PhilBench: Measuring Value Learning from Text

Gaurav Sett
Georgia Institute of Technology
gauravsett@gatech.edu

Abstract

Current alignment efforts are largely focused on getting AI to follow common norms. Strong alignment requires AI to confront complex and controversial topics. Typical alignment approaches are incapable of representing a distribution of beliefs. To achieve a democratic approach to alignment, we must develop new methods. Collecting democratic inputs from humans at scale is expensive and difficult. However, humans have already provided significant information about their values through writing. We can formulate language in many ways, so the utterances we choose reveals what we think is true, informative, relevant, and coherent. We propose PhilBench, a benchmark for value learning from text. We provide a dataset of text from philosophy papers. We repurpose the PhilPapers Survey, measuring the views of a sample of authors of these texts, as a test for AI. We encourage researchers to develop methods allowing AI to learn the values of these philosophers from their papers.

1 Introduction

Motivation...

Insights from pragmatics. (Grice 1975)

Benchmark overview.

2 Data

2.1 Philosophy Papers

2.2 PhilPapers Survey

3 Experiments

4 Discussion

Social Impacts Statement

Potential broader impact of their work, including its ethical aspects and future societal consequences.

References

Grice, Herbert P (1975). “Logic and conversation”. In: *Speech acts*. Brill, pp. 41–58.

Preprint. Under review.