A Brief Literature Review and Summary of Reinforcement Learning with Human Feedback

Maya Gambhir

January 3, 2026

1 Abstract

Reinforcement Learning with Human Feedback (RLHF) has emerged as a dominant technique for aligning large language models with human preferences. However, the process of preference aggregation remains difficult in the face of proven existing tradeoffs. This paper examines recent advances that bring some structure to preference aggregation in RLHF through various frameworks. We review two complementary lines of work: an axiomatic approach to preference aggregation working parallel to Arrow's impossibility theorem, and a diversity-aware framework that explicitly models the existence of distinct demographic or subpopulations with differentiated values and preferences. The former introduces desirable criteria for aggregators, and proves an impossibility result for certain methods, demonstrating that some types of aggregators cannot simultaneously satisfy all criteria. The latter presents MaxMin-RLHF, an algorithm that seeks to ensure equitable treatment of subpopulations by minimizing worst-case regret across them. Together, these approaches highlight the fundamental trade-offs between fairness, representational diversity, and rational aggregation in preference-based reinforcement learning. We conclude by identifying some open questions in alignment and the design of aggregation procedures that balance inclusivity with learnability and robustness.

2 Reinforcement Learning

Reinforcement learning is a framework for solving control tasks by building agents that learn from their environment through interactions. An agent takes an action a which then induces a state s and a reward r. While the specifics can vary between models, the main objective is to take actions that maximize the expected reward over time, with reward calculated as some function of the state and action.

Often, RL is presented as a **Markov Decision Process** (MDP). This allows the model to rely solely on the current state in deciding the next action, rather than the list of all previous states. The agent performs actions, receives rewards, and transitions to new states as outlined in 1. We outline some main components of RL below:

Policy The agent's decision-making process, mapping states to actions. Policies can be modeled using neural networks.

Reward Signal Measures the desirability of actions, with discount rates emphasizing short-term rewards, as they are often more predictable and more important.

Value Function Estimates the expected discounted return of a state when following a specific policy. The value of a state is defined as the expected discounted return starting from that state.

Model Predicts environment behavior based on actions, aiding in strategic decision-making.

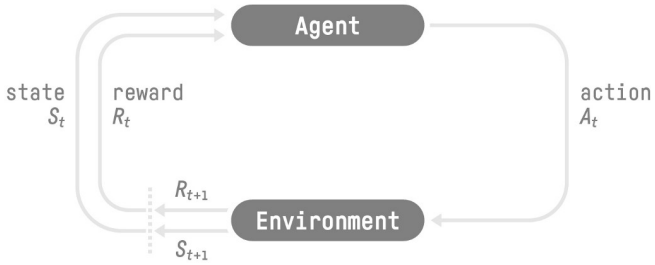

Figure 1: Via https://huggingface.co/learn/deep-rl-course/en/unit1/rl-framework

2.1 Q-Learning

Q-learning estimates the optimal action-value function $Q(s, a)$, representing the expected cumulative reward from state s after taking action a. The advantage function $A(s, a)$ captures the benefit of a specific action over others. **Policy based methods** train a policy (which is a neural network) directly, while **value based methods** train a value function to assign value to varying states, so that the agent may choose the actions that lead to those states. The value of a state is the expected discounted return/reward if the agent starts at that state and then follows some policy π.

3 Reinforcement Learning with Human Feedback (RLHF)

Oftentimes, an objective notion of reward and desirability is unattainable. Reinforcement Learning with Human Feedback leverages human preferences to guide RL models, transforming subjective and sometimes disparate human feedback into reward functions.

History As generative language models improved in utility and application, the question of alignment became paramount. Alignment ensures model responses are safe, factual and helpful to any given user. Previous attempts to tackle alignment involved pretraining a model and then engaging in supervised finetuning where the model is trained with a next token prediction objective over a dataset of annotated reference generations. While SFT is useful and broadly implemented in models today, this supervised training objective deviates from the actual goal of a model, which is to produce high quality and desirable outputs. Previous attempts to evaluate response quality, like ROUGE[Lin04] worked well for well defined tasks, but struggled to accurately reflect human preferences for more open ended ones. Thus, a process to train the models directly on human feedback was developed, first as an extension on the ROUGE score [BGM+19], and seen in its modern from in [OWJ+22] and [SOW+20].

3.1 Problem Formulation

Given initial states (prompts) and actions (completions), the objective is to maximize expected rewards:

$$J(\pi) = \mathbf{E}_{a,s}[r_0(s, a_t)]$$

where π is the policy, r_0 is the learned model of human preference, and s, a_t represent states and actions.

3.2 Annotation

In constrast to SFT, RLHF requires less annotation effort. For SFT, users must provide full high quality human-written responses while in RLHF, annotators simply rank existing responses. This leads to higher quality and more consistent responses as well, since the pool of possible annotations as compared to the pool of human written responses is much smaller.

3.3 Phases of RLHF in LLMs

Pre-training Starts with a pre-trained model.

Supervised Fine-Tuning (SFT) Optimizes model responses to user prompts.

Reward Model Training Utilizes human feedback to assign scalar rewards, often using pairwise preferences. Various regularization techniques are applied to improve robustness.

Policy Optimization Techniques like Proximal Policy Optimization (PPO) adjust the policy to balance utility and reward.

4 The Nature of Preferences and Linear Social Choice

It is important to discuss the sociological nature of human preferences as they relate to the annotation process for RLHF. Human preferences are still notably subjective, varying across time and contexts. Arrow's impossibility theorem highlights the challenges in aggregating these preferences, such as the difficulty of interpersonal comparison.

Arrow's Impossibility Theorem [Arr51] is a fundamental result in social choice theory. It states that no rank-order voting system can convert the ranked preferences of individuals into a community-wide ranking while simultaneously satisfying all of the following conditions:

1. **Unrestricted Domain (Universality):** The social welfare function should accept every possible set of individual preference orderings.
2. **Non-Dictatorship:** No single individual should possess the power to always determine the group's preference.
3. **Pareto Efficiency (Unanimity):** If every individual prefers alternative A over B, then the social preference should also rank A over B.
4. **Independence of Irrelevant Alternatives (IIA):** The social preference between any two alternatives A and B should depend only on individuals' preferences between A and B, not on their preferences involving other options.

Theorem (Arrow, 1951): If there are at least three alternatives and at least two voters, then no social welfare function can satisfy all four of the above conditions simultaneously.

This is a key consideration in applying a set of human preferences to create a satisfying single model for all users.

Preference Data Collecting representative preference data is difficult and costly. Preference data needs to be representative to avoid bias. Models like the Bradley-Terry model estimate the probability of one option being preferred over another:

$$P(i > j) = \frac{P_i}{P_i + P_j}$$

Our loss function for the reward model should satisfy these probability constraints.

Now we will look at a paper that outlines a parallel alignment difficulty to Arrow's Theorem and proposes a possible solution.

5 Axioms for Alignment from Human Feedback

This paper [GHM+24] by Ge et al. relies on axiomatic foundations to ensure consistent aggregation of preferences. While RLHF often aims to make a noisy estimate of a common ground truth, often humans have substantively different but valuable opinions and these opinions must be aggregated in a way that is is fair to the voters. This work aims to analyze whether existing aggregation algorithms satisfy certain axioms and whether better aggregation methods, with more axiomatic guarantees can be created.

This paper proposes an RLHF problem where candidates are identified by some known set of features and a linear reward function is created as a function of these features, designing a linear aggregation rule which creates a ranking induced by a linear function. This paradigm is known as *linear social choice*. The general arc of the work is to show how current loss-based methods fail the proposed axioms, and then designs a new linear aggregation rule known as *Leximax Copeland subject to PO*.

5.1 Core Axioms

The work proposes two core axioms it claims RLHF algorithms should satisfy.

Pareto Optimality (PO) This is achieved when, if option a is preferred over b universally (i.e. a is ranked higher than b in every voter ranking), a ranks higher in the final ranking. This draws distinct similarities to **Pareto Efficiency** in Arrow's Theorem.

Pairwise Majority Consistency (PMC) For any pair of candidates a, b, if the majority prefers a over b, the final ranking has a ranked above b.

5.2 The Linear Social Choice Model

Given a set of candidates c with associated feature vectors x_c, a reward function for candidates is induced by parameter vector θ as a dot product with the features. A parameter vector is *non-degenerate* if it induces a *feasible* ranking with no ties.

We then define a **parameter aggregation rule** as a function that takes in some profile/vector of voting rankings π and outputs a parameter vector s.t. r_π satisfies some desirable properties in the final ranking (such as the proposed axioms)

Notably, we look at $C1$ aggregation rules that output rankings that depend only on majority relationships.

5.3 Results

5.3.1 Loss Based Rules

The paper concludes that any loss-based aggregation rules that uses non-decreasing and convex loss function will fail the core axioms proposed.

Such a loss-based aggregation rule looks like the sum of some loss measure based on the number of times a is ranked above b, and the difference in their rewards based on some reward function r_θ. l could be as simple as binary cross entropy loss $l(x) = \ln(1 + e^x)$

$$L(\theta, \pi, l) = \sum_{a \neq b \in C} [\#\text{ of times a ranked above b}] l(r_\theta(b) - r_\theta(a))$$

PMC is rectified if we consider a majority based loss formulation where we replace [# of times a ranked above b] with a binary measure of whether a is ranked above b more often than not. However, it still fails PO.

This concludes with a theorem that all $C1$ linear rank aggravation rules fail PO.

5.3.2 Social Choice based Rules

Thus, a new aggregation rule must be proposed that satisfies our axioms. The proposed method combines Copeland scores with leximax strategies to improve axiomatic alignment:

- **Copeland Score:** Counts pairwise wins for each candidate.
- **Leximax Copeland:** Ranks candidates by maximizing the minimum score first, ensuring feasibility under parameter constraints.

LCPO (Leximax Copeland with Pareto Optimality) Imposes PO constraints on rankings to improve consistency. LCPO restricts rankings to avoid placing dominated alternatives above dominating ones.

LCPO satisfies PO and PMC as well as two new axioms: winner monotonicity and majority consistency.

1. **Majority Consistency** A candidate ranked first by the majority should be ranked first overall.
2. **Winner Monotonicity** Moving a candidate up in individual rankings should not harm their final aggregated ranking.

5.4 Conclusion

We see the value in the axioms this new scoring mechanism satisfies, however, it is notable that it may have some shortcomings, especially considering Arrow's theorem. Thus, while LCPO achieves important axiomatic guarantees like Pareto Optimality, Pairwise Majority Consistency, Majority Consistency, and Winner Monotonicity, it must necessarily compromise on other properties (e.g., IIA or full expressivity of preferences) due to this foundational impossibility. As such, the work makes a principled and practical tradeoff by selecting axioms that better align with the specific goals of RLHF, rather than aiming for an unattainable ideal.

This next paper tackles the inherent and important diversity of human preferences, and considers practical algorithmic solutions to maximize user satisfaction.

6 MaxMin-RLHF: Alignment with Diverse Human Preferences

[CQY+24]

6.1 Motivation

Alignment is a critical challenge in Reinforcement Learning from Human Feedback (RLHF). However, existing approaches often align with a single reward function, which fails to capture the diversity of human preferences. Some methods attempt to learn multiple reward functions but aggregate them arbitrarily, which does not adequately represent distinct preferences. A somewhat more effective approach is to design personalized models that account for diverse preferences explicitly, however this is relatively intractable and could perpetuate echo chambers of bias. This work highlights that single-utility RLHF does not sufficiently address preference diversity and proposes an alternative framework.

6.2 Contributions

The key contributions of this work include:

- An impossibility result demonstrating the fundamental limitations of aligning preferences using a single reward RLHF approach.
- The introduction of Max-Min RLHF, a method that models preferences as a mixture distribution, ensuring that diverse human preferences are better represented.
- An empirical study validating the effectiveness of the proposed Max-Min RLHF approach.

6.3 Background

We define the language model (LM) as mapping from vocabulary $\mathcal{V}$ to a policy π_θ, where:

$$Y \sim \pi_\theta(\cdot|X) \tag{1}$$

This formulation establishes the foundational model used for learning from human feedback.

6.4 Pipeline

The proposed approach follows a three-stage pipeline:

1. **Supervised fine-tuning**: The initial model is trained using supervised learning to establish a base policy π_{SFT}.
2. **Reward modeling**: A reward function R_θ is defined based on human preferences.
3. **Reinforcement learning fine-tuning**: The policy π_θ is updated using reinforcement learning to optimize alignment with the learned rewards.

6.5 Impossibility Result

To formally characterize the challenges in aligning diverse human preferences, we define the diversity metric as:

$$\text{Diversity}(i, j) = \text{TV}(p_i^{,} p_j^{)} \tag{2}$$

where TV represents the total variation distance between different subpopulation preferences. The probability distribution of outputs is denoted as:

$$p(y|x) = \gamma_1 p_1^* + \gamma_2 p_2^* \tag{3}$$

weighted by subpopulation preference sizes.

We establish a new minimization objective and derive a lower bound for reward mismatch:

$$|Q^* - Q^\theta| \tag{4}$$

which holds for a specific subpopulation. The alignment gap increases with greater diversity, particularly for smaller subpopulations with distinctive preferences. To mitigate this issue, we propose a substitute: the Max-Min RLHF framework.

6.6 Algorithms

The Max-Min RLHF approach is outlined in the following two algorithms:

6.6.1 Algorithm 1: Learning Diverse Preferences

1. Initialize policy π_θ and reward models for different subpopulations.
2. Collect feedback from diverse human subpopulations.
3. Train reward models separately for each subpopulation.
4. Construct a mixture of reward functions weighted by subpopulation proportions.
5. Optimize π_θ using reinforcement learning to balance rewards across subpopulations.

6.6.2 Algorithm 2: Max-Min Optimization for Policy Training

1. Initialize policy parameters and subpopulation reward models.
2. Define the objective as maximizing the minimum reward across subpopulations.
3. Perform policy updates using adversarial training to ensure robustness against preference shifts.
4. Iterate until convergence, ensuring diverse preferences are sufficiently represented.

6.7 Conclusion

This work highlights the fundamental limitations of single-reward RLHF in aligning diverse human preferences and introduces the Max-Min RLHF framework as an effective alternative. By modeling human preferences as a mixture distribution and optimizing policies with a max-min approach, we ensure better alignment with diverse human values. Empirical studies validate the proposed method, demonstrating its potential in real-world applications where preference diversity is critical.

7 Conclusion

It is clear that RLHF can and does enhance model alignment through structured integration of human feedback. The axiomatic nature of linear social choice provides a rigorous foundation, and also a framework for considering the tradeoffs a good model needs to consider. While diverse preferences can make RLHF difficult, work exists to combat this with different aggregation methods and a maxmin "patching" algorithm of poorly represented subgroups. One extension of this solution could be to consider finetuning "custom" models for groups or subgroups of users based on their preference data. If a practical tradeoff of efficiency to alignment exists in which "not too much" finetuning is required, this may be a useful appraoch.

Another point of future work in this area I find to be particularly interesting is the idea of RLHF robustness to jail-breaking. In other words, is there a problem setup in which some group or subgroup of annotators can achieve a more desirable ranking for themselves by misrepresenting their preferences? Intuitively, this does not seem to be a large issue at scale since individual preferences do not influence the final model too much. However, it is interesting to consider the possibility of a broader scale jailbreaking "attack" on a model using RLHF, as the consequences of a model producing misaligned or harmful information could be catastrophic.

References

[Arr51] Kenneth J. Arrow. *Social Choice and Individual Values*. Number 12 in Cowles Commission Monograph. John Wiley & Sons, New York, 1951.

[BGM+19] Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. Better rewards yield better summaries: Learning to summarise without references, 2019.

[CQY+24] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences, 2024.

[GHM+24] Luise Ge, Daniel Halpern, Evi Micha, Ariel D. Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. Axioms for ai alignment from human feedback, 2024.

[Lin04] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL 2004 Workshop on Text Summarization (WAS)*, Barcelona, Spain, 2004. Association for Computational Linguistics.

[OWJ+22] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[SOW+20] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc., 2020.