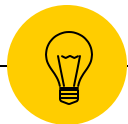


Axioms for AI Alignment from Human Feedback

Luise Ge¹, Daniel Halpern², Evi Micha², Ariel D. Procaccia², Itai Shapira², Yevgeniy Vorobeychik¹, Junlin Wu¹



NeurIPS 2024
Spotlight Presentation

¹ Washington University in St. Louis, ² Harvard University



Introduction

- ① AI alignment
- ② Reinforcement Learning with Human Feedback (RLHF)
- ③ Heterogeneous Preferences
- ④ Social Choice Theory

Our Model

- Set of “alternatives” / {prompt,response}

- Each alternative is associated feature vector $x_a \in \mathbb{R}^d$

- Dataset of pairwise comparisons

- Learn a parameterized reward

$$\hat{r}_\theta: \mathbb{R}^d \rightarrow \mathbb{R}$$

- Given a loss ℓ , minimize:

$$L(\theta) = \sum_{a \neq b} n_{a>b} \cdot \ell(r_\theta(b) - r_\theta(a))$$

- Set of participants/”voters”

- Each has their own unique reward

- Assume the reward model is **linear**:

$$\mathbb{H} = \{\langle \theta | \cdot \rangle \mid \theta \in \mathbb{R}^p\}$$

- understanding the relationship between individual rewards and the optimal reward



The Axiomatic approach

- ⊙ Axiomatic approach to study preference aggregation
- ⊙ **Pareto Optimality (PO)** If for every voter $r_{\theta_i}(a) \geq r_{\theta_i}(b)$ then
$$r_{\theta^*}(a) \geq r_{\theta^*}(b)$$
- ⊙ **Pairwise Majority Consistency (PMC)** If exist an ordering of the alternatives $c_1 > c_2 \dots > c_m$ such that c_i is preferred to c_j by a majority of voters whenever $i > j$, then $r_{\theta^*}(c_i) \geq r_{\theta^*}(c_j)$ if possible
- ⊙ Many others we can borrow from SC: monotonicity, majority consistency...



Results

Theorem. All reasonable loss-based aggregation method fail both PO and PMC

Theorem. There exist a linear aggregation rules that satisfy both PO and PMC



Discussion

- Growing Complexity Highlights RLHF's Limitations
- Need for Comparative Framework in Developing Robust Alignment Methods