Proportional Representation for Artificial Intelligence

Dominik Peters

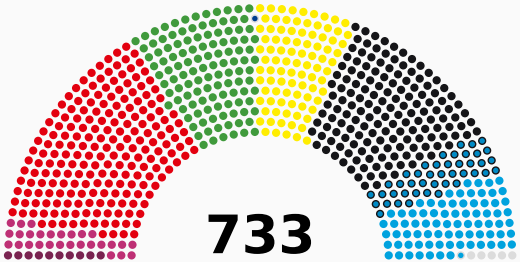
CNRS, LAMSADE, Université Paris Dauphine - PSL

2024-10-22
ECAI: Frontiers in AI
Santiago de Compostela

Overview

1. What is proportional representation?
2. A formal model of sequential decision making and proportionality.
3. Applications to emerging AI applications.

Proportional Representation



In politics, *proportional representation* refers to systems in which voters cast their ballot for a political party, and seats in parliament are allocated in proportion to vote count.

Goal: Parliament accurately reflects population.

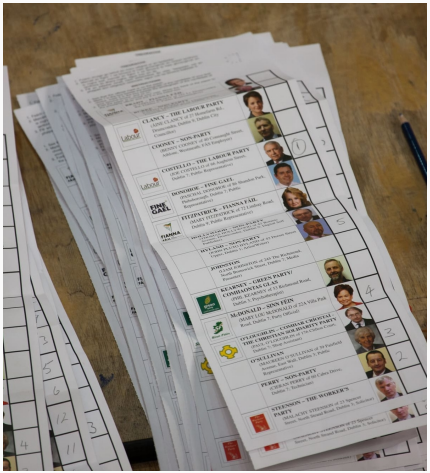
M. L. Balinski and H. P. Young. *Fair Representation: Meeting the Ideal of One Man, One Vote.* Yale University Press, 1982

F. Pukelsheim. *Proportional Representation: Apportionment Methods and Their Applications.* Springer, 2014

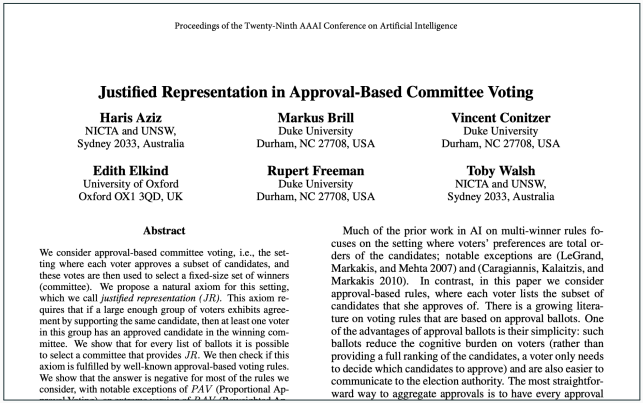
Proportional Representation: Ranking Candidates

But proportional representation also makes sense without parties: for example, in Ireland, voters rank candidates and the Single Transferable Vote (STV) rule ensures proportionality.

Goal: Each voter has approximately equal influence $\implies$ groups of voters with similar preferences have influence proportional to their size.



Proportional Representation: Formalization



Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Justified Representation in Approval-Based Committee Voting

Haris Aziz
NICTA and UNSW,
Sydney 2033, Australia

Markus Brill
Duke University
Durham, NC 27708, USA

Vincent Conitzer
Duke University
Durham, NC 27708, USA

Edith Elkind
University of Oxford
Oxford OX1 3QD, UK

Rupert Freeman
Duke University
Durham, NC 27708, USA

Toby Walsh
NICTA and UNSW,
Sydney 2033, Australia

Abstract

We consider approval-based committee voting, i.e., the setting where each voter approves a subset of candidates, and these votes are then used to select a fixed-size set of winners (committee). We propose a natural axiom for this setting, which we call *justified representation (JR)*. This axiom requires that if a large enough group of voters exhibits agreement by supporting the same candidate, then at least one voter in this group has an approved candidate in the winning committee. We show that for every list of ballots it is possible to select a committee that provides *JR*. We then check if this axiom is fulfilled by well-known approval-based voting rules. We show that the answer is negative for most of the rules we consider, with notable exceptions of *PAV* (Proportional Ap-

Much of the prior work in AI on multi-winner rules focuses on the setting where voters' preferences are total orders of the candidates; notable exceptions are (LeGrand, Markakis, and Mehta 2007) and (Caragiannis, Kalaitzis, and Markakis 2010). In contrast, in this paper we consider approval-based rules, where each voter lists the subset of candidates that she approves of. There is a growing literature on voting rules that are based on approval ballots. One of the advantages of approval ballots is their simplicity: such ballots reduce the cognitive burden on voters (rather than providing a full ranking of the candidates, a voter only needs to decide which candidates to approve) and are also easier to communicate to the election authority. The most straightforward way to aggregate approvals is to have every approval

Starting in 2015, AI researchers in *computational social choice* have started formalizing proportional representation as group fairness guarantees known as *justified representation* (JR) axioms, mostly studied for approval voting.

Haris Aziz et al. "Justified representation in approval-based committee voting". In: *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*. 2015, pp. 784–790

Proportional Representation: Approval Voting

Edvard Phragmén

Thorval N. Thiele

It turned out that two rules proposed by Phragmén and Thiele in the 1890s in Sweden satisfy strong versions of the JR axioms.

T. N. Thiele. “Om Flerfold Valg”. In: *Oversigt over det Kongelige Danske Videnskabernes Selskabs Fordhandlinger* (1895)

E. Phragmén. “Sur une méthode nouvelle pour réaliser, dans les élections, la représentation proportionnelle des partis”. In: *Öfversigt af Kongliga Vetenskaps-Akademiens Förhandlingar* 51.3 (1894), pp. 133–137

S. Janson. “Phragmén’s and Thiele’s election methods”. In: *arXiv:1611.08826* (2016)

Proportional Representation: Applications

Proportional representation can be applied to many collective decision making problems:

- Multi-winner voting ("choose k out of m candidates")

  > Martin Lackner and Piotr Skowron. *Multi-Winner Voting with Approval Preferences*. SpringerBriefs in Intelligent Systems. Springer, 2023. DOI: 10.1007/978-3-031-09016-5

- Aggregation of rankings

  > Patrick Lederer, Dominik Peters, and Tomasz Was. "The Squared Kemeny Rule for Averaging Rankings". In: *Proceedings of the 25th ACM Conference on Economics and Computation (EC)*. 2024

- Clustering

  > Xingyu Chen et al. "Proportionally fair clustering". In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. 2019, pp. 1032–1041

- *Participatory budgeting*

  > Dominik Peters, Grzegorz Pierczyński, and Piotr Skowron. "Proportional participatory budgeting with additive utilities". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 12726–12737

— Ballot Paper —

Total available budget: € 3 000 000.

Approve up to 4 projects.

- [x] Extension of the Public Library
  Cost: € 200 000
- [ ] Photovoltaic Panels on City Buildings
  Cost: € 150 000
- [x] Bicycle Racks on Main Street
  Cost: € 20 000
- [ ] Sports Equipment in the Park
  Cost: € 15 000
- [ ] Renovate Fountain in Market Square
  Cost: € 65 000
- [ ] Additional Public Toilets
  Cost: € 340 000
- [ ] Digital White Boards in Classrooms
  Cost: € 250 000
- [ ] Improve Accessibility of Town Hall
  Cost: € 600 000
- [x] Beautiful Night Lighting of Town Hall
  Cost: € 40 000
- [ ] Resurface Broad Street
  Cost: € 205 000

Participatory Budgeting: Project Selection

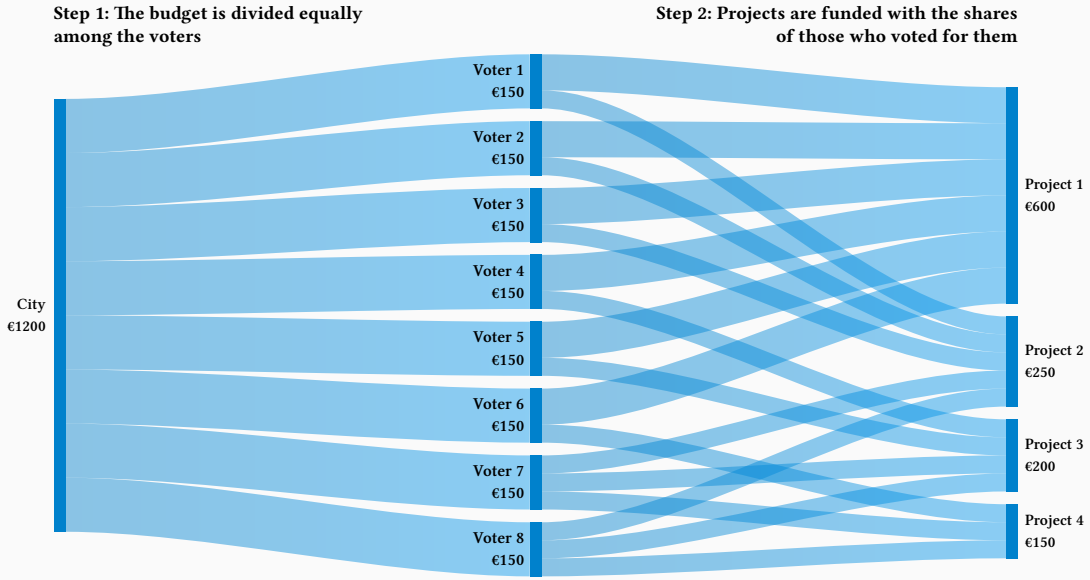
Given the votes, how to select the winning projects?

Standard method: Greedily take the most popular projects until money runs out.
Problem: too much money spent on similar projects in similar locations.

Alternative: The *Method of Equal Shares*

> Dominik Peters, Grzegorz Pierczyński, and Piotr Skowron. "Proportional participatory budgeting with additive utilities". In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 12726–12737

Participatory Budgeting: Standard Method vs. Method of Equal Shares



Participatory Budgeting: The Method of Equal Shares

2023:

Wieliczka
“Zielony milion”

Aarau
“Stadtidee"

Świecie

2024:

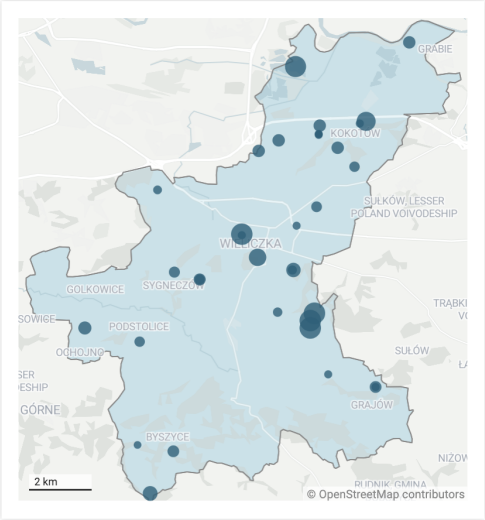
Assen
“Top Idea”

Winterthur
“Kultur Komitee”

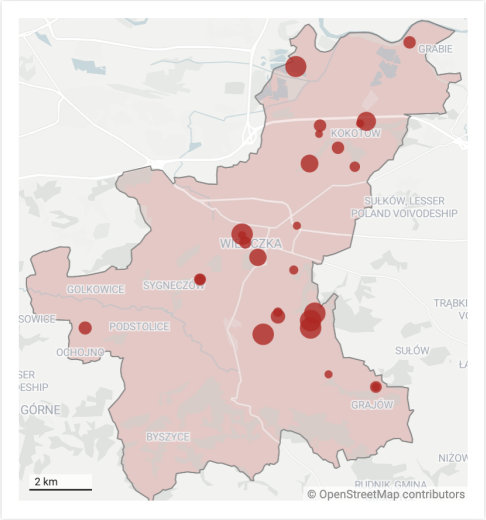
Powiat
Pruszków

More information:
https://equalshares.net/

Participatory Budgeting: The Method of Equal Shares



Method of Equal Shares



Standard voting method

Proportional Representation for AI

I will argue that proportionality can improve three emergent AI applications:

1. mixing the outputs of generative AI models such as LLMs,
2. training RLHF preference models based on labels from diverse raters,
3. the model of "virtual democracy" in which voters are represented by preference models that cast votes on their behalf.

Sequential Decision Making

These three applications build on a simple model of sequential decision making.

- $R = \{1, 2, \ldots, T\}$ is a set of T *rounds* (maybe online, maybe offline).
- In each round $j \in R$, a set C_j of *alternatives* is available.
- We need to make a *decision* $d_j \in C_j$ in each round j.
- $N = \{1, 2, \ldots, n\}$ is a set of *voters*.
- Each $i \in N$ *approves* a set $A_j^i \subseteq C_j$ of alternatives in each round $j \in R$.
  - Future work: generalize beyond 0/1 approval.

> Martin Lackner. "Perpetual Voting: Fairness in Long-Term Decision Making". In: *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*. 2020, pp. 2103–2110

> Nikhil Chandak, Shashwat Goel, and Dominik Peters. "Proportional Aggregation of Preferences for Sequential Decision Making". In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*. 2024, pp. 9573–9581

Example

Round	1	2	3	4
Voter 1	$\{a, b\}$	$\{a, b\}$	$\{a, b\}$	$\{a, b\}$
Voter 2	$\{a, c\}$	$\{a, c\}$	$\{a, c\}$	$\{a, c\}$
Voter 3	$\{d\}$	$\{d\}$	$\{e\}$	$\{e\}$
Voter 4	$\{d\}$	$\{d\}$	$\{f\}$	$\{f\}$

Example

Round	1	2	3	4
Voter 1	$\{a, b\}$	$\{a, b\}$	$\{a, b\}$	$\{a, b\}$
Voter 2	$\{a, c\}$	$\{a, c\}$	$\{a, c\}$	$\{a, c\}$
Voter 3	$\{d\}$	$\{d\}$	$\{e\}$	$\{e\}$
Voter 4	$\{d\}$	$\{d\}$	$\{f\}$	$\{f\}$
Greedy	a	a	a	a

Example

Round	1	2	3	4
Voter 1	$\{a, b\}$	$\{a, b\}$	$\{a, b\}$	$\{a, b\}$
Voter 2	$\{a, c\}$	$\{a, c\}$	$\{a, c\}$	$\{a, c\}$
Voter 3	$\{d\}$	$\{d\}$	$\{e\}$	$\{e\}$
Voter 4	$\{d\}$	$\{d\}$	$\{f\}$	$\{f\}$
Greedy	a	a	a	a
Phragmén	a	d	a	f
Thiele	d	d	a	a

Example

Round	1	2	3	4
Voter 1	$\{a, b\}$	$\{a, b\}$	$\{a, b\}$	$\{a, b\}$
Voter 2	$\{a, c\}$	$\{a, c\}$	$\{a, c\}$	$\{a, c\}$
Voter 3	$\{d\}$	$\{d\}$	$\{e\}$	$\{e\}$
Voter 4	$\{d\}$	$\{d\}$	$\{f\}$	$\{f\}$
Greedy	a	a	a	a
Phragmén	a	d	a	f
Thiele	d	d	a	a

What can we do with this model?

Merging Outputs of LLMs

How many emojis do you want in your response?

Please give recommendations what to do as a tourist in Santiago de Compostela.

As a tourist in Santiago de Compostela, be sure to explore the magnificent Santiago Cathedral, walk the historic streets of the Old Town, and enjoy local Galician cuisine, especially fresh seafood and traditional dishes like pulpo a la gallega.

Merging Outputs of LLMs

How many emojis do you want in your response?

none lots

Please give recommendations what to do as a tourist in Santiago de Compostela.

In Santiago de Compostela, ⭐ explore the iconic Cathedral ⛪, wander the charming Old Town streets 🏘️, snap pics of plazas 📸, taste delicious Galician octopus 🐙 and wine 🍷, enjoy the green parks 🌳, feel the spirit of the Camino pilgrims 🚶 🏞️, and shop for souvenirs 🎁 and local crafts 🧺!

Merging Outputs of LLMs

- Large language models such as GPT-4 generate text sequentially.
- At each step, they give a probability distribution over the next *token* (word fragment).

```
As a tourist in Santiago de Compostela, be sure to explore the
 magnificent Santiago Cathedral, walk the historic streets of the Old
 Town, and enjoy local Galician cuisine, especially fresh seafood and
 traditional dishes like pulpo a la gallega.
```

- There are many LLMs: different models (GPT-4, Claude, Gemini, Llama, etc.) each with different strengths and personalities.
- Even more via fine-tuning and via changing the system prompt.
- How can we merge their outputs?

Merging Outputs of LLMs

- Let's consider a collection of n LLMs, possibly with weights $w_1, \ldots, w_n$, each responding to the same prompt.
  - system prompts can differ
- Idea: Each token is a round, and each LLM votes for the tokens it thinks most likely.
- If we use Phragmén, it will mix the outputs according to the weights.
- Applications:
  - Compromise documents
  - Customizing style and tool use
  - Ethical decision-making
  - Avoiding hallucination
- Paper discusses interesting technical implementation challenges 💰

Reinforcement Learning from Human Feedback (RLHF)

- *Reinforcement learning from human feedback* (RLHF) is used by major AI labs to align and steer their LLMs.
- Human labelers are shown a prompt and possible responses to that prompt.
- They indicate their preferences over the responses via pairwise comparisons.
- Labels are then used to train a *preference model*.
- The preference model specifies rewards used in reinforcement learning.

> Paul F Christiano et al. "Deep Reinforcement Learning from Human Preferences". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017

RLHF: Problems

Recent survey about open problems in RLHF:

> *"RLHF is typically formulated as a solution for aligning an AI system with a single human, but humans are highly diverse in their preferences. Evaluators often disagree. Attempting to condense feedback from a variety of humans into a single reward model without taking these differences into account is thus a fundamentally misspecified problem."*

> Stephen Casper et al. "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback". In: *Transactions on Machine Learning Research* (2023). URL: https://openreview.net/forum?id=bx24KpJ4Eb

Recent position paper:

> *"Methods from social choice should be applied to address questions such as which humans should provide input and how it should be aggregated and used."*

> Vincent Conitzer et al. "Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback". In: *Proceedings of the 41st International Conference on Machine Learning (ICML)*. 2024

Note: quotes edited for brevity.

Reinforcement Learning from Collective Human Feedback

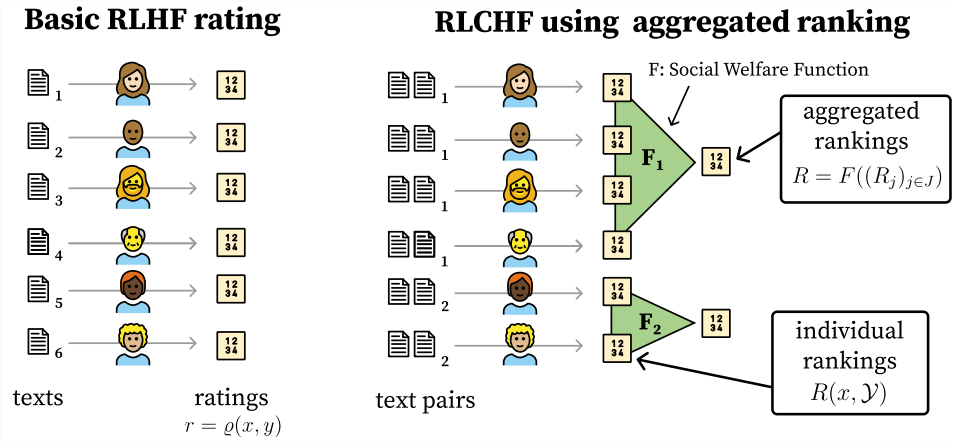


Figure from

> Vincent Conitzer et al. “Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback”. In: *Proceedings of the 41st International Conference on Machine Learning (ICML)*. 2024

RLCHF with Proportional Representation

- Caspar et al. note that when annotators disagree, "the majority wins, potentially disadvantaging under-represented groups".

> Stephen Casper et al. "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback". In: *Transactions on Machine Learning Research* (2023). URL: https://openreview.net/forum?id=bx24KpJ4Eb

- RLCHF does not address this issue, because each prompt is treated independently.
- Imagine 60% of raters dislike emojis, while 40% enjoy them 😍. The majority always votes against emoji-containing responses.

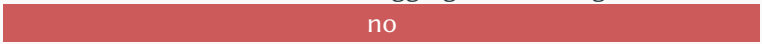
RLCHF with Proportional Representation

- Caspar et al. note that when annotators disagree, "the majority wins, potentially disadvantaging under-represented groups".

> Stephen Casper et al. "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback". In: *Transactions on Machine Learning Research* (2023). URL: https://openreview.net/forum?id=bx24KpJ4Eb

- RLCHF does not address this issue, because each prompt is treated independently.
- Imagine 60% of raters dislike emojis, while 40% enjoy them 😍. The majority always votes against emoji-containing responses.

- Standard social choice: 100% of aggregated rankings will advise against emojis! 🤯😡



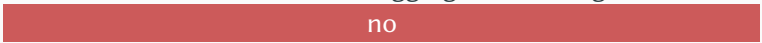
RLCHF with Proportional Representation

- Caspar et al. note that when annotators disagree, "the majority wins, potentially disadvantaging under-represented groups".

> Stephen Casper et al. "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback". In: *Transactions on Machine Learning Research* (2023). URL: https://openreview.net/forum?id=bx24KpJ4Eb

- RLCHF does not address this issue, because each prompt is treated independently.
- Imagine 60% of raters dislike emojis, while 40% enjoy them 😍. The majority always votes against emoji-containing responses.

- Standard social choice: 100% of aggregated rankings will advise against emojis! 🤯😡



- Idea: use a proportional aggregation method, where each prompt is a "round".
  $\Longrightarrow$ use emojis on 40% of prompts.

Virtual Democracy

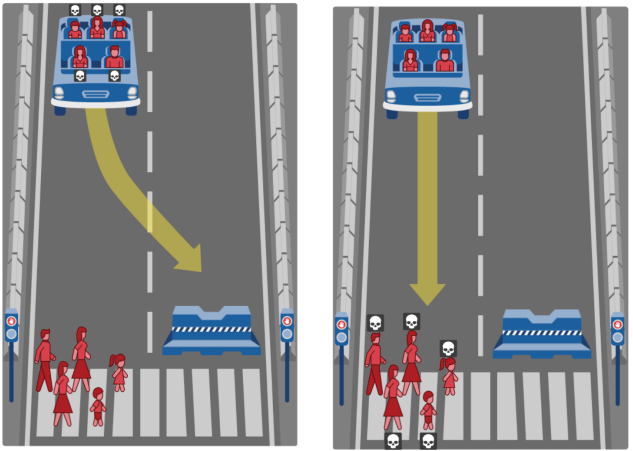
- Most natural way to combine social choice theory with AI agents is to use AI to let voters "outsource" the tasks of forming and reporting preferences.
- Each voter trains a personal preference model.
- Useful when a group of people need to make an extremely large number of decisions.
- This idea has been termed *virtual democracy*.
- Has been applied to kidney exchange and allocating food donations.

Ritesh Noothigattu et al. "A voting-based system for ethical decision making". In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*. 2018, pp. 1587–1594. DOI: 10.1609/aaai.v32i1.11512

Rachel Freedman et al. "Adapting a kidney exchange algorithm to align with human values". In: *Artificial Intelligence* 283 (2020), p. 103261

Min Kyung Lee et al. "WeBuildAI: Participatory framework for algorithmic governance". In: *Proceedings of the ACM on Human-Computer Interaction (HCI)* 3 (2019)

Virtual Democracy for the Moral Machine



Edmond Awad et al. “The moral machine experiment”. In: *Nature* 563.7729 (2018), pp. 59–64

Virtual Democracy for the Moral Machine

- **As a thought experiment**, let's consider how the car could make ethical decisions by letting humans from around the world vote over what's the right action.
- I'm not advocating actually doing this.
- Each user gave 14 pairwise comparisons, not enough.
- So we treat users from the same country as the same person and learn a preference model on their responses.
- Voters = countries.

Ritesh Noothigattu et al. "A voting-based system for ethical decision making". In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*. 2018, pp. 1587–1594. DOI: 10.1609/aaai.v32i1.11512

Nikhil Chandak, Shashwat Goel, and Dominik Peters. "Proportional Aggregation of Preferences for Sequential Decision Making". In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*. 2024, pp. 9573–9581

Virtual Democracy: Tyranny of the Majority?

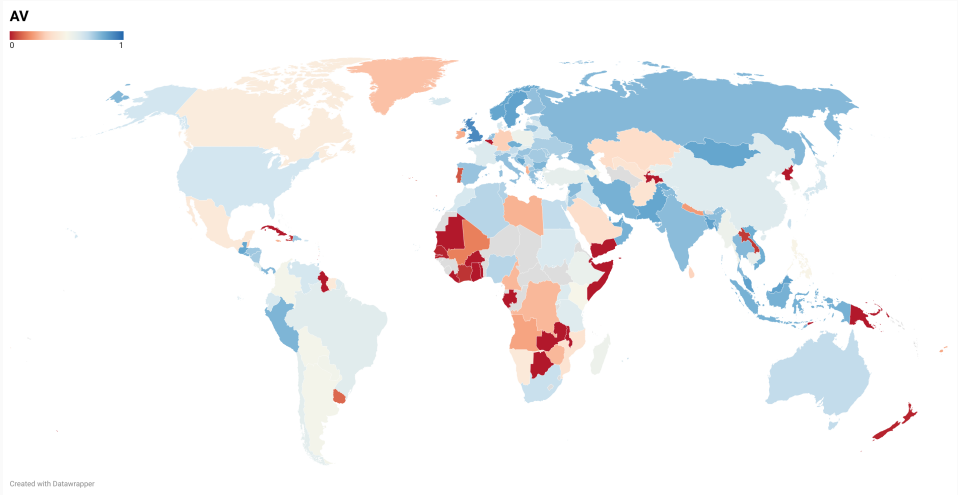
- Experiment: generate a sequence of dilemmas, and predict the vote of each country.
- Then, analogously to Noothigattu et al., take the most commonly voted-for action.
- Problem: "tyranny of the majority" – majority view will be followed in every decision.

> Michael Feffer, Hoda Heidari, and Zachary C. Lipton. "Moral machine or tyranny of the majority?" In: *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*. 2023, pp. 5974–5982

- Idea: Use Phragmén proportional rules to make the decisions instead, so that every view is followed an appropriate fraction of time.

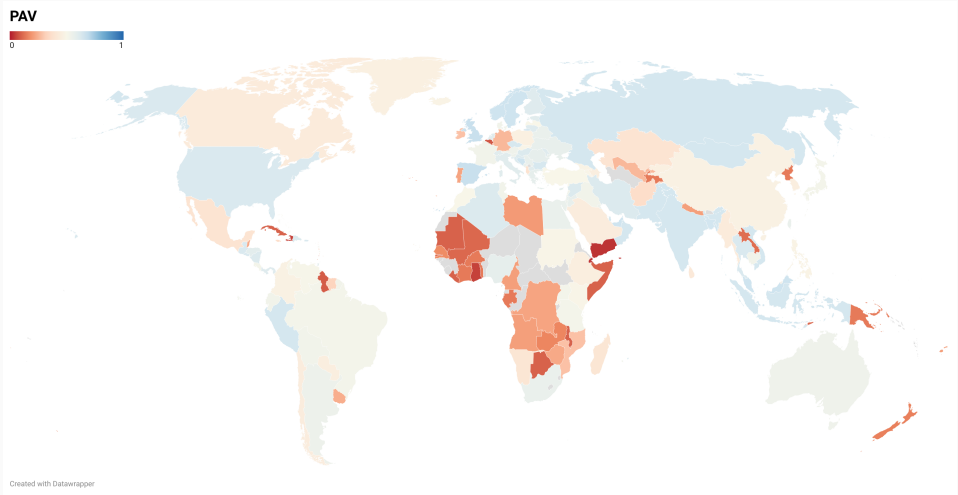
> Nikhil Chandak, Shashwat Goel, and Dominik Peters. "Proportional Aggregation of Preferences for Sequential Decision Making". In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*. 2024, pp. 9573–9581

Virtual Democracy for the Moral Machine: Majority



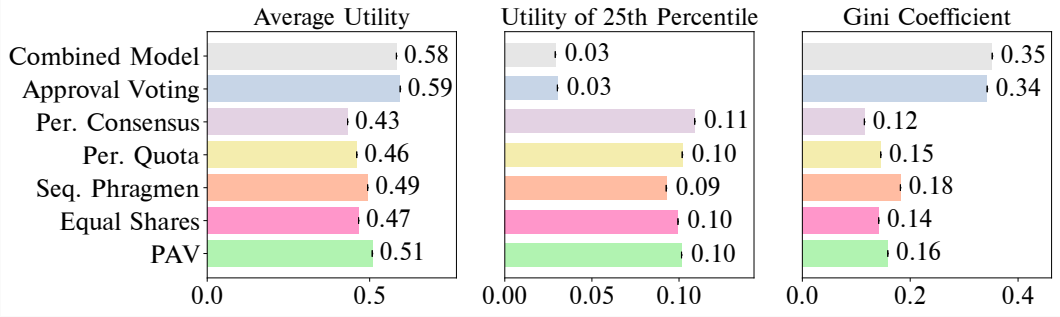
Nikhil Chandak, Shashwat Goel, and Dominik Peters. “Proportional Aggregation of Preferences for Sequential Decision Making”. In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*. 2024, pp. 9573–9581

Virtual Democracy for the Moral Machine: Proportionality



Nikhil Chandak, Shashwat Goel, and Dominik Peters. “Proportional Aggregation of Preferences for Sequential Decision Making”. In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*. 2024, pp. 9573–9581

Virtual Democracy for the Moral Machine: Comparison



Nikhil Chandak, Shashwat Goel, and Dominik Peters. "Proportional Aggregation of Preferences for Sequential Decision Making". In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*. 2024, pp. 9573–9581

Proportional Representation for Artificial Intelligence

Dominik Peters

CNRS, LAMSADE, Université Paris Dauphine - PSL

2024-10-22
ECAI: Frontiers in AI
Santiago de Compostela