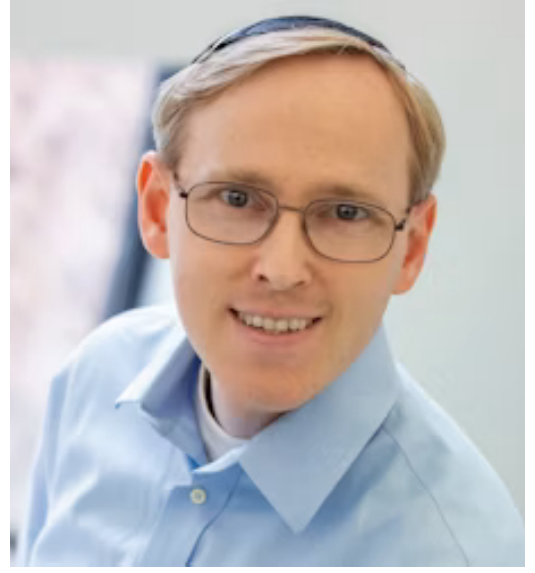
Linear Social Choice with Few Queries: A Moment-Based Approach

Luise Ge*, Daniel Halpern#, Gregory Kehne*, Yevgeniy Vorobeychik*

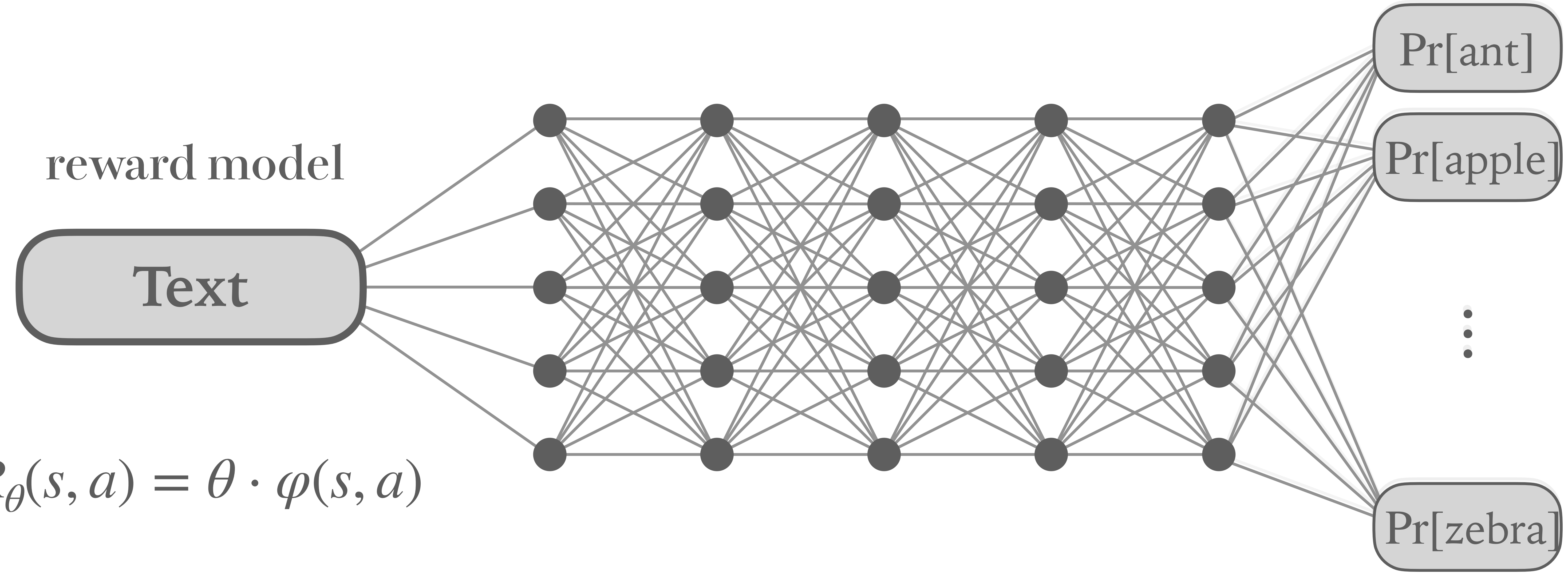
Washington University in St. Louis*
Google Research → Toronto#

What is Linear Social Choice?

Axioms for AI Alignment From Human Feedback [Ge et. al. 2024]

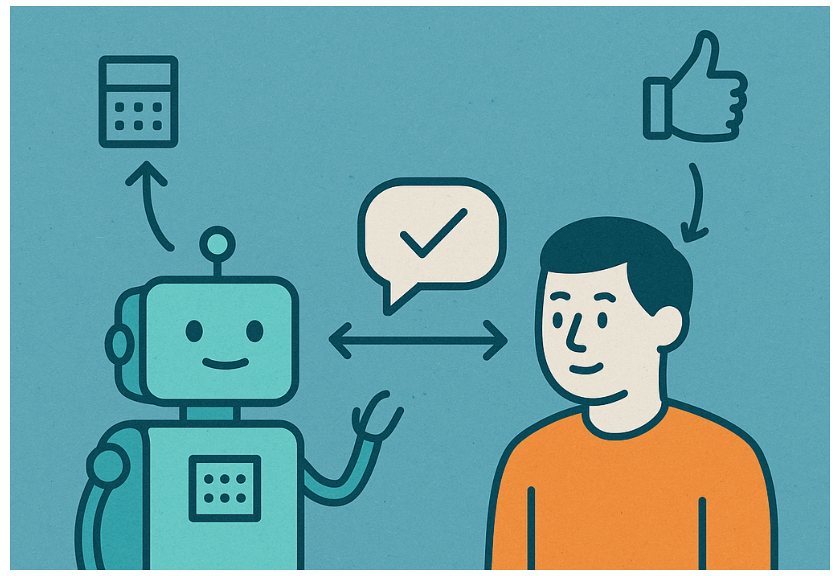
Each voter v and each candidate c has an embedding in $\mathbb{R}^d$

Each voter's utility is parametric (specifically linear): $u_v(c) = \theta_v^T c$

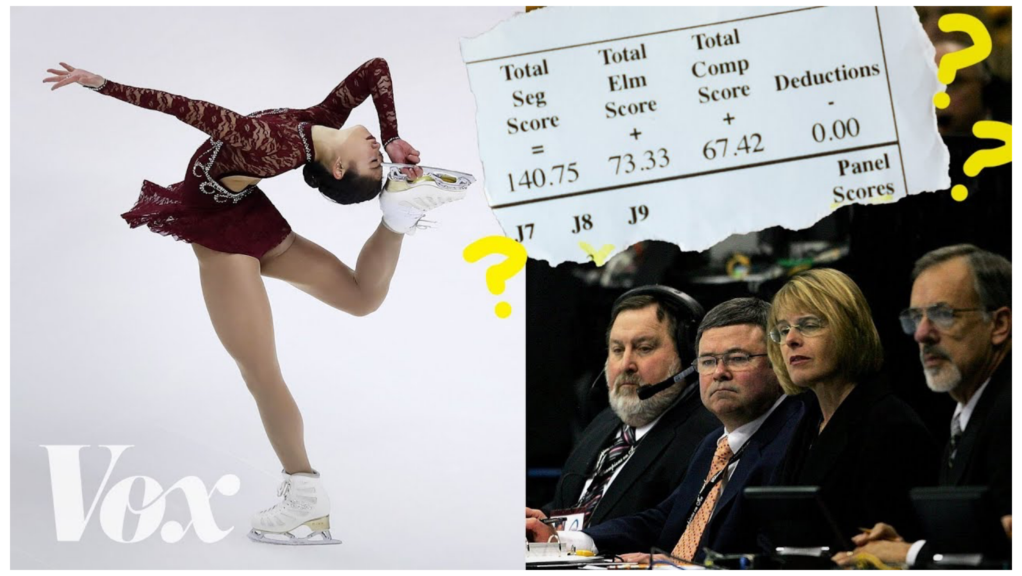


$$R_\theta(s, a) = \theta \cdot \varphi(s, a)$$

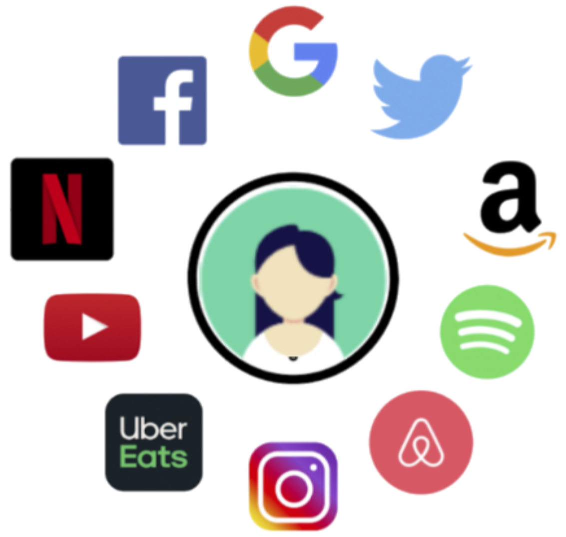
Linear social choice is more common than you think.



AI alignment

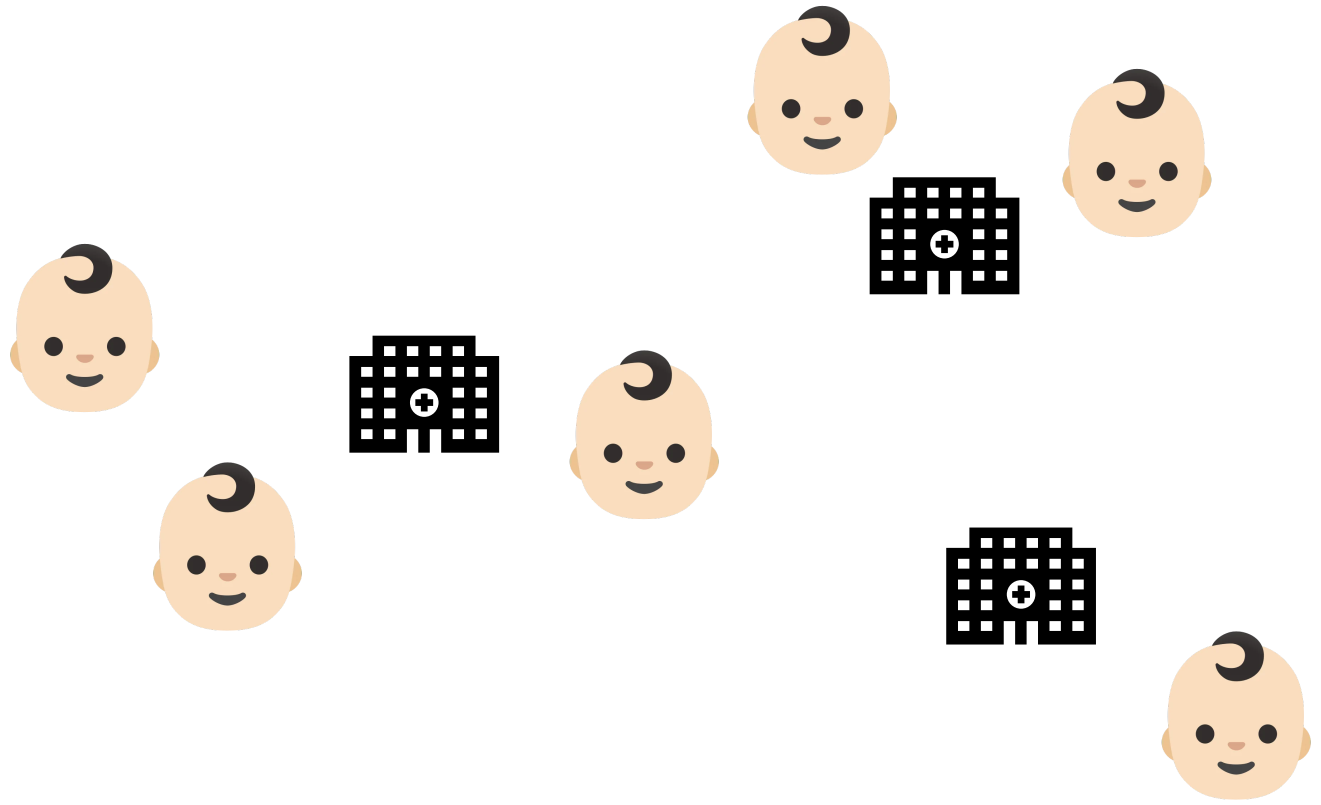


Multi-Objective Rubrics



Collaborative Filtering

There are not many paradigms in social choice.



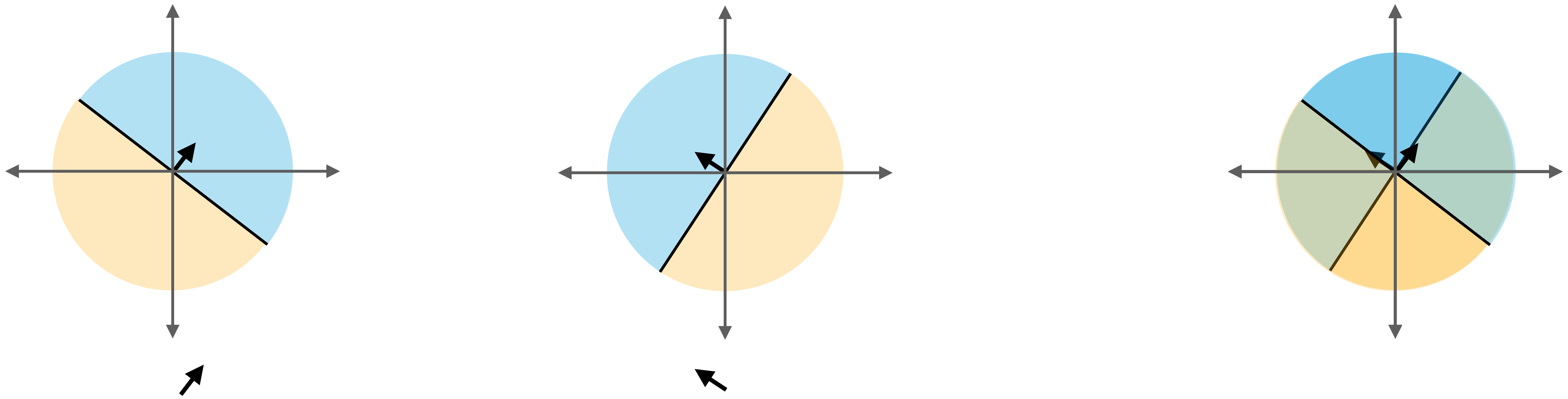
Facililty location problem—

disutillity= dist(v,c)

The information bottleneck for pluralistic alignment

every user has their own utility parameters $\theta \sim \Theta$

If we can ask each user many many pairwise preference queries,
we can identify each user's linear utillity weights under normalization [Zhu et. al 2023; Ge et. al 2024].

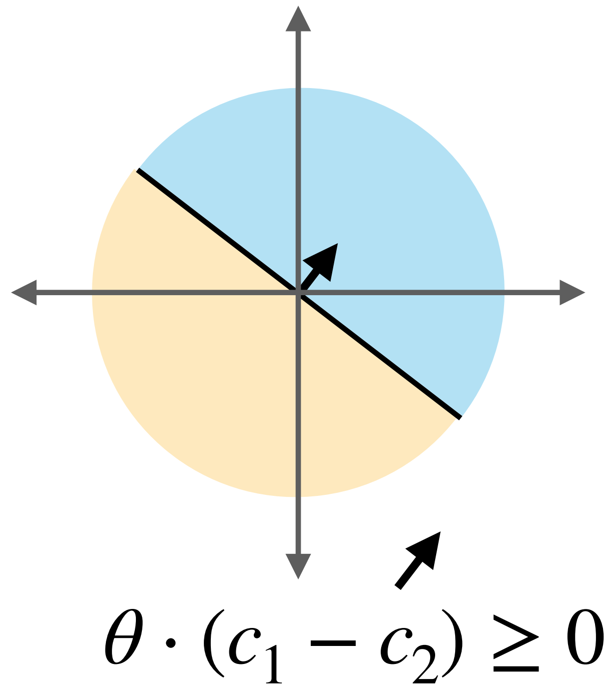


$\theta \cdot (c_1 - c_2) \geq 0$ $\qquad$ $\theta \cdot (c_3 - c_4) \geq 0$

The information bottleneck for pluralistic alignment

every user has their own utility parameters $\theta \sim \Theta$

If we can ask each user many many pairwise preference queries,
we can identify each user's linear utillity weights under normalization [Zhu et. al 2023; Ge et. al 2024].



However, in a typical alignment dataset, the preferences are annoymous.
One pairwise comparison tells you very little about a vector's position⋯

Research Question: What can we learn about the distribution Θ under few queries?

Knowledge about a distribution?

mean; variance; skewness$\cdots$ moments!

$$m_k = \int x^k d\mu(x)$$

$$m_k = \int \theta^{\otimes k} d\mu(\theta)$$

Hausdorff moment problem:
For a distribution of mass or probability on a bounded interval, *the collection of all the moments (of all orders, from 0 to ∞) uniquely determines the distribution.*

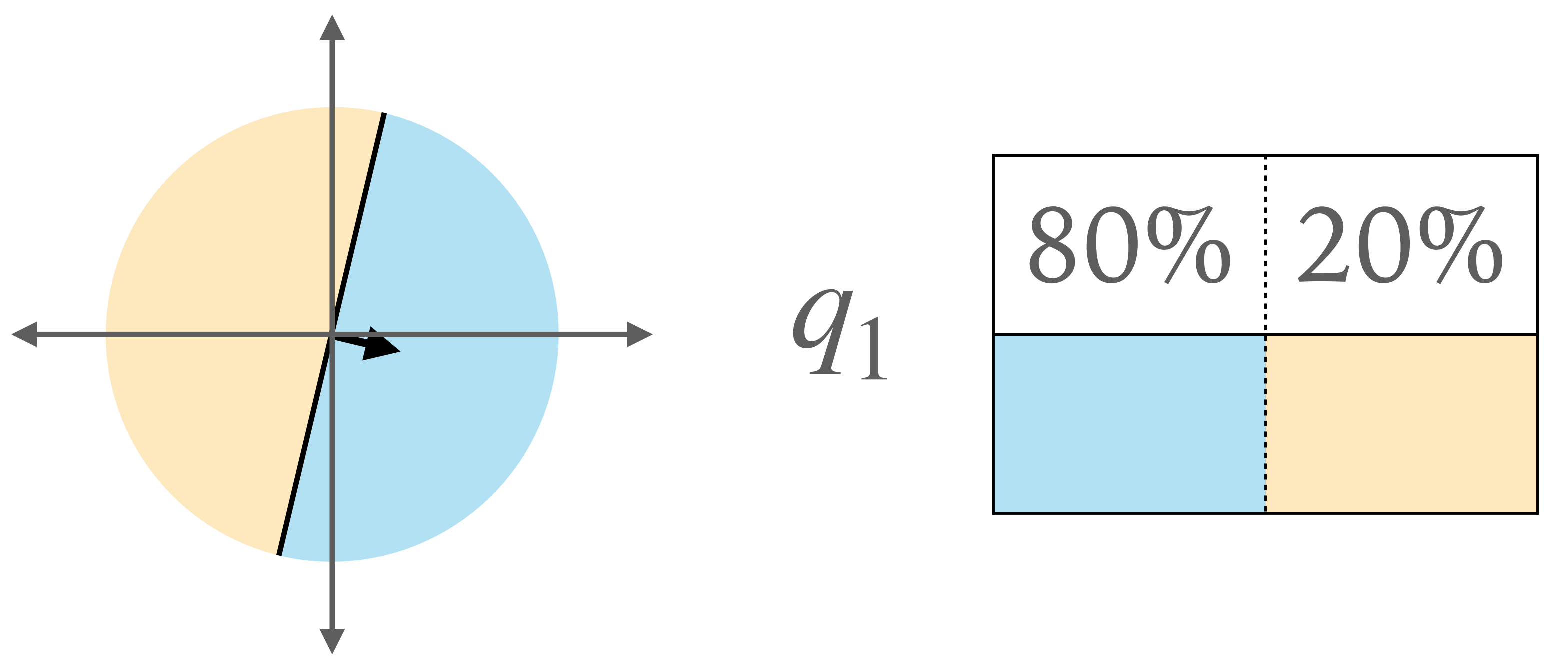
Usage of First Moment

- Goal: maximize welfare $r(x, y) = \mathbb{E}_{\theta \sim \Theta}[\theta \cdot \varphi(x, y)]$
- Suffices to learn $\bar{\theta} = \mathbb{E}_{\theta \sim \Theta}[\theta]$, and set $r(x, y) = \bar{\theta} \cdot \varphi(x, y)$

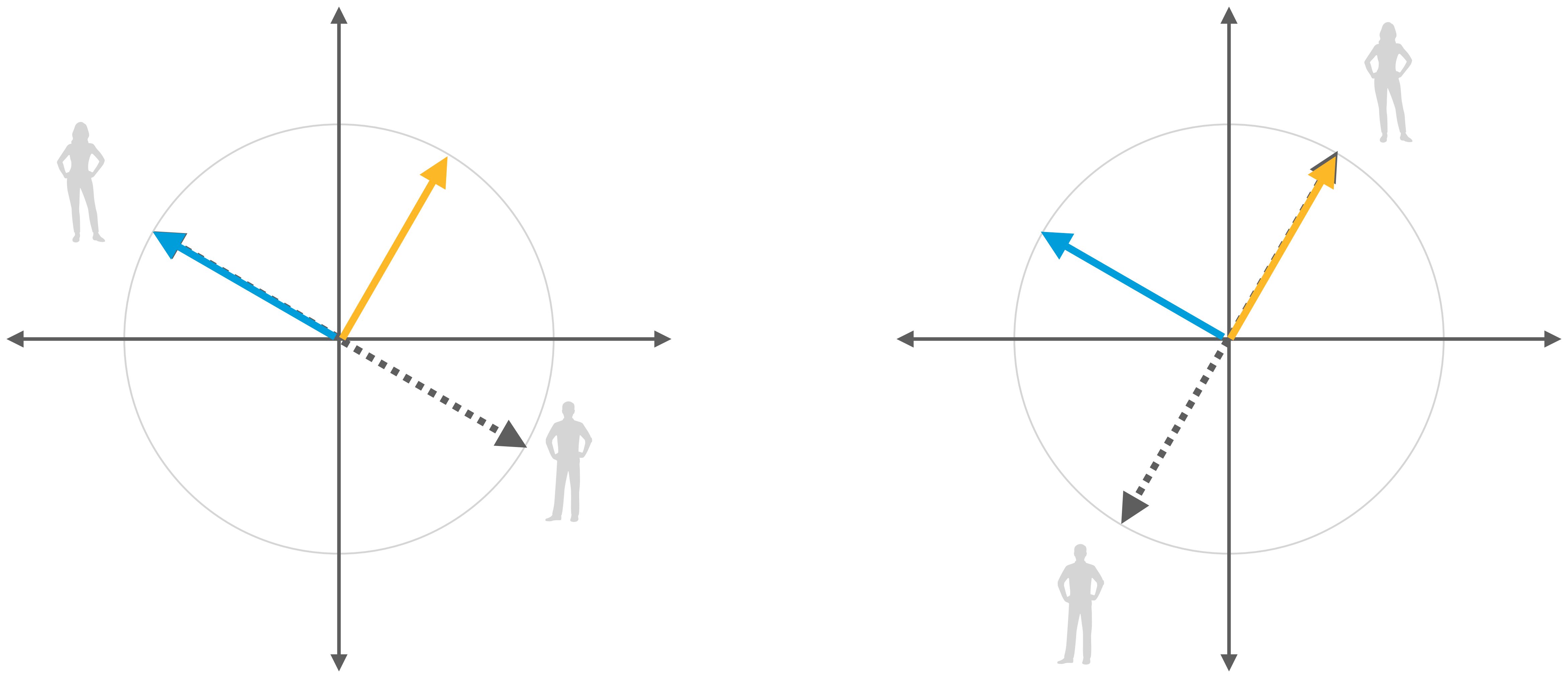
Theorem: This is possible with one query per voter!

proof trick: exploitation of symmetry

- uniformly random query direction
- uniformly random voters



Beyond Mean Estimation and Welfare Maximizing



- These two populations respond 50/50 to all queries so are indistinguishable
- Should we pick y_1 or y_2?
- 50% utility 1 and 50% utility -1 vs 100% 0

More egalitarian objectives requires higher moments and hence more queries per voter!

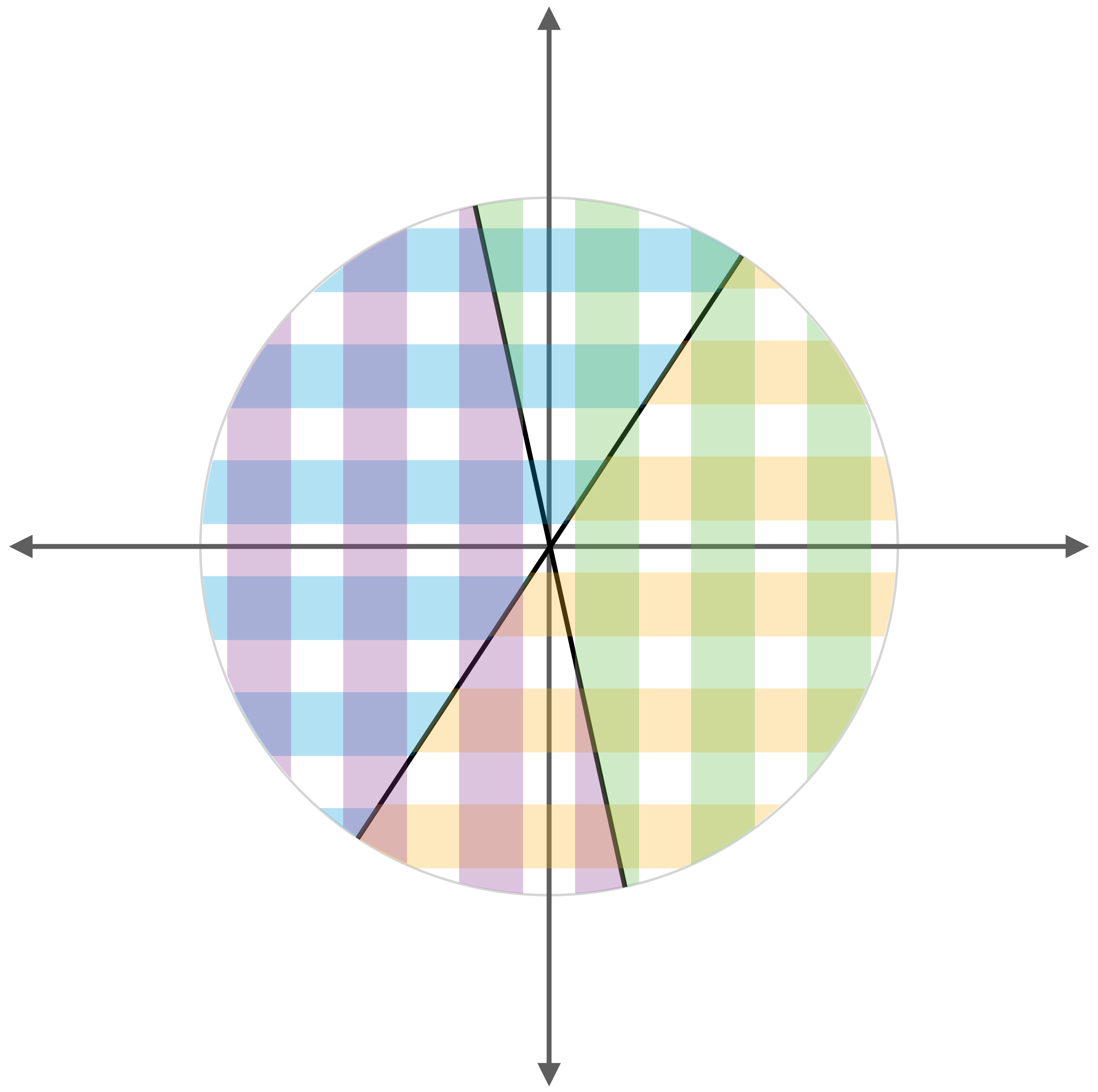
- Dispersion-adjusted: (mean utility) – $\gamma \cdot$ (stddev utility)
- Quantile-based: Maximize the 10th percentile utility
- Concave Utility: Maximize $\mathbb{E}[f(\text{utility})]$

K queries per person

- Fractions of people that prefers a_1 to b_1 **and** a_2 to b_2, …

TWO QUERIES PER PERSON

- Is y_1 better than y_2 for x and is y_1' better than y_2' for x'?



	y_1	y_2
y_1'	40%	16%
y_2'	13%	31%

Two Query possibility

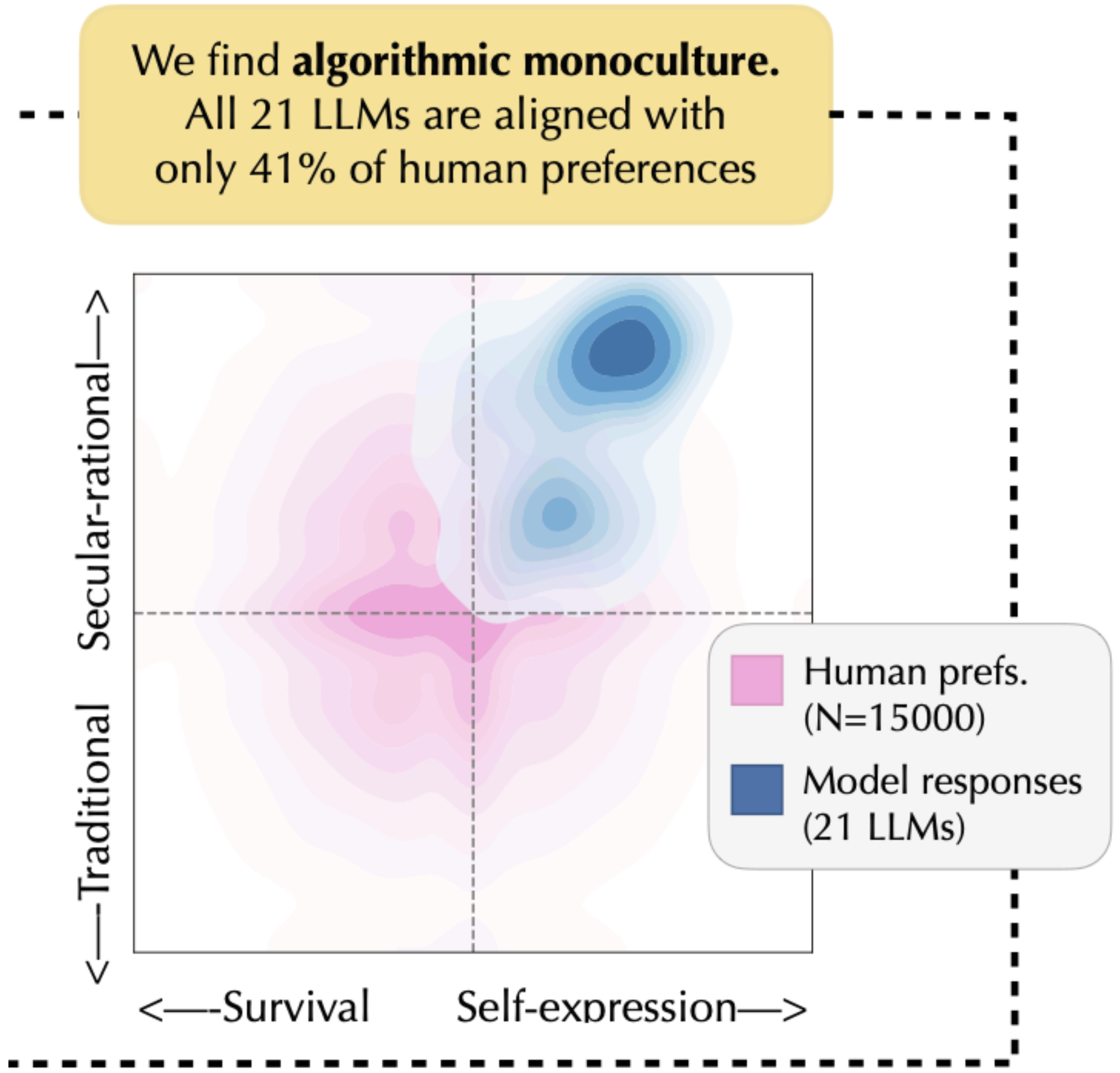
❋ Theorem: Eliciting just two pairwise comparisons per user allows us to *exactly* identify the underlying preference distribution Θ

- Can optimize arbitrary* functions of the voters!
- Need $\approx O(d^{3k+2})$ samples to estimate the k'th moments of Θ
- All moments uniquely identify Θ
- Proof using spherical harmonics

❋ Theorem: Alternatively, eliciting k pairwise comparisons per user allows us to estimate the k-th moment with $O(k(2\pi)^k d^{\lfloor k/2 \rfloor})$ samples.

Conclusions

➤ The query itself should not be biased → Better data generation/collection protocols?



[Zhang et. al 2025] Community Alignment Dataset

➤ We should get to know more about each voter: two is already much better than one! (Recent papers also show under very different preference models [Cherapanamjeri et al., 2026, Chidambaram et al., 2026])