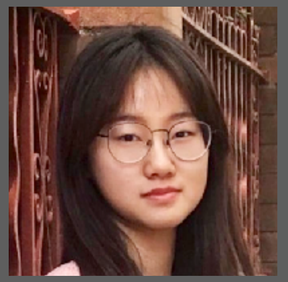
Distortion of AI Alignment

Does Preference Optimization Optimize for Preferences?

Paul Gölz

Nika Haghtalab



Kunhe Yang

LLM Training Pipeline

"supervised fine-tuning"

"reinforcement learning from human feedback"

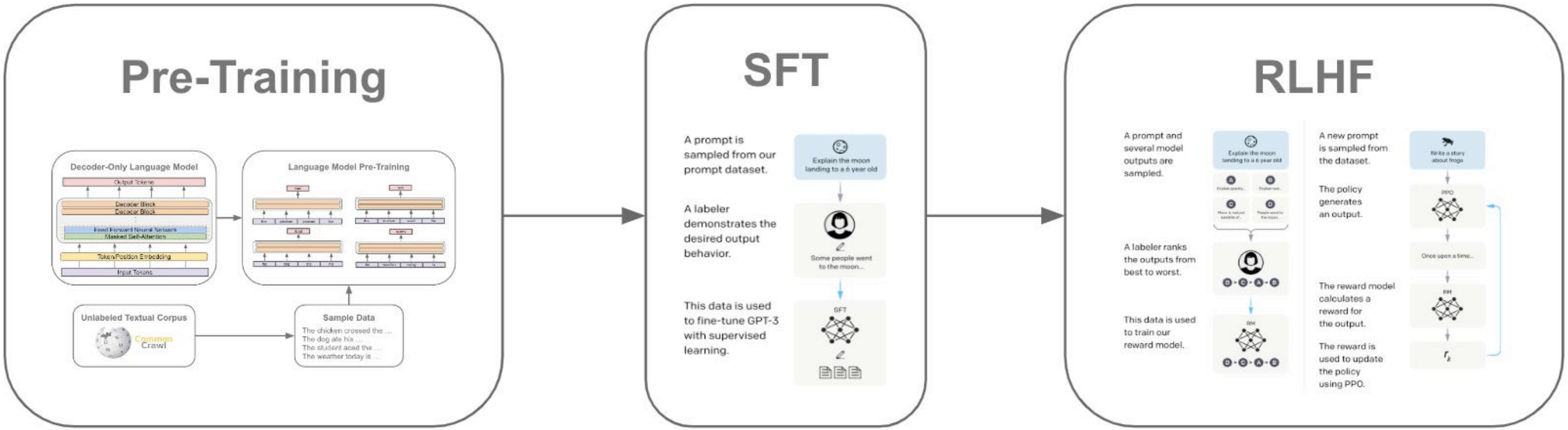


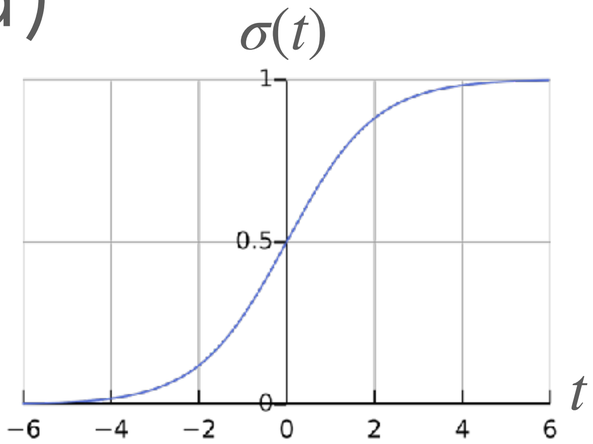
Image: https://cameronrwolfe.substack.com/p/understanding-and-using-supervised

Bradley–Terry Estimation

Given a prompt, each LLM response x has reward $r_x \in \mathbb{R}$.

Comparing responses x, y, a rater prefers x with probability $\sigma(r_x - r_y)$. (σ = logistic sigmoid)

Estimate rewards via maximum likelihood: $\max_{\vec{r}} \sum_{x_i \succ_i y_i} \log \sigma(r_{x_i} - r_{y_i})$.



RLHF updates model towards higher rewards.

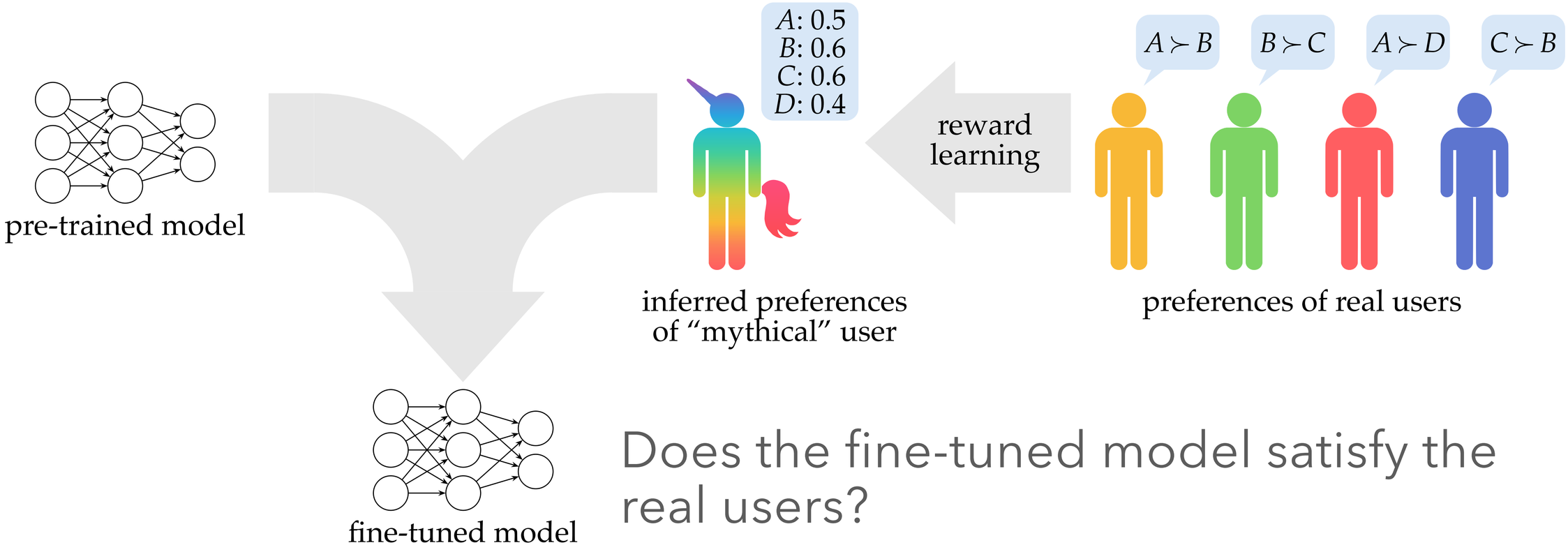
Preference Heterogeneity

If the Bradley–Terry assumption was correct, the maximum likelihood estimator (MLE) is consistent, so will recover true scores for $n \to \infty$ comparisons.

⚠️ Model assumes that all differences in preferences can be modeled as random noise.

What if raters disagree on values & preferences?

RLHF Ignores Preference Heterogeneity



Does the fine-tuned model satisfy the real users?

A question for social choice theory!

Voting Setting

Classic Distortion

n voters, m alternatives.

Voter i has utility $u_i(x) \in [0,1]$ for alternative x.* We only see induced ordering of alternatives σ_i.

Voting rule f takes in $\vec{\sigma}$, returns alternative $f(\vec{\sigma})$.

Maximize $SW(x, \vec{u}) = \sum_i u_i(x)$.

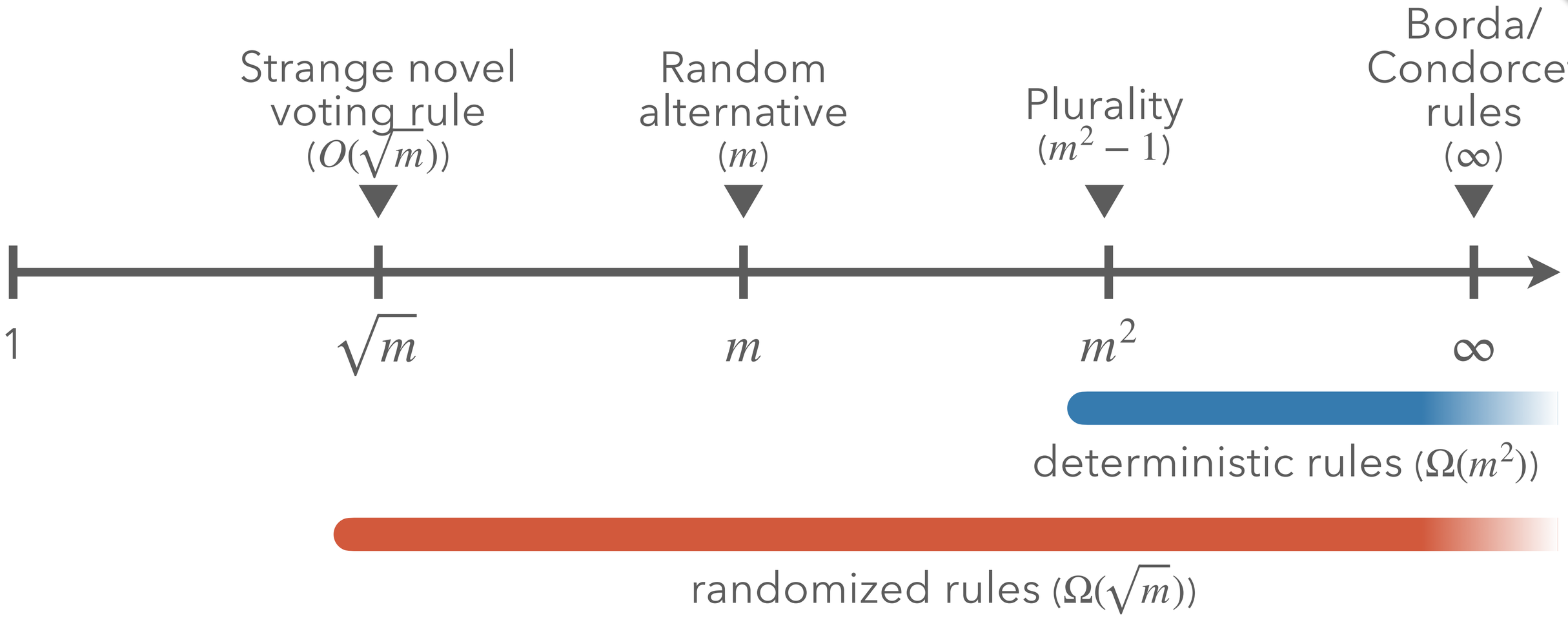
(or expected value, if f returns a distribution over alternatives.)

$$distortion(f) := \sup_{\vec{u} \rhd \vec{\sigma}} \frac{\max_x SW(x, \vec{u})}{SW(f(\vec{\sigma}), \vec{u})}.$$

*PR06 additionally assume that $\sum_x u_i(x) = 1$ for each i, we won't.

[PR06] Procaccia & Rosenschein. The distortion of cardinal preferences in voting.

Classic Distortion Bounds



[PR06] Procaccia & Rosenschein. The distortion of cardinal preferences in voting.
[BCH12] Boutilier et al. Optimal social choice functions.
[EKP+24] Ebadian et al. Optimized distortion and proportional fairness in voting.

Classic Distortion Bounds

Strange novel voting rule ($O(\sqrt{m})$)

Random alternative (m)

Plurality ($m^2 - 1$)

Borda/ Condorcet rules (∞)

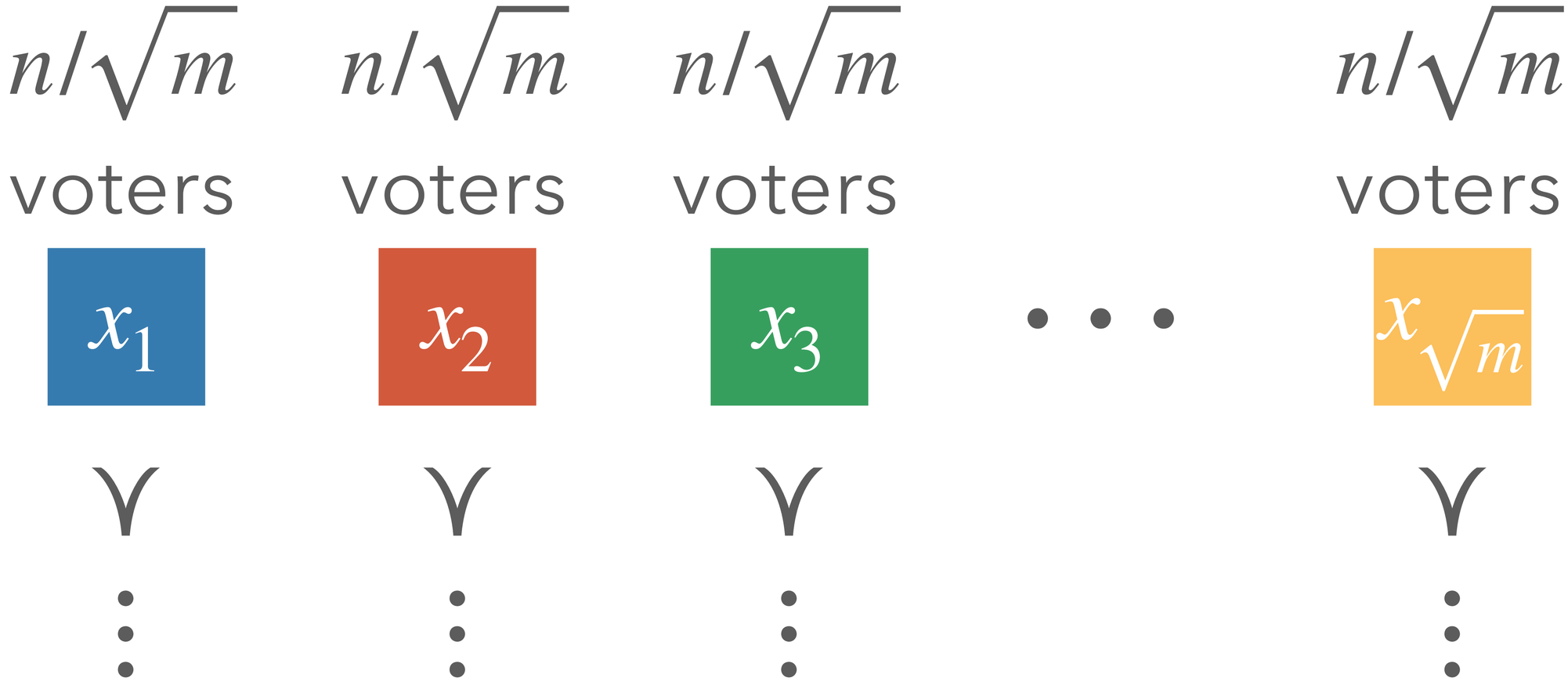
rality of Borda count) that are easy for voters to understand and satisfy appealing axiomatic prop-
erties. A significant barrier to the modern optimization-based approaches, which focus on quan-
titative objectives such as distortion or proportional fairness, is that they often yield rules that
are difficult to understand (and sometimes difficult to compute). Significant challenges lie ahead
in paving the path for increased practicability of such approaches: Can we design simple rules

[EKP+24] Ebadian et al. Optimized distortion and proportional fairness in voting.

randomized rules ($\Omega(\sqrt{m})$)

[PR06] Procaccia & Rosenschein. The distortion of cardinal preferences in voting.
[BCH12] Boutilier et al. Optimal social choice functions.
[EKP+24] Ebadian et al. Optimized Distortion and Proportional Fairness in Voting.

Classic Distortion Lower Bound



Bloc whose alternative is least likely to win cares only about top-ranked, all other blocs indifferent.

Our Distortion Model

Continuum of voters, m alternatives.

Voter i has utility $u_i(x) \in [0,\beta]$ for alternative x. We only see a Bradley–Terry comparison between $x, y \sim \mu$ for rewards $u_i(x)$.

Similar assumption in *metric* distortion [GS25].

Voting rule f takes in pairwise comparisons, returns alternative. Maximize $SW(x,\vec{u}) = \sum_i u_i(x)$.

$$distortion(f) := \sup_{\vec{u}} \frac{\max_x SW(x,\vec{u})}{SW(f,\vec{u})}.$$

expected SW over $n \to \infty$ random comparisons

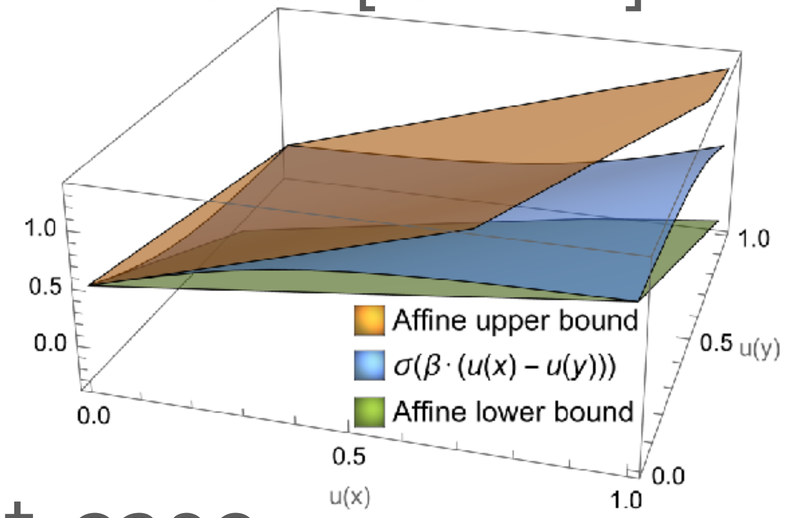
[GS25] Goyal & Sarmasakar. Metric distortion under probabilistic voting.

Why Might Randomness Help?

Can't recover utilities (too little data per voter).

If, instead of comparison, agent gave us alternative x with probability $u_i(x)$, easy to maximize SW [CP11].

For us, adding up pairwise comparisons loses information, but preserves affine bounds.



cf. smoothed analysis/beyond worst-case.

[CP11] Caragiannis & Procaccia. Voting almost maximizes social welfare despite limited communication.

Distortion Lower Bound

Theorem. No voting rule guarantees distortion better than $\frac{\beta}{2} \cdot \frac{1+\exp(-\beta)}{1-\exp(-\beta)} \approx \frac{\beta}{2}$ for large m.

🌍 **one:**

minority: $(1,0,\cdots,0)$

majority: $(0,\epsilon,\cdots,\epsilon)$

🌏 **two:**

minority: $(0,1,0,\cdots,0)$

majority: $(\epsilon,0,\epsilon,\cdots,\epsilon)$

• • •

Balance population such that all pairs are majority-tied, good alternative indistinguishable.

Distortion Lower Bound

Theorem. No voting rule guarantees distortion better than $\frac{\beta}{2} \cdot \frac{1+\exp(-\beta)}{1-\exp(-\beta)} \approx \frac{\beta}{2}$ for large m.

(Relies on assumption that each voter provides single pairwise comparison. For several comparisons, extends to voting rules that satisfy Condorcet loser or Condorcet consistency.)

Distortion Upper Bounds

Bradley–Terry assumption leads to distortion bounds that are constant in m!

Theorem. Borda and Copeland voting rules have $O(\beta^2)$ distortion.

Theorem. Maximal lotteries voting rule has exactly $\frac{\beta}{2} \cdot \frac{1+\exp(-\beta)}{1-\exp(-\beta)} \approx \frac{\beta}{2}$ distortion.

Alignment Setting

Alignment vs. Voting

Difference 1: Must return distribution over responses for every prompt, generalize preferences over prompts. Won't model this.

Difference 2: Returned distribution π over responses must be close to given reference policy π_{ref}, in the sense that

$$D_{KL}(\pi \| \pi_{ref}) \leq \tau.$$

Kullback–Leibler divergence

Image: https://francisbach.com/von-neumann-entropy/

How To Generalize Distortion?

What is the right benchmark for our policy?

KL-ball around reference policy

$$distortion(f) := \sup_{\vec{u},\pi_{ref},\tau} \frac{\max_{\pi^* \in B_\tau(\pi_{ref})} SW(\pi^*, \vec{u})}{SW(f, \vec{u})}.$$

expected SW over $n \to \infty$ comparisons

If τ is large enough that $B_\tau(\pi_{ref})$ the whole simplex, recover old definition, so lower bound applies.

Alignment Method 1: RLHF

Estimate the Bradley–Terry rewards r_x using MLE.

Then, optimize $\max_{\pi \in B_\tau(\pi_{ref})} \mathbb{E}_{x \sim \pi}[r_x]$.

In practice, using reinforcement learning on neural network, we assume perfect maximization.
Also captures *direct preference optimization* (DPO).

Rewards are ordered as Borda scores [SLH24], so coincides with Borda rule in social choice setting.

[SLH24]: Siththaranjan et al. Distributional Preference Learning.

Distortion of RLHF

Unconstrained (Borda)	With KL constraint	Comparisons not i.i.d.
$\geq (1 - o(1))\,\beta$ $\leq O(\beta^2)$	$\geq e^{\Omega(\beta)}$	unbounded in β

Idea: MLE optimizes rewards to predict frequent comparison pairs well, but KL constraint makes only rare comparison pairs matter.

Alignment Method 2: NLHF

Proposed alignment method: Nash learning from human feedback (NLHF) [MVC+23].

Let $M_{x,y}$ be the fraction of x, y comparisons won by x. NLHF policy is

$$\operatorname{argmax}_{\pi_1 \in B_\tau(\pi_{ref})} \min_{\pi_2 \in B_\tau(\pi_{ref})} \mathbb{E}_{x_1 \sim \pi_1, x_2 \sim \pi_2}[M_{x_1,x_2}].$$

In unconstrained setting, coincides with maximal lotteries voting rule.

[MVC+23]: Munos et al. Nash learning from human feedback.

Distortion of NLHF

Theorem. The distortion of NLHF is $\frac{\beta}{2} \cdot \frac{1+\exp(-\beta)}{1-\exp(-\beta)} \approx \frac{\beta}{2}$, even in the alignment setting.

Proof.

Recall: $\pi_{NLHF} = \text{argmax}_{\pi_1 \in B_\tau(\pi_{ref})} \min_{\pi_2 \in B_\tau(\pi_{ref})} \mathbb{E}_{x_1 \sim \pi_1, x_2 \sim \pi_2}[M_{x_1,x_2} - \frac{1}{2}]$.

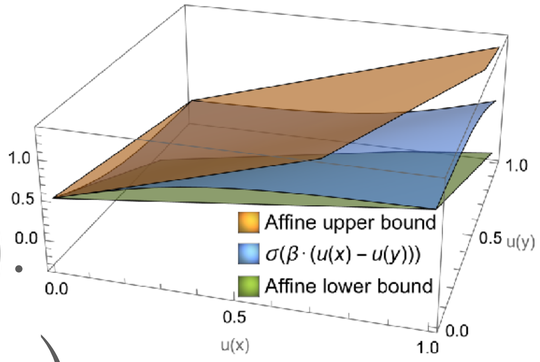
Symmetric zero-sum game over compact domain, so value 0.

Let the benchmark $\pi^* := \text{argmax}_{\pi^* \in B_\tau(\pi_{ref})} SW(\pi^*, \vec{u})$.

$\mathbb{E}_{x_1 \sim \pi_{NLHF}, x_2 \sim \pi^*}[M_{x_1,x_2} - \frac{1}{2}] \geq 0$

Distortion of NLHF

Theorem. The distortion of NLHF is $\frac{\beta}{2} \cdot \frac{1+\exp(-\beta)}{1-\exp(-\beta)} \approx \frac{\beta}{2}$, even in the alignment setting.



Proof (cont.).

Affine bound: $\sigma(u_i(x) - u_i(y)) - \frac{1}{2} \leq \beta \cdot \left(\frac{1}{4} u_i(x) - \frac{1-e^{-\beta}}{2\beta(1+e^{-\beta})} u_i(y) \right)$.

In infinite-sample limit: $M_{x,y} - \frac{1}{2} \leq \beta \cdot \left(\frac{1}{4} SW(x) - \frac{1-e^{-\beta}}{2\beta(1+e^{-\beta})} SW(y) \right)$.

From last slide: $\mathbb{E}_{x_1 \sim \pi_{NLHF}, x_2 \sim \pi^*}[M_{x_1,x_2} - \frac{1}{2}] \geq 0$

Hence, $\frac{1}{4} SW(\pi_{NLHF}) - \frac{1-e^{-\beta}}{2\beta(1+e^{-\beta})} SW(\pi^*) \geq 0.$

Hence, $\frac{SW(\pi^*)}{SW(\pi_{NLHF})} \leq \frac{\beta}{2} \frac{1+e^{-\beta}}{1-e^{-\beta}}$.

□

Distortion of NLHF

Theorem. The distortion of NLHF is $\frac{\beta}{2} \cdot \frac{1+\exp(-\beta)}{1-\exp(-\beta)} \approx \frac{\beta}{2}$, even in the alignment setting.

Perfectly matches lower bound.

Continues to hold if comparisons are not i.i.d., or for other convex divergence constraints.

Take-Home Messages

Prescriptions of distortion model in voting become reasonable with Bradley–Terry assumption.

Social choice can point out failure modes for heterogeneous alignment, possible alternatives.

https://paulgoelz.de/papers/alignment.pdf