safety labels provided by humans.

3.3 Unified Objective of PreSa

We integrate both preference alignment and safety alignment components into a single constrained optimization problem:

$$\min_{\pi_\theta} L_{\text{pref}}(\pi_\theta, D) \quad \text{s.t.}, \mathbb{E}_{(\sigma, y_\sigma)\sim D}\left[p\left(Y = y_\sigma | \sigma; \pi_\theta\right)\right] \geq \delta \quad (2)$$

To solve this constrained optimization problem, we employ the Lagrangian method to convert the constrained primal problem into an unconstrained dual form:

$$\min_{\pi_\theta} \max_{\nu \geq 0} L_{\text{pref}}(\pi_\theta, D) + \nu \cdot \left(\delta - \mathbb{E}_{(\sigma, y_\sigma)\sim D}\left[p\left(Y = y_\sigma | \sigma; \pi_\theta\right)\right]\right) \quad (3)$$

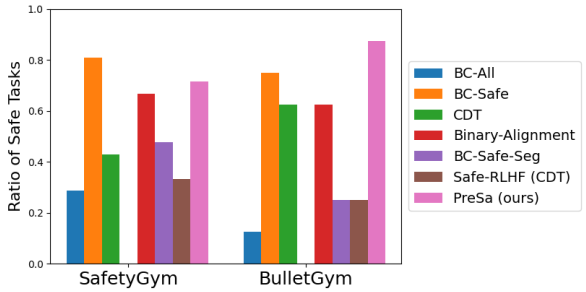

Figure 1: Ratio of safe agents learned by different approaches.

where $\nu \geq 0$ is the Lagrange multiplier. This formulation directly learns a policy that balances reward preference and safety compliance, without explicit reward and cost modeling and a separate constrained RL stage.

4 EXPERIMENTS

We evaluate PreSa on continuous control benchmarks derived from SafetyGym [8, 18], BulletGym [5], and MetaDrive [13]. Synthetic human feedback is generated from offline datasets by deriving preferences from cumulative rewards and safety labels from ground-truth cost thresholds. We also conduct experiments with real human feedback in autonomous driving scenarios [12]. We compare PreSa against: 1) Offline Safe RL baselines using ground-truth reward and cost; 2) Offline Safe POHF baselines such as behavior cloning on safe segments and Safe RLHF adapted to continuous control tasks.

Results. Across tasks, PreSa consistently learns policies that satisfy safety constraints while achieving higher rewards than competing Offline Safe POHF methods. Notably, PreSa matches or exceeds the safety performance of offline safe RL baselines that have access to ground-truth cost functions, despite relying only on human feedback, as shown in Figure 1. With real human feedback in driving tasks, PreSa achieves higher task completion scores and a higher proportion of safe behaviors compared to Safe RLHF and preference-only baselines.

Notably, the Offline Safe RL baselines have access to ground truth data, while PreSa relies solely on human feedback which has much less information. Despite this, PreSa matches or exceeds the performance of these baselines for safety alignment, highlighting its effectiveness.

5 CONCLUSION

In this paper, we introduce Offline Safe POHF, a framework for learning policies using human pairwise preferences and binary safety labels for each trajectory segment, without access to ground truth rewards or costs. We propose PreSa, which integrates preference alignment and safety alignment into a unified constrained optimization objective, as the safety alignment module defines a feasible policy set. PreSa learns a policy directly from offline human feedback, without the needs for reward or cost models or constrained RL. Empirical results with synthetic and real human feedback show that PreSa outperforms Offline Safe POHF baselines as well as matches or surpasses offline safe RL methods with ground truth rewards and costs.

ACKNOWLEDGMENTS

This research/project is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (Award Number: AISG2-RP-2020-016).

REFERENCES

[1] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).

[2] 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).

[3] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe RLHF: Safe Reinforcement Learning from Human Feedback. In *The Twelfth International Conference on Learning Representations*.

[4] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306* (2024).

[5] Sven Gronauer. 2022. Bullet-safety-gym: A framework for constrained reinforcement learning. (2022).

[6] Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W Bradley Knox, and Dorsa Sadigh. 2024. Contrastive Preference Learning: Learning from Human Feedback without Reinforcement Learning. In *The Twelfth International Conference on Learning Representations*.

[7] Joey Hejna and Dorsa Sadigh. 2024. Inverse preference learning: Preference-based rl without a reward function. *Advances in Neural Information Processing Systems* 36 (2024).

[8] Jiaming Ji, Jiayi Zhou, Borong Zhang, Juntao Dai, Xuehai Pan, Ruiyang Sun, Weidong Huang, Yiran Geng, Mickel Liu, and Yaodong Yang. 2023. Omnisafe: An infrastructure for accelerating safe reinforcement learning research. *arXiv preprint arXiv:2305.09304* (2023).

[9] Yachen Kang, Diyuan Shi, Jinxin Liu, Li He, and Donglin Wang. 2023. Beyond reward: offline preference-guided policy optimization. In *Proceedings of the 40th International Conference on Machine Learning*. 15753–15768.

[10] W Knox, S Hatgis-Kessell, S Booth, S Niekum, P Stone, and A Allievi. 2024. Models of Human Preference for Learning Reward Functions. *Transactions on Machine Learning Research* (2024).

[11] Kimin Lee, Laura M Smith, and Pieter Abbeel. 2021. PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training. In *International Conference on Machine Learning*. PMLR, 6152–6163.

[12] Edouard Leurent. 2018. An Environment for Autonomous Driving Decision-Making. https://github.com/eleurent/highway-env.

[13] Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. 2022. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence* 45, 3 (2022), 3461–3475.

[14] Runze Liu, Fengshuo Bai, Yali Du, and Yaodong Yang. 2022. Meta-reward-net: Implicitly differentiable reward learning for preference-based reinforcement learning. *Advances in Neural Information Processing Systems* 35 (2022), 22270–22284.

[15] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[16] Jongjin Park, Younggyo Seo, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. 2022. SURF: Semi-supervised Reward Learning with Data Augmentation for Feedback-efficient Preference-based Reinforcement Learning. In *International Conference on Learning Representations*.

[17] 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).

[18] Alex Ray, Joshua Achiam, and Dario Amodei. 2019. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708* 7, 1 (2019), 2.

[19] Daniel Shin, Daniel S Brown, and Anca D Dragan. 2021. Offline preference-based apprenticeship learning. *arXiv preprint arXiv:2107.09251* (2021).

[20] 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.

[21] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. 2017. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research* 18, 136 (2017), 1–46.

[22] Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425* (2023).

[23] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).