AI Alignment Via Power-Mean Elicitation

Extended Abstract

Chloé Becquey
University of Massachusetts Amherst
Amherst, United State
cbecquey@umass.edu

Cyrus Cousins
University of Massachusetts Amherst
Amherst, United State
originalcyruscousins@gmail.com

Yuhui Wei
University of Massachusetts Amherst
Amherst, United State
sophiawei28@gmail.com

Chang Zeng
University of Massachusetts Amherst
Amherst, United State
changzeng@umass.edu

Yair Zick
University of Massachusetts Amherst
Amherst, United State
yairzick@gmail.com

ABSTRACT

Both people and AI make decisions based on their innate/learned social values. These are often complex and difficult to elicit directly. We propose a comparison-query framework for eliciting welfare (or malfare) concepts that ε-approximately recovers the underlying power-mean concept using only a logarithmic number of queries in the search space. Experiments with large language models demonstrate that the approach converges rapidly and reveals informative differences in LLM social preferences.

KEYWORDS

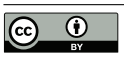
Preference Elicitation; Power Means; AI Alignment

ACM Reference Format:
Chloé Becquey, Cyrus Cousins, Yuhui Wei, Chang Zeng, and Yair Zick. 2026. AI Alignment Via Power-Mean Elicitation: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/UAUY5393

1 INTRODUCTION

Large Language Models (LLMs) [3, 9, 12] demonstrate exceptional capabilities in various natural language processing tasks, such as question-answering and reasoning; their decision-making processes are often guided by training on massive and diverse sources of data, as well as interactions with their users. Recent investigations show that these systems may also exhibit social biases in their responses [5, 7, 11]; a notable recent example is Grok, the AI chatbot used on the social media platform X. Following a software update, Grok started exhibiting alarming behavior, spewing hate speech and idolizing Adolf Hitler [4][1]. Were there any underlying social values that Grok was following at the time? Prior efforts to reduce model bias focus on fine-tuning models via specific datasets. Recent works [10, 13] indicate that LLMs can adopt various personas based on user instructions. However, the fundamental question remains: do LLMs maintain consistent social values?

[1]the issue has since been fixed.



This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/UAUY5393

Social dilemmas are a well-studied method of recovering social preferences: users are presented with a sequence of questions, each asking them to rank two alternatives [1, 2, 8]. We can now apply these methods to recover the social preferences of LLMs. Our goal is to design an economically grounded and algorithmically efficient method that

elicit social justice concept via iterated social dilemmas.

Our Contributions. We propose a framework for *malfare concept elicitation* via a sequence of comparison queries, formulating elicitation as a search task in which a decision maker selects preferred options. For power-mean malfare concepts, our method achieves logarithmic query complexity, establishing matching upper and lower bounds of $\Theta(\log \frac{\log p/q}{\varepsilon})$ queries to obtain an ε-accurate estimate when the true power mean value p^* is in the interval $[p, q]$. We further present a lightweight sampling scheme for generating binary social dilemmas and demonstrate experimentally, across several LLM frameworks, that the elicitation procedure converges rapidly and offers insights into the social preferences of various LLM frameworks, highlighting the practical effectiveness of our approach.

2 MALFARE POWER-MEAN ELICITATION VIA APPROXIMATE SEARCH

A *decision-maker* (or *actor*) evaluates outcomes using some *loss* (or *malfare*) function $\mathcal{M}^* : O \to \mathbb{R}_{\geq 0}$, where each *outcome* $o \in O$ induces disutilities $s_g(o) \geq 0$ for entities $G = \{1, \ldots, n\}$; when the outcome is clear from context, we write s_g for brevity. We assume that $\mathcal{M}^*$ belongs to the weighted power-mean family

$$\mathcal{M}^*(o) = \mathrm{M}_{p^*}(\boldsymbol{s}; \boldsymbol{w}) = \sqrt[p^*]{\sum_{g \in G} w_g s_g^{p^*}},$$

where the weight vector $\boldsymbol{w}$ is known and the aggregation parameter $p^* \geq 1$ is unknown. The elicitation task therefore reduces to identifying the latent parameter p^* using adaptive outcome comparisons.

We introduce the *Approximate Search with Binary Queries* (ASBQ) framework, a extended version of binary search to elicit a malfare concept $\hat{\mathcal{M}}$ that ε-approximates $\mathcal{M}^*$ the decision-maker's *true malfare concept*. Due to the nonlinear relationship between p and $\mathrm{M}_p(\boldsymbol{s}; w)$, rather use ℓ_1 distance between p, we define the supremum distance metric to gauge the difference between power mean concepts.

Definition 2.1 (Supremum Distance on Malfare Concepts)**.** Given two malfare functions $\mathcal{M}(\cdot)$ and $\mathcal{M}'(\cdot)$, their *supremum distance* is

$$\Delta\left(\mathcal{M}(\cdot), \mathcal{M}'(\cdot)\right) \doteq \sup_{o \in O} |\mathcal{M}(o) - \mathcal{M}'(o)|.$$

If both $\mathcal{M}$ and $\mathcal{M}'$ are power means with the same weight vector $\boldsymbol{w}$, i.e., $\mathcal{M}(\cdot) = \mathrm{M}_p(\cdot;\boldsymbol{w})$ and $\mathcal{M}'(\cdot) = \mathrm{M}_q(\cdot;\boldsymbol{w})$, we simply write $\Delta(p, q;\boldsymbol{w}) \doteq \Delta(\mathrm{M}_p(\cdot;\boldsymbol{w}), \mathrm{M}_q(\cdot;\boldsymbol{w}))$. Let W denote the set of all potential partial sums for the weights vector $\boldsymbol{w}$, i.e., $W \doteq \{\sum_{i \in I} w_i \mid I \subseteq \{1, 2, \ldots, g\}\}$. Given two power mean values $p \geq q \geq 1$ and a weight vector $\boldsymbol{w}$, the supremum distance between power mean values is characterized by

$$\Delta(p, q;\boldsymbol{w}) = \sup_{w \in W} \sqrt[p]{w} - \sqrt[q]{w}. \tag{1}$$

Since the supremum operation is subadditive rather than linear, bounding the required number of search queries (the *query complexity*) is challenging. To address this issue, we introduce the *Additive Supremum Distance Measure* $\Delta_{\uparrow}$, which upper bounds the supremum distance Δ.

Definition 2.2 (Additive Supremum Distance Measures)**.** Let $\bar{\mathbb{R}}$ be $\mathbb{R} \cup \{\pm\infty\}$. A function $\Delta_{\uparrow}(\cdot, \cdot;\boldsymbol{w}) : \bar{\mathbb{R}} \times \bar{\mathbb{R}} \times \triangle_n \to \bar{\mathbb{R}}_{0+}$ is an *additive supremum distance bound function* if for all $\boldsymbol{w} \in \triangle_n$, the following properties hold:

(1) **Dominance**: $\Delta_{\uparrow}(p, q;\boldsymbol{w}) \geq \Delta(p, q;\boldsymbol{w})$;
(2) **Additivity**: $\Delta_{\uparrow}(p, q;\boldsymbol{w}) = \Delta_{\uparrow}(p, r;\boldsymbol{w}) + \Delta_{\uparrow}(r, q;\boldsymbol{w})$;
(3) **Symmetry**: $\Delta_{\uparrow}(p, q;\boldsymbol{w}) = \Delta_{\uparrow}(q, p;\boldsymbol{w})$.

The (unique) *additive supremum distance function* $\Delta^*_{\uparrow}(\cdot, \cdot;\boldsymbol{w})$ satisfies properties (1)–(3), as well as:

(4) **Minimality**: for any other function $\Delta'_{\uparrow}$ satisfying properties (1)–(3) above: $\Delta^*_{\uparrow}(p, q;\boldsymbol{w}) \leq \Delta'_{\uparrow}(p, q;\boldsymbol{w})$.

Let $N_\varepsilon(p, q;\boldsymbol{w})$ be the number of queries required to find an ε-approximation of M_{p^*} within the interval $[p, q]$, assuming that $p^* \in [p, q]$. We derive the following asymptotic bounds for the query complexity.

Theorem 2.3 (Minimax Query Complexity Bounds on Δ). *Given two power mean values $p \geq q \geq 1$, a weight vector $\boldsymbol{w}$ and $\varepsilon > 0$,*

$$\left\lceil \log_2 \left\lceil \frac{\Delta(p, q;\boldsymbol{w})}{2\varepsilon} - 1 \right\rceil \right\rceil \leq N_\varepsilon(p, q;\boldsymbol{w}) \leq \left\lceil \log_2 \left\lceil \frac{\Delta^*_{\uparrow}(p, q;\boldsymbol{w})}{2\varepsilon} \right\rceil \right\rceil.$$

In addition, there exists some ε_0 such that for all $\varepsilon \leq \varepsilon_0$:

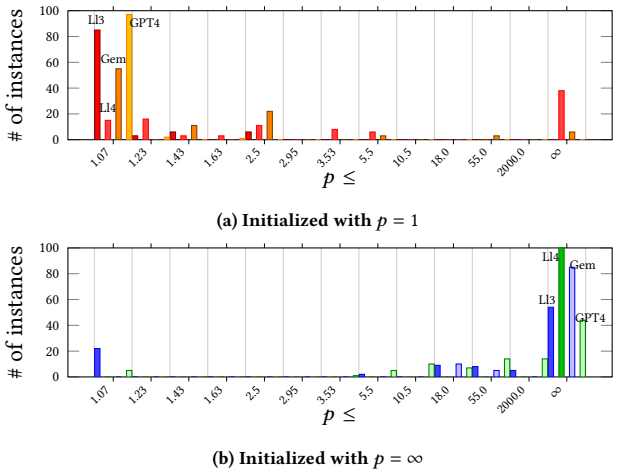
$$N_\varepsilon(p, q;\boldsymbol{w}) = \left\lceil \log_2 \left\lceil \frac{\Delta^*_{\uparrow}(p, q;\boldsymbol{w})}{2\varepsilon} \right\rceil \right\rceil \pm \Theta(1).$$

Furthermore, with $g \to \infty$ and uniform weights on each group:

$$N_\varepsilon\left(p, q; \left\langle \frac{1}{g}, \ldots \frac{1}{g} \right\rangle\right) = \Theta\left(\log \frac{\log(p/q)}{\varepsilon}\right).$$

3 EXPERIMENTS

We evaluate our ASBQ framework on several large language models (LLMs): GPT 4.1 (GPT-4.1-2025-04-14), Gemma 3 (Gemma-3-27b-it), and Meta's Llama 3 (Llama-3.3-70B-Instruct) and 4 (Llama-4-Scout-17B-16E-Instruct) with the sampling temperature to 0 for all models. We conduct 100 experiments with search initialized at $p = 1$ and $p = \infty$. In all of our experiments, we set the error parameter to be $\varepsilon = 10^{-4}$.



(a) Initialized with $p = 1$

(b) Initialized with $p = \infty$

Figure 1: p-value distribution across models. Each bar denotes the number of instances that converged to a p^* value that is at most the value indicated on the x axis. Ll3 denotes the Llama-3 model; Ll4 is the Llama-4 model; Gem is the Gemma-3 model, and GPT4 is the GPT-4.1 model.

For each experiment, we present the social dilemma with fix the group weights to the LLM and, given the LLM's responses, iteratively generate outcome pairs characterized by group disutilities. By varying the weights across runs, we probe the robustness of the LLM's preferences under different contextualizations of the same underlying scenario.

Our experimental design is guided by the following hypotheses:

(H1) LLMs converge to consistent malfare concepts when evaluated under the ASBQ elicitation framework.
(H2) Different LLMs exhibit similar malfare concepts.

On average, our ASBQ framework takes ≤ 10 queries to identify an approximate power mean within $\varepsilon = 10^{-4}$ error, and no more than 25 queries in total.

Results. Our key observation is that the p^* value that we converge to is initialization-dependent. In the disaster relief scenario, when we initialize the search with $p = 1$, all LLMs converge to approximately utilitarian power means (Figure 1a). However, when presented with the same scenario but initializing $p = \infty$, LLMs tend to be more egalitarian (Figure 1b). Intuitively, the scenarios we initially generate tend to make the LLMs 'stick' to their initial position, rather than align towards a consistent concept of justice. This could be explained by the sycophantic tendency of LLMs [6]; LLMs tend to appease users, adhering to positions they believe the user wants to hear. Addressing this issue requires a different approach to model evaluation; for example, injecting additional 'noise' into the scenario generation by asking the LLMs to answer a number of random social dilemmas, in order to avoid this anchoring effect. Our results rejects Hypothesis **(H2)**; LLMs exhibit significantly different behaviors across scenarios; for example, in the disaster scenario, Llama 3 and Gemma 3 take a more utilitarian approach than Llama 4, which tends more egalitarian.

REFERENCES

[1] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine experiment. *Nature* 563 (2018), 59–64.

[2] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576.

[3] Gemma-AI. 2025. Gemma 3 Technical Report.

[4] Dylon Jones. 2025. Why Grok Fell in Love With Hitler. *Politico* (2025).

[5] Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2025. Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception. In *Proceedings of the 31th International Conference on Computational Linguistics (COLING)*. 10634–10649.

[6] Lars Malmqvist. 2025. Sycophancy in Large Language Models: Causes and Mitigations. In *Proceedings of the 2025 Intelligent Computing: Computing Conference (CompCon)*. 61–74.

[7] Garg Nikhil, Schiebinger Londa, Jurafsky Dan, and Zou James. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[8] Ritesh Noothigattu, Snehalkumar 'Neil' S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. 2018. A voting-based system for ethical decision making. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*. 1587–1594.

[9] OpenAI. 2025. ChatGPT (4.1).

[10] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role-Play with Large Language Models. *Nature* 623 (2023), 493–498.

[11] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic Biases in LLM Simulations of Debates. In *Proceedings of the 29th Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 251–267.

[12] Llama 3 Team. 2024. The Llama 3 Herd of Models.

[13] Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. In *Proceedings of the 62th Finding of the Association for Computational Linguistics (ACL)*. 14743–14777.