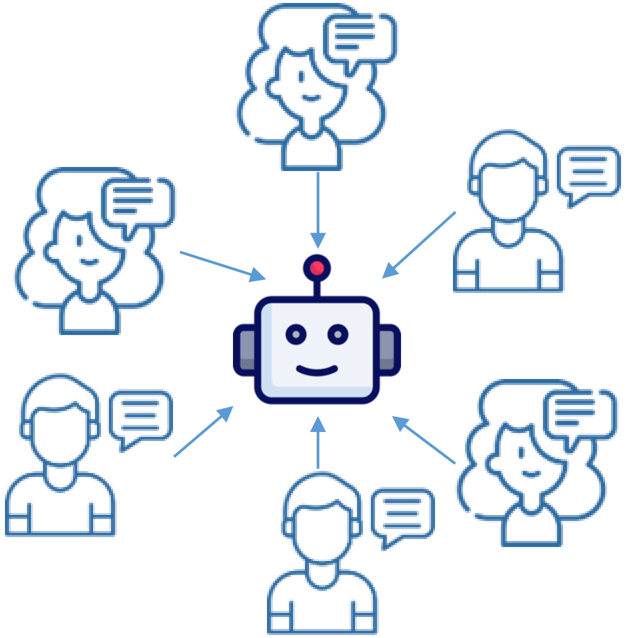
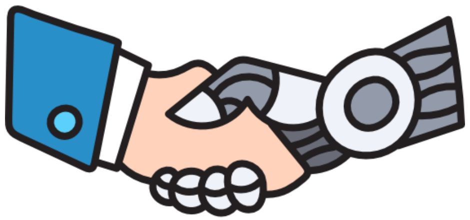
Towards Pluralistic Alignment: From Axiomatic Foundations to Pairwise Calibration



Evi Micha

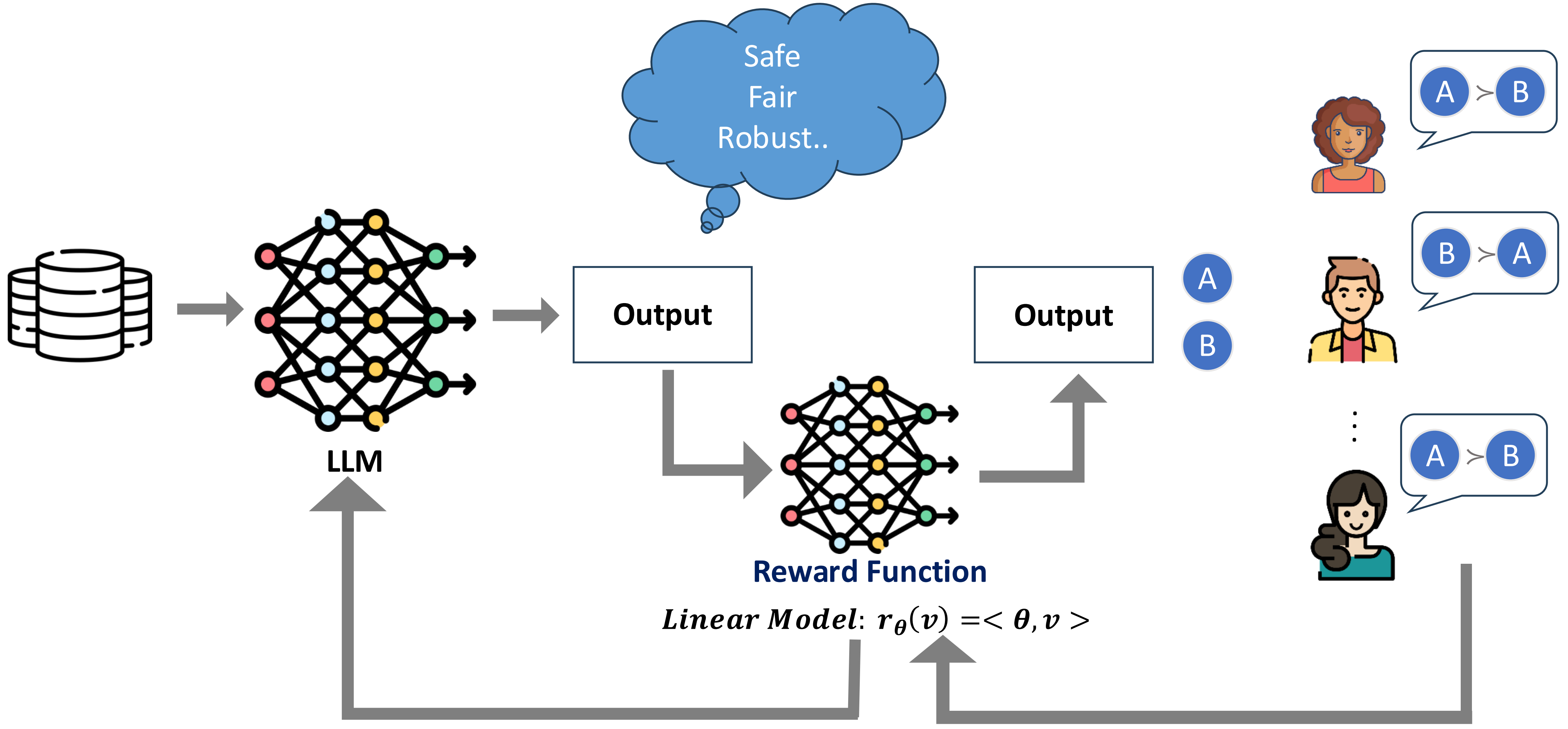
University of Southern California

AI Alignment

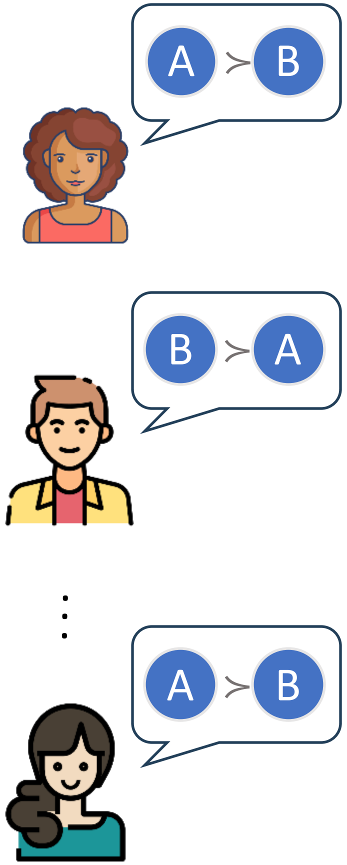


…AI alignment involves ensuring that an AI system's objectives match those of its designers…
(wikipedia)

Reinforcement Learning with Human Feedback



Random Utility Models



BTL Model

A ≻ B $\qquad \dfrac{e^{r_\theta(A)}}{e^{r_\theta(A)} + e^{r_\theta(B)}}$

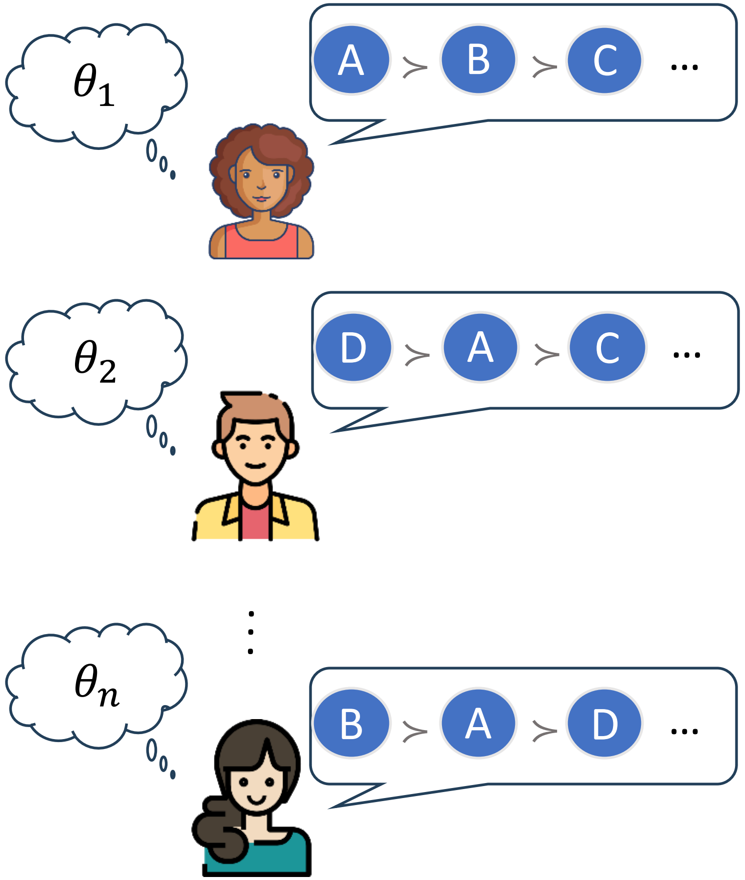
B ≻ A $\qquad \dfrac{e^{r_\theta(B)}}{e^{r_\theta(A)} + e^{r_\theta(B)}}$

Profile of Ordinal Preferences

$$inf_\theta \ L(\theta;\pi) = inf_\theta \ \sum_{A\neq B} n_{A\succ B}(\pi) \cdot ln(1 + e^{r_\theta(B)-r_\theta(A)})$$

Number of voters in π that prefer A to B

Heterogeneous Preferences



Axiomatic Approach

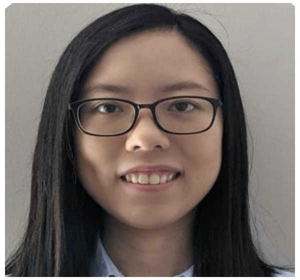
Luise Ge

Daniel Halpern

Ariel Procaccia

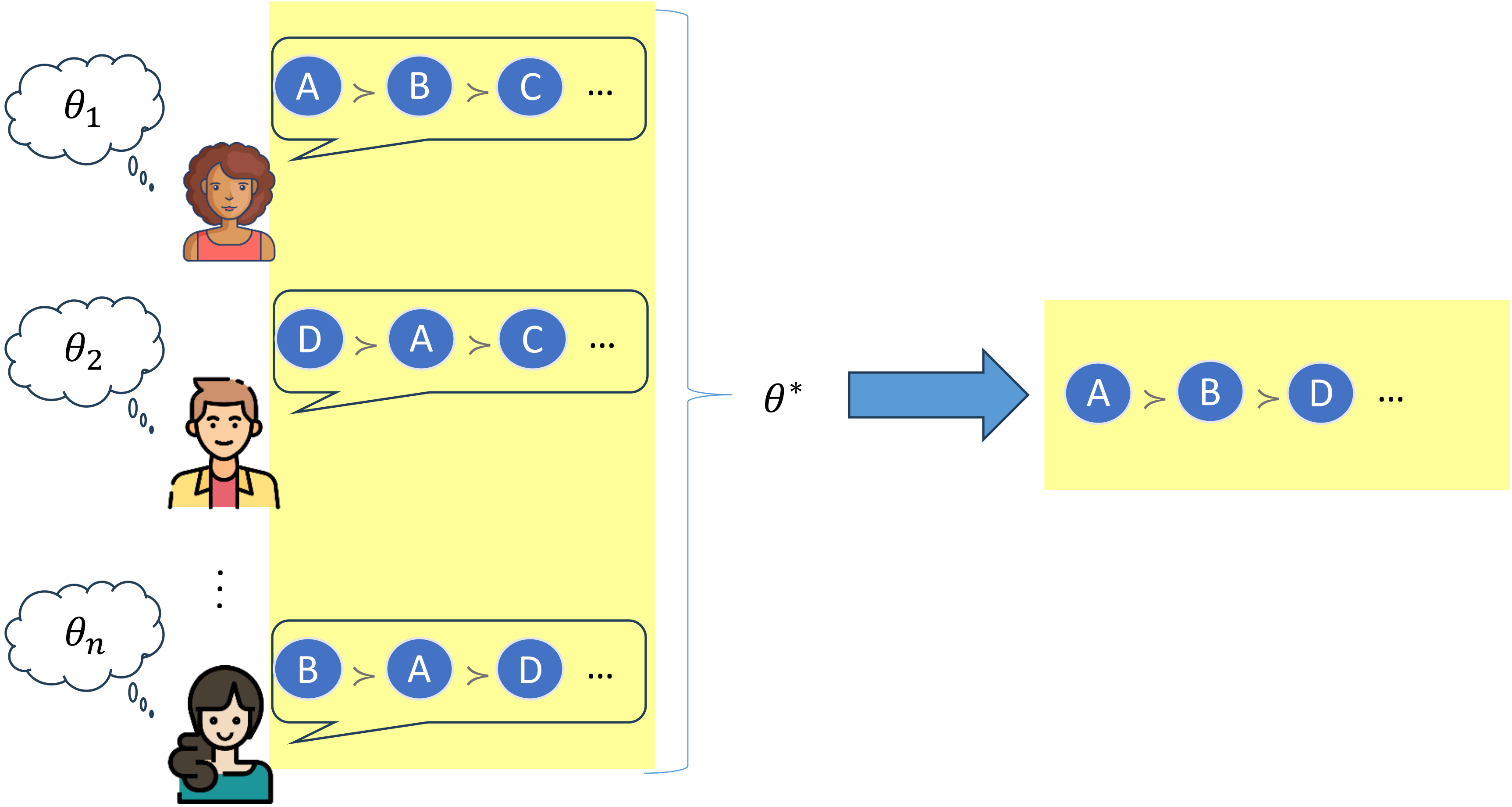
Itai Shapira

Yevgeniy Vorobeychik



Junlin Wu

Heterogeneous Preferences



Linear Model: $r_\theta(v) = < \theta, v >$

Linear Social Choice

A $v_A = [20,0,0]$

B $v_B = [0,20,0]$

C $v_C = [0,10,10]$

D $v_D = [0,0,1]$

E $v_E = [1,0,0]$

$\theta = [\theta_1, \theta_2, \theta_3] \Rightarrow A \succ B \succ C \succ D \succ E$

Linear Social Choice

A $v_A = [20,0,0]$

B $v_B = [0,20,0]$

C $v_C = [0,10,10]$

D $v_D = [0,0,1]$

E $v_E = [1,0,0]$

$\theta = [\theta_1, \theta_2, \theta_3]$ ➡ A $\succ$ B $\succ$ C $\succ$ D $\succ$ E

$\theta_1 > \theta_2$

Linear Social Choice

A $v_A = [20,0,0]$

B $v_B = [0,20,0]$

C $v_C = [0,10,10]$

D $v_D = [0,0,1]$

E $v_E = [1,0,0]$

$\theta = [\theta_1, \theta_2, \theta_3]$ ➡ A $\succ$ B $\succ$ C $\succ$ D $\succ$ E

$\theta_1 > \theta_2$

$\theta_2 > \theta_3$

Linear Social Choice

A $v_A = [20,0,0]$

B $v_B = [0,20,0]$

C $v_C = [0,10,10]$

D $v_D = [0,0,1]$

E $v_E = [1,0,0]$

$\theta = [\theta_1, \theta_2, \theta_3]$ ➡ A $\succ$ B $\succ$ C $\succ$ D $\succ$ E

$\theta_1 > \theta_2$

$\theta_2 > \theta_3$

$\theta_3 > \theta_1$

Linear Social Choice

A $v_A = [20,0,0]$

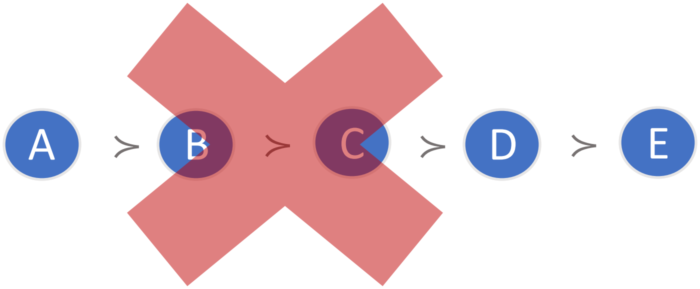
B $v_B = [0,20,0]$

C $v_C = [0,10,10]$

D $v_D = [0,0,1]$

E $v_E = [1,0,0]$

$\theta = [\theta_1, \theta_2, \theta_3]$ ➡



Linear Rank Aggregation Rules

Axiomatic Approach

Goals:

- What axioms are satisfied by aggregation methods used by existing RLHF algorithms?
- Are there alternative aggregation methods that offer stronger axiomatic guarantees?

- **Pareto Optimality:** A linear rank aggregation rule f satisfies Pareto optimality if, whenever every voter prefers candidate a over candidate b, then candidate a is ranked higher than candidate b in the output ranking

- **Pairwise Majority Consistency (PMC):** A ranking σ is called a PMC ranking for profile π if for all a, b ∈ C, $a \succ_\sigma b$ if and only if a majority of voters rank $a \succ b$. A linear rank aggregation rule satisfies PMC if, when a PMC ranking σ exists for the input profile π and σ is feasible, then $f(\pi) = \sigma$

Loss-Based Rules

A loss function $\ell: \mathbb{R} \rightarrow \mathbb{R}$

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

BTL model: $\ell(x) = \ln(1 + e^{x})$

Theorem (informal): If a linear rank aggregation rule f optimizes a loss function that is either nondecreasing and weakly convex, or strictly convex then f ***fails PO and PMC***

A Social Choice Based Rule

- **Leximax Copeland subject to PO**

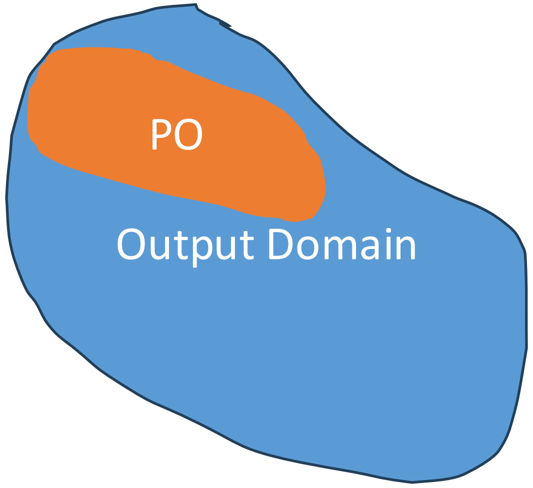
σ_1	σ_2	…	σ_n
1	2		3
2	1		2
3	3		$m-1$
⋮	⋮		⋮
m	$m-1$		m

Copeland →

σ^*
2
1
3
⋮
$m-1$

→

σ'



A Social Choice Based Rule

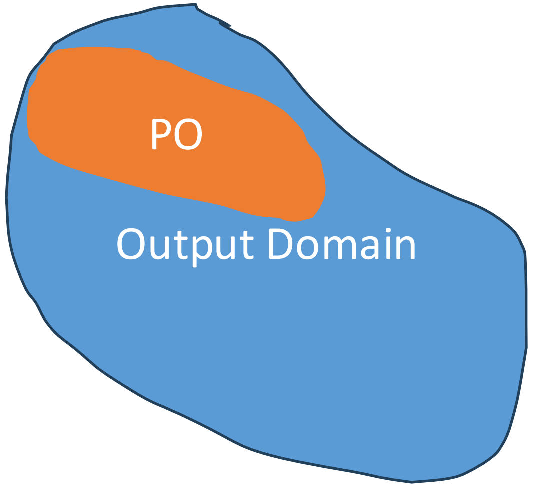
- **Leximax Copeland subject to PO**

σ_1	σ_2	…	σ_n
1	2		3
2	1		2
3	3		$m-1$
⋮	⋮		⋮
m	$m-1$		m

Copeland

σ^*
2
1
3
⋮
$m-1$

σ'



A Social Choice Based Rule

- **Leximax Copeland subject to PO**

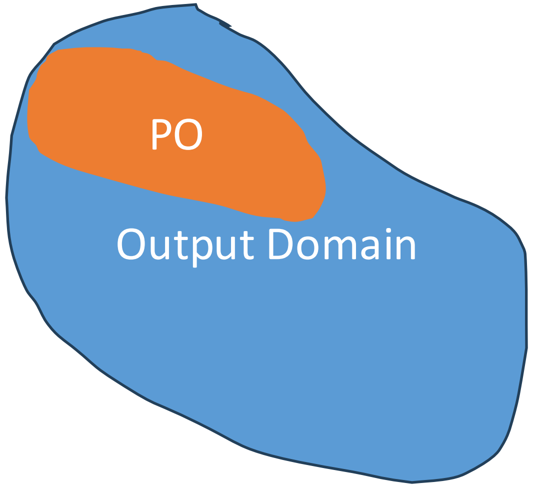
σ_1	σ_2	…	σ_n
1	2		3
2	1		2
3	3		$m-1$
⋮	⋮		⋮
m	$m-1$		m

Copeland →

σ^*
2
1
3
⋮
$m-1$

→

σ'



A Social Choice Based Rule

- **Leximax Copeland subject to PO**

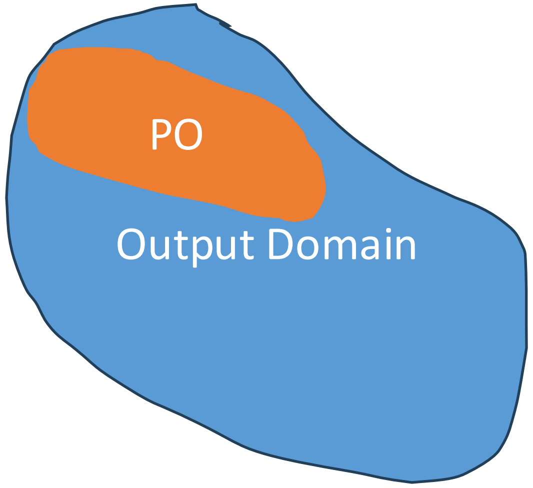
σ_1	σ_2	…	σ_n
1	2		3
2	1		2
3	3		$m-1$
⋮	⋮		⋮
m	$m-1$		m

Copeland →

σ^*
2
~~1~~
3
⋮
$m-1$

→

σ'
1



A Social Choice Based Rule

- **Leximax Copeland subject to PO**

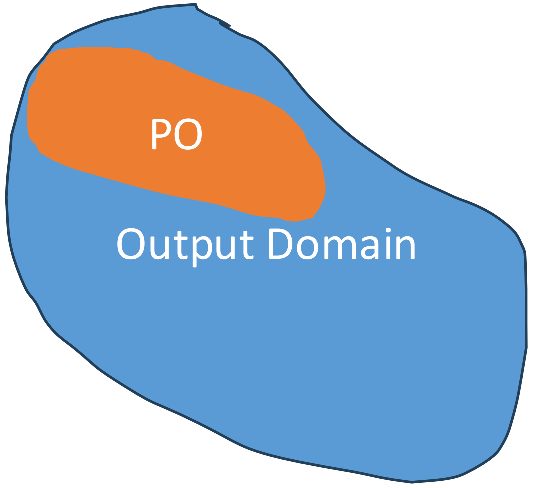
σ_1	σ_2	…	σ_n
1	2		3
2	1		2
3	3		$m-1$
$\vdots$	$\vdots$		$\vdots$
m	$m-1$		m

Copeland →

σ^*
2
1
3
$\vdots$
$m-1$

→

σ'
1



A Social Choice Based Rule

- **Leximax Copeland subject to PO**

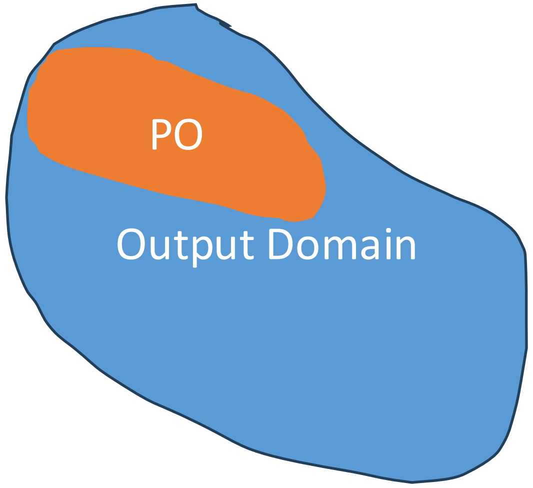
σ_1	σ_2	…	σ_n
1	2		3
2	1		2
3	3		$m-1$
⋮	⋮		⋮
m	$m-1$		m

Copeland →

σ^*
2
~~1~~
3
⋮
$m-1$

→

σ'
1



A Social Choice Based Rule

- **Leximax Copeland subject to PO**

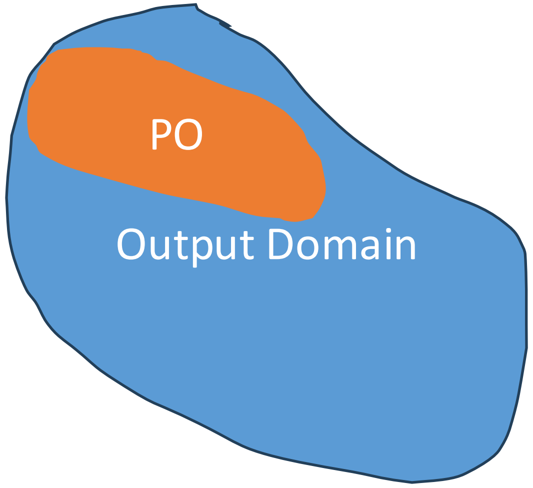
σ_1	σ_2	…	σ_n
1	2		3
2	1		2
3	3		$m-1$
⋮	⋮		⋮
m	$m-1$		m

Copeland →

σ^*
2
~~1~~
~~3~~
⋮
$m-1$

→

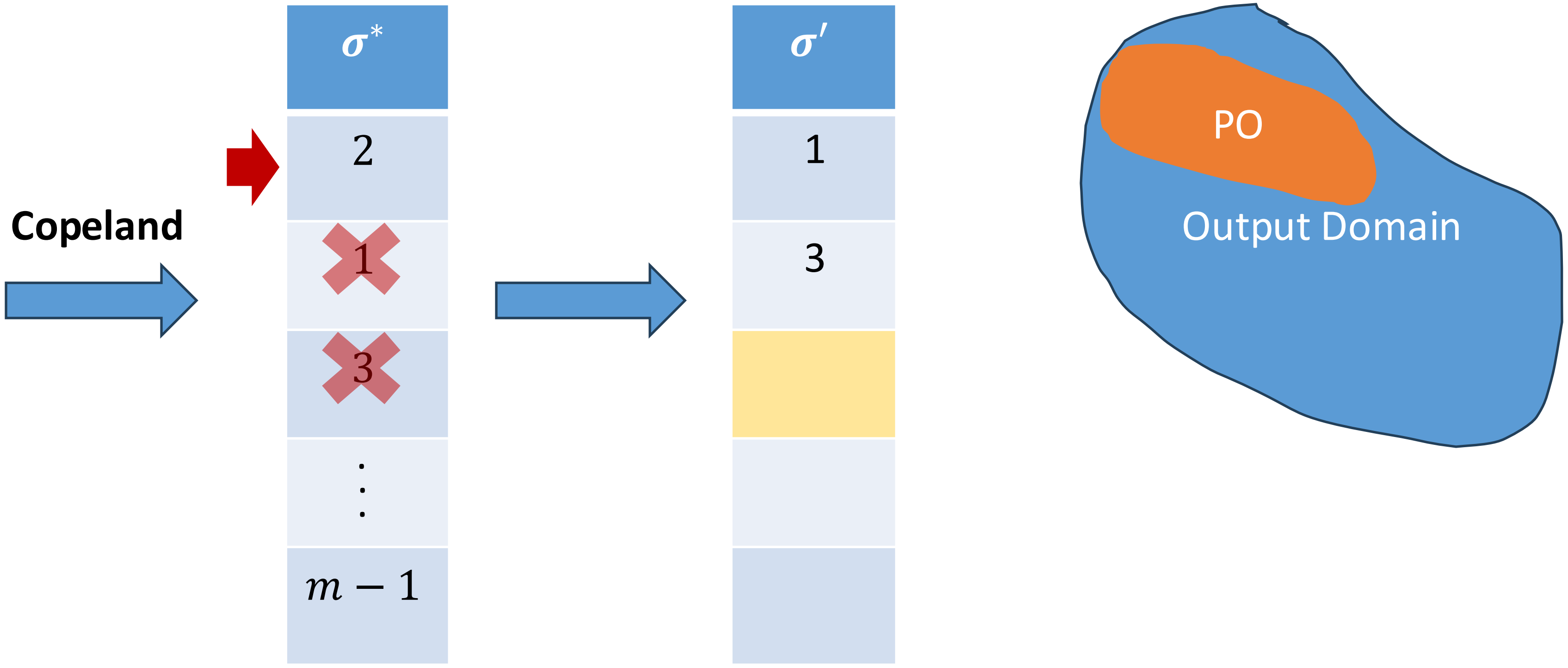
σ'
1
3



A Social Choice Based Rule

- **Leximax Copeland subject to PO**

σ_1	σ_2	…	σ_n
1	2		3
2	1		2
3	3		$m-1$
⋮	⋮		⋮
m	$m-1$		m



A Social Choice Based Rule

- **Leximax Copeland subject to PO**

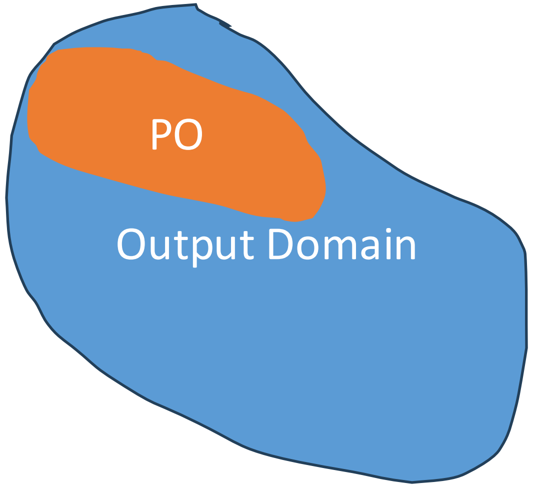
σ_1	σ_2	…	σ_n
1	2		3
2	1		2
3	3		$m-1$
⋮	⋮		⋮
m	$m-1$		m

Copeland →

σ^*
~~2~~
~~1~~
~~3~~
⋮
$m-1$

→

σ'
1
3
2



A Social Choice Based Rule

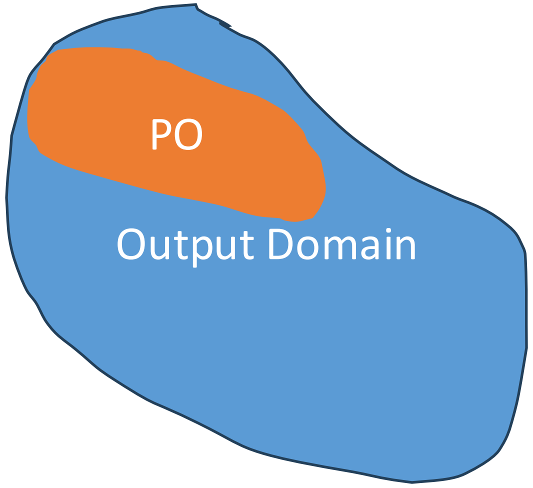
- **Leximax Copeland subject to PO**

σ_1	σ_2	…	σ_n
1	2		3
2	1		2
3	3		$m-1$
⋮	⋮		⋮
m	$m-1$		m

Copeland

σ^*
2
1
3
⋮
$m-1$

σ'
1
3
2
⋮
m



A Social Choice Based Rule

- **Theorem:** Leximax Copeland subject to PO **satisfies**
  a) PO
  b) PMC

A Social Choice Based Rule

- **Theorem:** Leximax Copeland subject to PO **satisfies**
  a) PO
  b) PMC
  c) majority consistency
  d) winner monotonicity

- **Majority Consistency:** A linear rank aggregation rule f satisfies majority consistency if when a candidate a is ranked first by a majority of voters in the input profile, a is ranked first in the output ranking

- **Winner Monotonicity:** A linear rank aggregation rule f satisfies winner monotonicity if, when a candidate a is ranked first in the output ranking, elevating a in any voter's preference does not cause a to lose their top position in the updated aggregate ranking

A Social Choice Based Rule

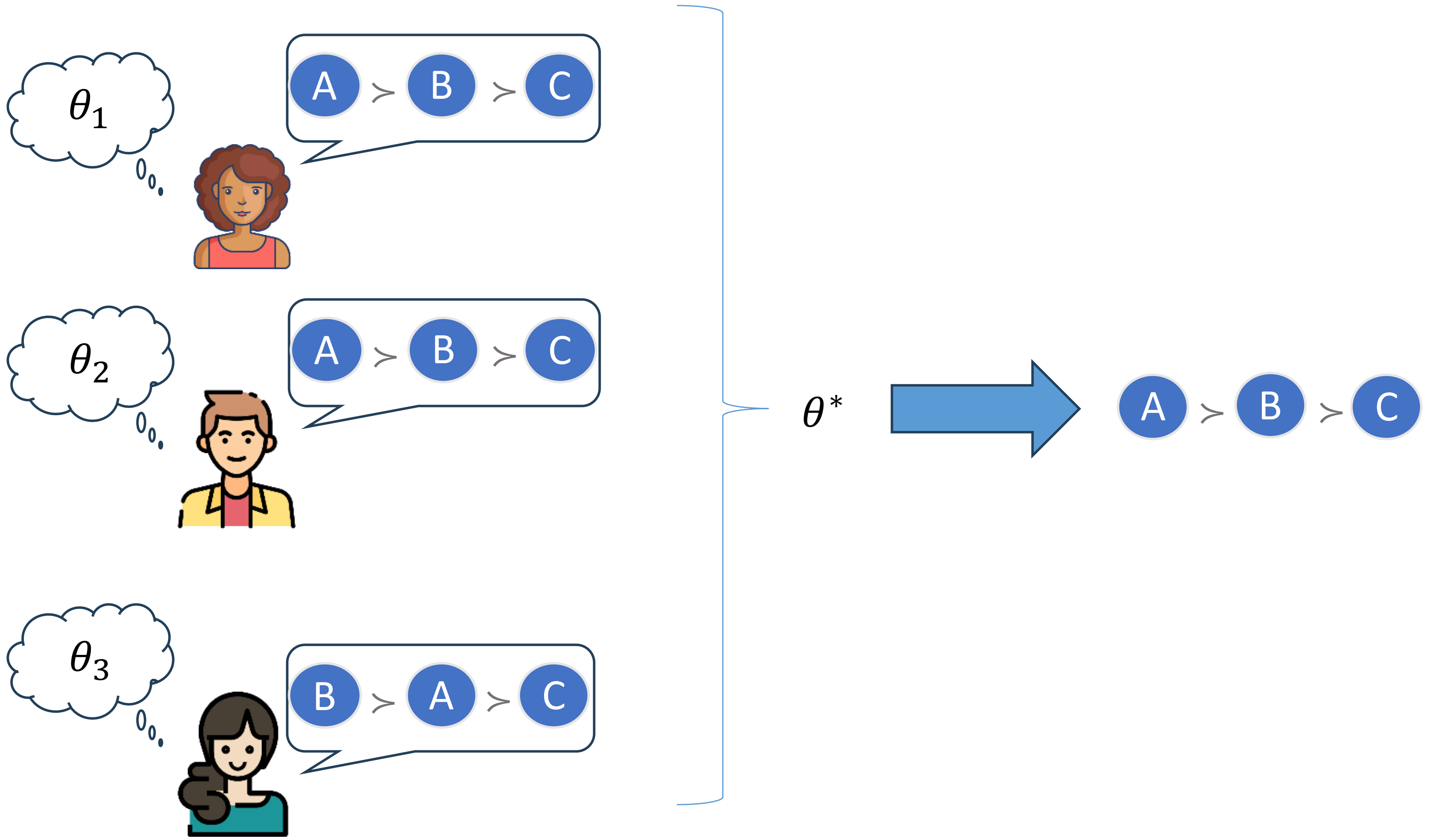
- **Theorem:** Leximax Copeland subject to PO **satisfies**
  a) PO
  b) PMC
  c) majority consistency
  d) winner monotonicity

  and can be implemented in polynomial time by solving $O(m^2)$ small linear programs

- **Majority Consistency:** A linear rank aggregation rule f satisfies majority consistency if when a candidate a is ranked first by a majority of voters in the input profile, a is ranked first in the output ranking

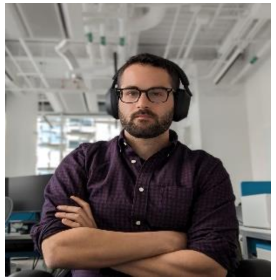
- **Winner Monotonicity:** A linear rank aggregation rule f satisfies winner monotonicity if, when a candidate a is ranked first in the output ranking, elevating a in any voter's preference does not cause a to lose their top position in the updated aggregate ranking

Heterogeneous Preferences



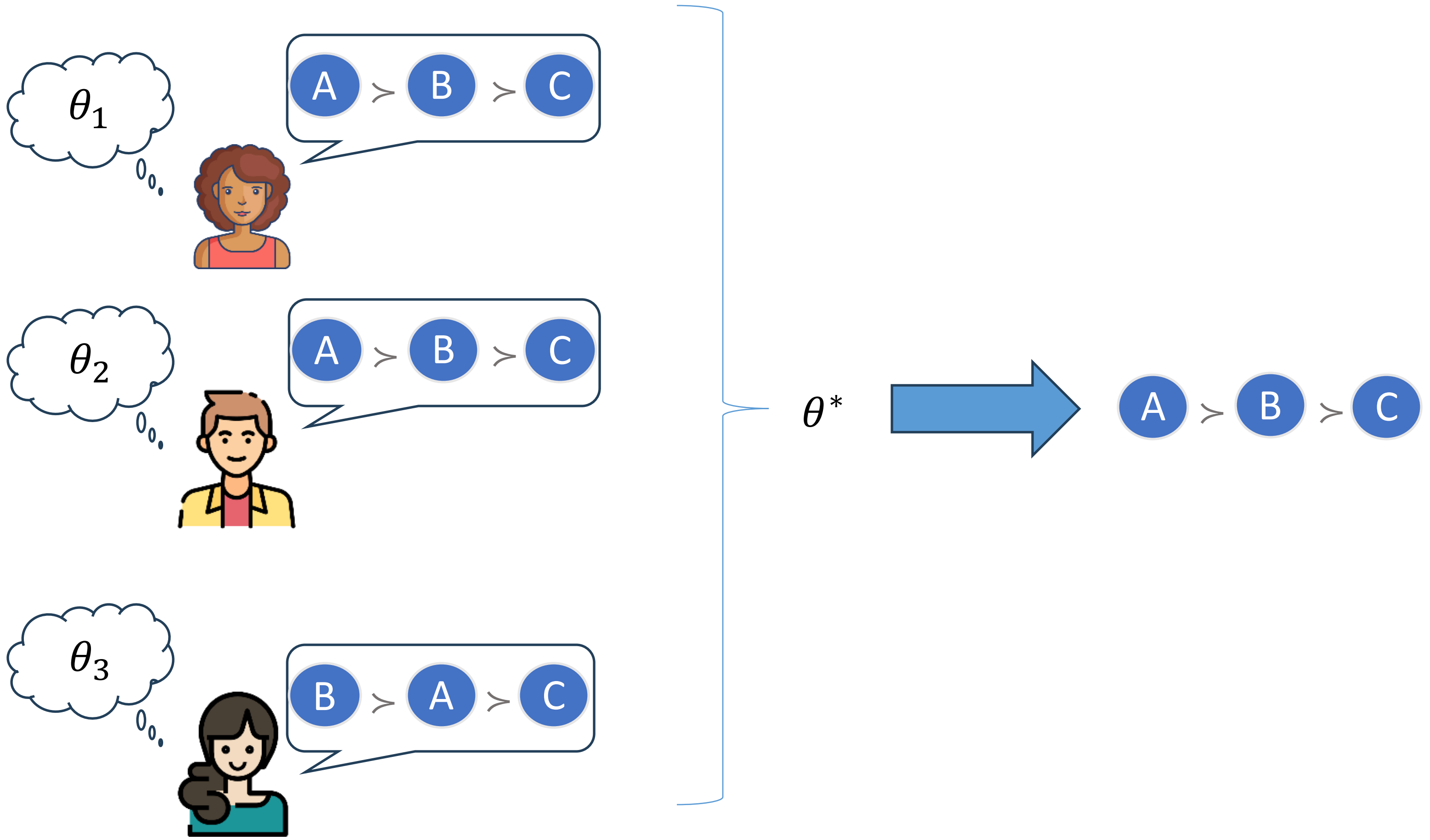
Pairwise-Calibrated Ensemble of Reward Functions



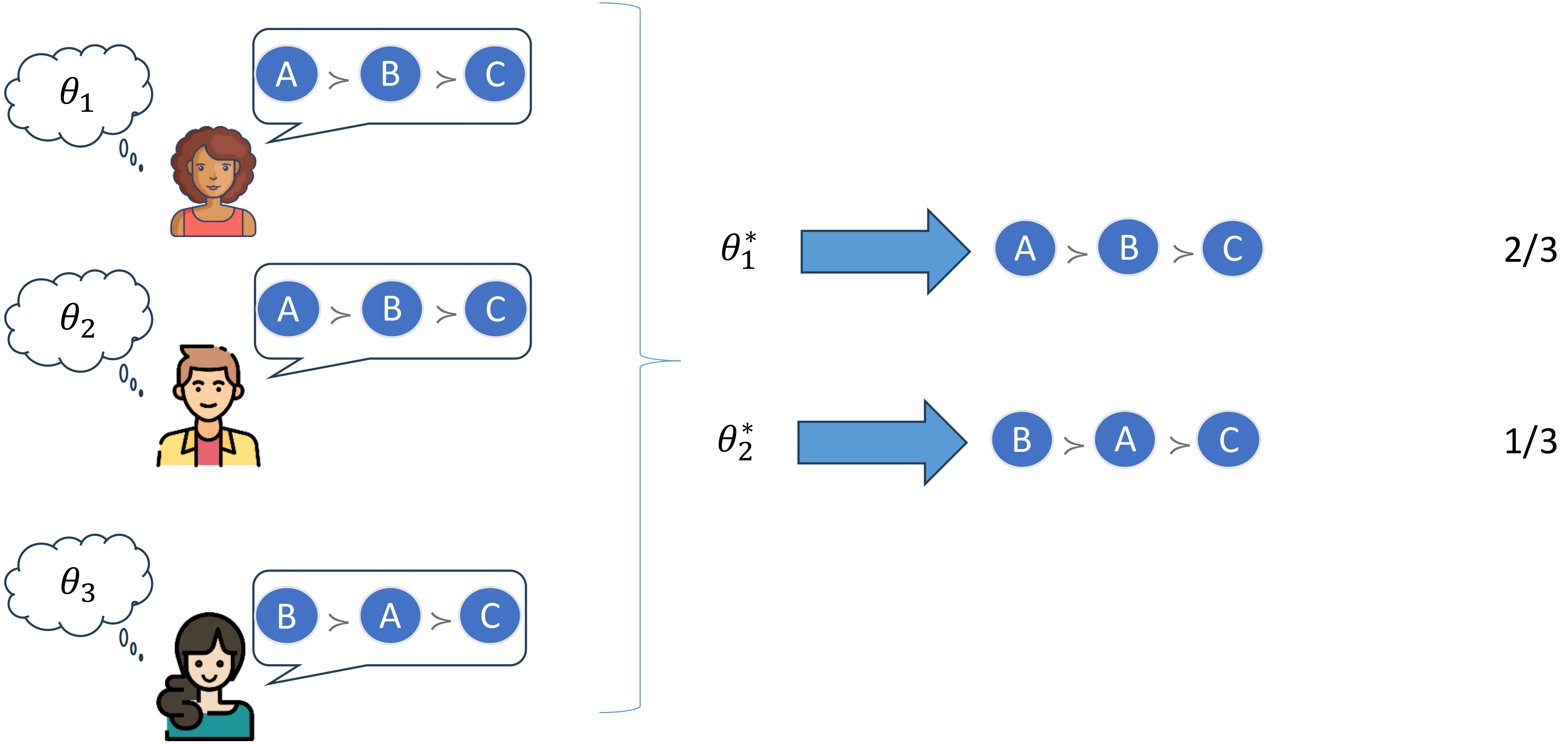


Daniel Halpern Ariel Procaccia Itai Shapira

Heterogeneous Preferences



Heterogeneous Preferences

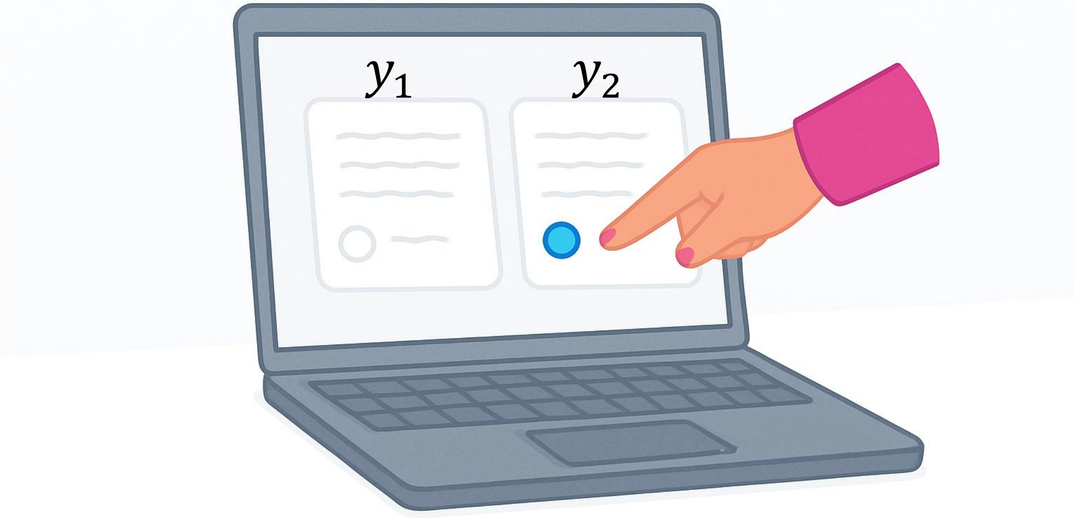


Pairwise-Calibrated Ensemble

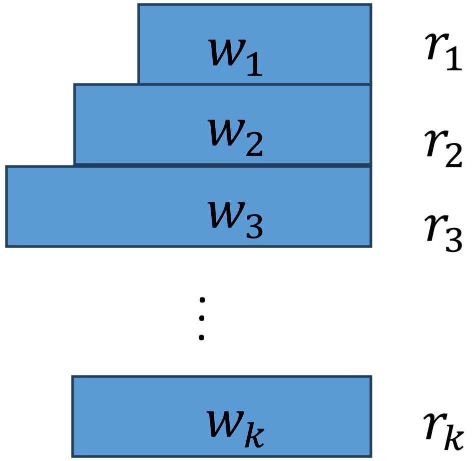
Pairwise Comparisons

$x \sim D_x$

$y_1, y_2 \sim \pi_0(\cdot\,|x)$



Ensemble of k reward functions



$$\Pr_{i\sim N}(y_1 \succ_i y_2|x) = \Pr_{r_j \sim D_w}\left(r_j(y_1|x) > r_j(y_2|x)\right)$$

Pairwise-Calibrated Ensemble

Position: A Roadmap to Pluralistic Alignment

Taylor Sorensen [1] Jared Moore [2] Jillian Fisher [1 3] Mitchell Gordon [1 4] Niloofar Mireshghallah [1]
Christopher Michael Rytting [1] Andre Ye [1] Liwei Jiang [1 5] Ximing Lu [1] Nouha Dziri [5] Tim Althoff [1]
Yejin Choi [1 5]

Abstract

With increased power and prevalence of AI systems, it is ever more critical that AI systems are designed to serve *all*, i.e., people with diverse values and perspectives. However, aligning models to serve *pluralistic* human values remains an open research question. In this piece, we propose a roadmap to pluralistic alignment, specifically using large language models as a test bed. We identify and formalize three possible ways to define and operationalize pluralism in AI systems: 1) *Overton pluralistic* models that present a spectrum of reasonable responses; 2) *Steerably pluralistic* models that can steer to reflect certain perspectives; and 3) *Distributionally pluralistic* models that are well-calibrated to a given population in distribution. We also formalize and discuss three possible classes of *pluralistic benchmarks*: 1) *Multi-objective* benchmarks, 2) *Trade-off steerable* benchmarks that incentivize models to steer to arbitrary trade-offs, and 3) *Jury-pluralistic* benchmarks that explicitly model diverse human ratings. We use this framework to argue that current alignment techniques may be fundamentally limited for pluralistic AI; indeed

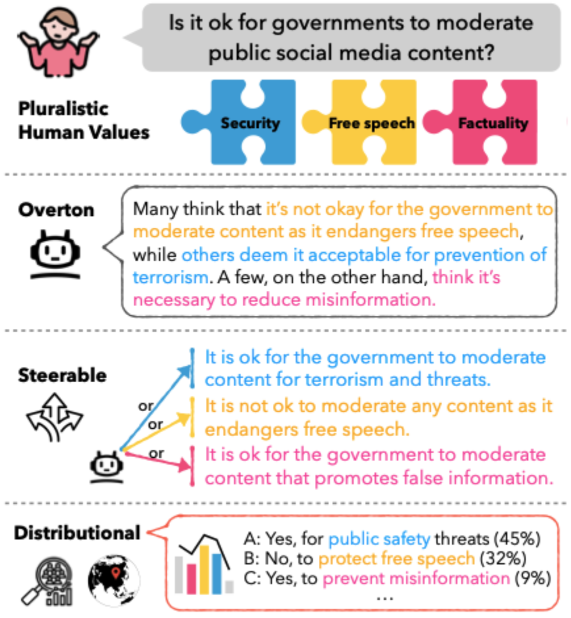


Figure 1. Three kinds of pluralism in models.

Theoretical Results

The goal is to design an ensemble that:

- Satisfies *pairwise calibration*
- Has *small support*
- Excludes *outliers*
    - No ranking has a Kemeny score significantly worse than the optimal ranking

- **Proposition:** A pairwise-calibrated ensemble with support $\min(m, n)$ always exists
- **Theorem:** Finding a pairwise-calibrated ensemble is an NP-hard problem
- **Theorem:** For any $\epsilon > 0$, there exists a ϵ-pairwise-calibrated ensemble with support $O(\epsilon^{-1})$
- **Theorem:** For any $\epsilon > 0$ and $\beta \geq 2$, there exists a $\left(\sqrt{\epsilon} + \frac{1}{\beta-1}\right)^2$ -pairwise-calibrated ensemble that does not contain $(\beta, (\beta + 1) \cdot \sqrt{\epsilon})$-outliers
- **Theorem (informal):** Pairwise calibration can be learned with a limited number of pairwise comparisons

Experiments

Name	Pref. Pairs	Unique Prompts	Annotation	Avg # Annots.
MultiPref	9,413	4,791 / 532	Human annotators	4.0
PersonalLLM	263,256	9,402 / 1,000	Model-based scores	10
HelpSteer2	21,000	10,000 / 1,000	Human annotators	3.5
Reddit TL;DR	3,217	729 / 845	Human annotators	7.56