Axioms for AI Alignment from Human Feedback

Luise Ge[1], Daniel Halpern[2], Evi Micha[2], Ariel D. Procaccia[2], Itai Shapira[2], Yevgeniy Vorobeychik[1], Junlin Wu[1]

NeurIPS 2024
Spotlight Presentation

[1] Washington University in St. Louis, [2] Harvard University

Introduction

- AI alignment
- Reinforcement Learning with Human Feedback (RLHF)
- Heterogeneous Preferences
- 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 compressions
- Learn a parameterized reward

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

- Given a loss ℓ, minimize:

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

- Set of participants/”voters”
- Each has their own unique reward
- Assume the reward model is **linear**:

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

- 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 $c_1 > c_2 \ldots > c_m$ 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