Axioms for AI Alignment from Human Feedback

Evi Micha

University of Southern California
evi.micha@usc.edu

The alignment of AI models with human values is widely recognized as a critical challenge in the development of safe and beneficial AI systems. Ensuring that these models operate in ways that reflect and respect human preferences, ethics, and social norms is essential for minimizing risks and maximizing their positive impact on society. Human feedback plays a key role in achieving alignment between AI models and human values. By incorporating input from people during training, AI systems can learn to better understand and reflect human preferences, ethical principles, and societal norms.

One widely used approach for achieving AI alignment is *reinforcement learning with human feedback (RLHF)*, which has been applied in various domains, including robotics and recommendation systems, and recently, has gained significant attention as a method for fine-tuning large language models (LLMs). The RLHF process involves training a reward model using a pre-trained LLM, which is then employed to fine-tune the existing LLM. Typically, human feedback is provided through pairwise comparisons between different options, and a reward function is estimated via maximum likelihood, assuming an underlying random utility model, such as the Bradley-Terry (BT) model.

The assumption of an underlying random utility model implies that humans share a common, unobservable ground truth, and the observed comparisons are merely noisy estimates of this true preference. Is this the "right" way of aggregating individual preferences towards a socially desirable reward function, however? To answer this question, we draw on *social choice theory*, a field that studies collective decison making through a mathematical lens. The maximum likelihood estimation approach is in line with a storied body of work that assumes that different human participants have preferences stemming from noisy estimation of a common ground truth, and the goal is to learn this ground truth as accurately as possible. But this is not the case when it comes to questions of AI alignment, where individuals can have legitimate differences of opinion rooted in different values or priorities.

We argue that when preferences are truly heterogeneous, the *axiomatic approach* which rose to prominence in social choice may be more suitable. This approach analyzes the desirability of aggregation methods through their satisfaction of certain axioms that capture notions of consensus, fairness, and economic efficiency. Specifically, we are interested in the axiomatic properties of aggregation methods that take ordinal preferences as input and output a reward function. We address the following two research questions: *What axioms are satisfied by aggregation methods used by existing RLHF algorithms? And are there alternative aggregation methods that offer stronger axiomatic guarantees?*

To evaluate different aggregation methods, we adapt fundamental axioms from social choice theory. The first is *Pareto optimality (PO)*, which requires that if a candidate a is ranked above candidate b in *every* input ranking, then the resulting ranking should rank a above b. This is seen as a basic requirement and is satisfied by *every* standard voting method in the classical setting. The second axiom is *pairwise majority consistency (PMC)*: If there exists a reward function that generates a ranking where, for each pair of candidates, a majority of voters agree with the ranking, then the resulting ranking should match that ranking. This axiom is an extension of *Condorcet consistency* to rankings, and is satisfied by some, but not all, standard voting methods in the classical setting.

We will first focus on a significant negative result of the BTL model, which is widely applied in practice. In particular, we will discuss that when the goal is to design a linear reward function, the BTL model fails both PMC and PO. This result suggests that the prevailing practice of RLHF is flawed from an axiomatic viewpoint. We will also generalize this negative result to a family of loss-based rules for cases where the loss function is weakly convex and non-decreasing or strictly convex.

In the light of the above negative result, we will ask whether there are different approaches that can satisfy these two natural axioms. We will discuss that there exists an aggregation rule that we call *leximax Copeland subject to PO* which satisfies both PO and PMC. Moreover, we will show that it also satisfies two additional ones, *majority consistency* and *winner monotonicity*. This result indicates that while there are methods that satisfy multiple desirable axioms, an algorithm that is widely applied in practice fails to satisfy any of them. Therefore, it may be time to reconsider which algorithms are used in the presence of diverse preferences.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.