Proportional Representation for Artificial Intelligence

Dominik Peters[a,*]

[a]CNRS, LAMSADE, Université Paris Dauphine - PSL
ORCID (Dominik Peters): https://orcid.org/0000-0001-9418-7571

Abstract. In recent years, the computational social choice community has increasingly studied the topic of proportional representation. This topic is particularly relevant for political elections, with many countries basing their voting systems on this principle, especially in Europe. However, the ideas behind proportional representation are also relevant in many other domains, including applications in artificial intelligence. We discuss a model of sequential decision making with proportional rules, and how it can be used for three applications: for merging the outputs of several large language models, for improving reinforcement learning from human feedback (RLHF), and for virtual democracy.

1 Introduction

In the realm of political elections, *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. This is typically achieved using *apportionment methods* which have become the subject of beautiful mathematical theory [38, 11, 64, 34]. Proportional representation can also be achieved in settings that allow voters to cast more expressive ballots. For example, the *Single Transferable Vote* (STV) is a method that allows voting for individual candidates instead of parties. It is used, for example, in Ireland [70]. Another class of methods based on approval voting were developed in the 1890s in Sweden by Phragmén and Thiele [69, 62, 39].

Starting around 2015, researchers in computational social choice have begun studying these rules formally [49], by defining axiomatic representation guarantees [6, 66, 18, 60], proving axiomatic characterizations and impossibilities [58, 44, 48], and by proposing new rules [59, 5, 7]. They have also extended the concept of proportional representation to new domains, such as voting for participatory budgeting [61, 19, 8, 65] and for budget allocation [9, 17, 31, 30, 2], as well as rank aggregation [67, 50] and online decision making [47]. In these contexts, proportional representation is usually formalized as saying that every cohesive group of agents must obtain a utility at least proportional to the group size. Thus, proportionality becomes a *group fairness guarantee* for all groups that have similar preferences.

In this informal article, I will argue that the principle of proportional representation can also be usefully applied to classic and emerging applications in Artificial Intelligence. An example of a classic AI topic is clustering, where these ideas lead to *proportionally fair clustering* [24, 53, 52, 42], in which any collection of sufficiently many similar data points is guaranteed to be matched to a nearby centroid. I will explain how proportionality can improve three emergent kinds of AI applications: (1) mixing the outputs of generative AI models, (2) training preference models (such as the ones used as part of reinforcement learning from human feedback (RLHF)) based on labels provided by diverse decision makers, and (3) the model of "virtual democracy" in which voters are represented by preference models that cast votes on their behalf.

For each of these applications, I will argue that a natural starting point for reasoning about proportional representation is provided by a model I recently studied with my coauthors Nikhil Chandak and Shashwat Goel at AAAI 2024, titled *Proportional Aggregation of Preferences for Sequential Decision Making* [23]. I will begin by briefly explaining our model and then explain how it can be used for the three AI applications mentioned.

2 Sequential Decision Making

We consider a set $R = \{1, 2, \dots, T\}$ of T *rounds*. In each round $j \in R$, we need to make a *decision*, by selecting exactly one alternative from a set C_j of alternatives available in round j. Depending on the context, the rounds may happen sequentially and we may not know in advance what alternatives will be available in each round (giving us an *online* setting that is also known as *perpetual voting* [46]), or the rounds could indicate separate issues that do not have a temporal character and that are known from the start (giving us an *offline* setting that is also known as *public decision making* [26, 4]).

We will want to make our decisions based on the preferences of a set $N = \{1, 2, \dots, n\}$ of *voters*. These preferences could take many forms, such as a cardinal (i.e., numerical) utility that each voter assigns to each alternative. However, in our initial work [23], we focus on the simple case of *approval votes*, where each voter just indicates in each round which alternatives the voter approves. This is equivalent to binary 0/1 utilities. We will represent these preferences as approval sets $A_j^i \subseteq C_j$, one for each agent $i \in N$ and each round $j \in R$, consisting of alternatives in round j approved by voter i.

A *voting rule* is given as input the sets of available candidates and all voters' approval sets, and must make a decision in each round, giving rise to a *decision sequence* $D = (d_1, \dots, d_T) \in C_1 \times \cdots \times C_T$. In the online setting, the information about the next round (i.e., the available alternatives and the approval sets) may only be revealed to the voting rule after the decision for the current round has been finalized. Given a decision D, for a voter $i \in N$, we write $U_D^i = |\{j \in R : d_j \in A_j^i\}|$ for the number of decisions in D that i approves, and we treat U_D^i as i's *utility*.

Phragmén's rule is a voting rule was first proposed in 1894 in the context of parliamentary elections [62, 39] but which can be naturally adapted to work for sequential decision making [47]. On a high level, this method works by keeping track how often each voter was happy

* Email: dominik.peters@lamsade.dauphine.fr. This is an invited contribution as part of the *Frontier in AI* series.

with decisions taken in the past, and in the next round will make a decision that is popular with voters who have not yet approved many decisions. We have shown that Phragmén's rule satisfies the axiom of *strong Proportional Justified Representation* (strong PJR) [21]. In particular, this axiom requires that if there is a group of 30% of all voters who are cohesive (in each round, they commonly approve at least one alternative) will be satisfied by the decision in at least 30% of the rounds. It provides appropriately weaker guarantees to groups who are cohesive in only some of the rounds.

3 Merging Outputs of LLMs

Large language models such as GPT-3 and GPT-4 [1, 20] work by generating text sequentially, at each step proposing a probability distribution over the next *token* (typically a fragment of a word). The models are pretrained on large text corpuses and later fine-tuned and adjusted via preference models to exhibit desirable behavior [22]. They can be further adjusted at runtime via their system prompts. These steps can lead to models with different "personalities" and different strengths.

One can imagine a variety of situations in which a user might want to use several of these models simultaneously to be able to combine their strengths. Some user interfaces allow this by sending the user prompt to several models and showing their outputs side-by-side. But one could instead offer to *combine* the outputs of the models by letting models *vote* over the next token. Specifically, given a user prompt, we could query each model for their suggested probability distribution of the next token and interpret this as a vote (for example, by letting the model "approve" all tokens above a certain threshold probability). Then a voting method such as Phragmén's rule would select the winning token. We then repeat the process by re-querying the models for the following token. Note that the different models can easily be assigned different weights according to user preferences.

Here are some scenarios in which this approach could prove useful:

- *Tool choice*: Products such as ChatGPT use models that have access to several *tools* such as code interpretation, image generation, and web searching. Users differ in their preferences regarding how often each of these tools should be applied. Using an appropriate system prompt, one could obtain an ensemble of models, one for each tool, each instructed to use their assigned tool as often as possible when it is appropriate. The user could then assign weights to the individual models to control the frequency with which each tool is used.
- *Compromise documents*: For drafting documents that should reflect a compromise between different members of a committee (e.g., a governmental body), one could instantiate a model for each member, instructing it to include as much as possible the views of that member in the produced document.
- *Ethical decision-making in AI*: For LLMs that are used as agents that take high-stakes actions automatically, it is important to ensure its decisions are ethical [16]. One could consider using models tuned to different ethical frameworks to generate responses to ethical questions, ensuring proportional representation of various moral perspectives.
- *Avoiding hallucinations*: In principle, one could hope that a voting approach would increase the overall capability and accuracy of the system. In particular it could reduce *hallucinations* [40] (responses containing false information presented as facts), for two reasons: First, models predicting the correct answer may assign higher confidence to their suggested tokens. Second, different hallucinating models would likely produce different completions, while those generating factual statements are likely to agree with each other and thus be boosted by the voting process.

In the first three suggested applications, using a proportional voting rule is important to accurately reflect the range of the models that are voting. For the last application, theoretically one might expect majoritarian rules to produce more accurate outputs; it could be worthwhile to compare proportional and majoritarian rules in this context.

There are many challenges to address, though. One class of challenges concerns the process of creating the individual voter models and ensuring that they behave as expected (correctly implementing their assigned "role"), which would involve some prompt engineering. For contexts involving ethical decision making, deciding on the weights of the models also involves difficult questions. But there are also several technical challenges with implementing this approach:

- *Unit of voting*: To perform the voting step, we need to decide whether to aggregate models' views on tokens or to use larger units such as words, sentences, or paragraphs. Using tokens is a natural choice given how the models work internally, but it will make it difficult for lower-weight models to generate coherent passages containing the views of that model, since they can be overruled by higher-weight models mid-sentence. Using tokens also makes it difficult to combine models that internally use different tokenizers (each training company tends to use its own tokenization scheme, and updates them over time). On the other hand, using larger units comes with its own problems. In particular, it would become unlikely that the same sentence would receive votes from more than one model (even if that is because they chose minor wording differences). This would severely reduce any efficiency gains from voting by weakening its ability to discover agreement between different perspectives.
- *Conditioning*: In LLMs, the predicted next token depends on all the tokens that came before it. Thus, the outcomes of the votes in prior rounds affect the voting patterns in future rounds. This is desirable to produce coherent outputs. However, consider a lower-weight model that was tasked with incorporating a certain view or a certain style in the outputs, but which "lost" in many of the previous rounds of voting. That model might look at the text that has been generated thus far and conclude that in this instance, it appears to have decided not to press its point of view. It might therefore tend to "vote with the majority" in future rounds. For a toy example, suppose one of the models has been instructed to enrich its answers with as many emojis as possible. However, it might happen that hundreds of tokens get generated without any emojis. The model might then prefer not to add emojis in future rounds, since this would reduce the consistency of the overall answer. It is unclear how best to deal with this effect, and how big it is.
- *Cost efficiency*: Running an ensemble of n LLMs will cost at least n times as much as running just a single model. In fact, it might cost a lot more, especially when depending on APIs. Common APIs don't allow adaptive querying and so require a new call for each round of voting, including retransmitting the entire context each time. It would be interesting to find cheaper methods to implement the voting approach.

A final technical challenge is that the proportional methods we have studied [23] take as input approval sets, while in the present context it would be more natural to directly take the token probabilities (logits) as input. Thus, it would be interesting to study generalizations of the sequential decision making model to such more expressive input types.

Overall, it would be exciting to develop some prototype applications to determine the feasibility of the approach and to get a better sense of how voting leads to changes in the outputs of LLMs.

4 RLHF

Reinforcement learning from human feedback (RLHF) [25] is a method used by the major AI labs to align and steer their large language models. A typical instantiation involves human labelers being presented with a prompt and possible responses to that prompt. They are asked to indicate their preferences over these responses, such as via pairwise comparisons. These labels are then used to train a *preference model* which is used to specify rewards used in the reinforcement learning process. RLHF aims to get the model to output responses that would have received favorable evaluations from the preference model, and hence from the human labelers.

The literature has identified several limitations and challenges to RLHF, as it is currently implemented. A recent survey about open problems in RLHF by Casper et al. [22] explains (Section 3.2.1):

> RLHF is typically formulated as a solution for aligning an AI system with a single human, but humans are highly diverse in their preferences, expertise, and capabilities [15, 57]. Evaluators often disagree: Stiennon et al. [68], Ouyang et al. [56], and Bai et al. [10] report annotator-annotator and annotator-researcher agreement rates from 63% to 77%, while Biyik and Sadigh [14] find distinct clusters of human feedback. 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. Moreover, current techniques model differences among evaluators as noise rather than potentially important sources of disagreement [13]. As a result, when preferences differ, the majority wins, potentially disadvantaging under-represented groups [63, 32, 43].

The authors of the survey [22] suggest that this issue could be "improved by algorithms that explicitly model multiple evaluators [35, 29, 28, 36, 12], that tune models to individuals [45], or that use more sophisticated aggregation strategies [55]". In a recent position paper by Conitzer et al. [27] titled "*Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback*", the authors argue that "methods from social choice should be applied to address questions such as which humans should provide input, what type of feedback should be collected, and how it should be aggregated and used." In particular, they suggest that incorporating a collective decision making step can address some of the shortcomings of RLHF when it comes to incorporating diverse preferences. Conitzer et al. [27] term this approach "Reinforcement Learning from Collective Human Feedback" (RLCHF). It involves letting *several* human raters rank responses to the same prompt and then aggregating their rankings. These aggregates are then used to train the preference model.

However, the RLCHF approach as described does not address the issue identified by Casper et al. [22] that the majority wins and minority groups may be under-represented. This is because the RLCHF proposal involves ranking aggregations that are done *independently* across prompts. Thus, a majority of raters with similar opinions might "win" in the aggregation again and again on many input prompts. For example, suppose that a majority of 60% the raters strongly dislikes emojis, while a minority of 40% enjoys them. The majority will always vote against emoji-containing responses, and standard social choice methods will implement majority wishes. Hence 100% of the aggregated rankings will advise against emojis, even though this is a position held by only 60% of the population.

To address this concern, one could use a proportional aggregation method. Each prompt becomes a "round" in which the raters give their preferences over possible responses. Crucially, in this approach, we keep track of how the same rater responded across different prompts, rather than treating them separately. In the toy example, we would expect that on about 40% of prompts, this approach would recommend the inclusion of emojis.

To understand whether and how proportional representation is useful for improving RLHF, a lot of work needs to be done. For starters, we need preference datasets where preference judgments are annotated by the identity of the labeler, so we can link their responses across prompts. Once such a dataset is assembled, we need to adapt or design aggregation rules that are compatible with the dataset. We can then test whether using proportional methods affect the behavior of the trained model, or whether perhaps proportionality goes "missing" in some step of the procedure, such as when fitting the preference model or during RL. On a more conceptual level, it would be interesting to think about what proportionality should mean in a setting like this, where "preferences" refer to a mixture of objective quality judgments (which response correctly answers the questions?) and subjective preferences (which response do I like more, or better reflects my values?). It would also be worth formally studying questions about elicitation of preference judgments, since each labeler can only consider a small number of prompts, and how elicitation interacts with the proportionality goal [37, 18].

5 Virtual Democracy

In the preceding two sections, we have discussed how social choice and proportional representation can be useful in the context of training and using large language models. But more fundamentally, the 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. These tasks can be handled by preference models that have been trained previously on the opinions of that voter. This is particularly interesting in cases where a group of people need to make an extremely large number of decisions, and where automation is the only way to do collective decision making at scale. This idea has been termed *virtual democracy* [55, 41].

The basic procedure is that we initially learn voters' preferences over a space of potential alternatives (specified by feature vectors), for example based on pairwise comparisons. After preference models are trained, each decision is made by letting the models vote on the decision maker's behalf, by predicting their preferences. This approach has led to proof-of-concept systems that automate moral decisions faced by autonomous vehicles [55], kidney exchanges [33], collective decision making directly from natural language preferences [54] and allocation of food donations [51]. However, a recent paper by Feffer et al. [32] points out that the voting rule adopted by many of these papers may lead to a "tyranny of the majority": if a majority of voters have similar preferences, those preferences will prevail at every future decision. In our paper on sequential decision making [23], we show preliminary evidence that using proportional voting methods could avoid this issue. We obtain this evidence by loosely following the experimental setup of Noothigattu et al. [55]. They used the moral machine dataset [3] which consists of moral judgment reported by millions of participants in the context of a self-driving car getting into a crash. They train a preference model for each participant and then let these models vote on future moral judgments. In our experiment,

we use proportional voting methods to implement the voting step, and find that this leads to decisions that take multiple perspectives into account, while Borda-style voting methods such as the one trialled by Noothigattu et al. [55] make majoritarian choices. Similarly, we find that if instead of training an individual preference model, we train a single preference model on a combined dataset of all responses, we again obtain more majoritarian decisions.

6 Conclusion

We have seen several possible applications of the principle of proportional representation to current topics of interest in artificial intelligence. I have suggested that voting rules for sequential decision making appear useful for many of these topics. I hope this inspires further work on this model, both theoretically and through evaluation of different rules on real data and models.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 technical report. *arXiv:2303.08774*, 2023.

[2] S. Airiau, H. Aziz, I. Caragiannis, J. Kruger, J. Lang, and D. Peters. Portioning using ordinal preferences: Fairness and efficiency. *Artificial Intelligence*, 314:103809, 2023. doi: 10.1016/j.artint.2022.103809.

[3] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan. The moral machine experiment. *Nature*, 563 (7729):59–64, 2018.

[4] H. Aziz and B. E. Lee. Sub-committee approval voting and generalized justified representation axioms. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 3–9, 2018. doi: 10.1145/3278721.3278739.

[5] H. Aziz and B. E. Lee. The expanding approvals rule: Improving proportional representation and monotonicity. *Social Choice and Welfare*, 54 (1):1–45, 2020. doi: 10.1007/s00355-019-01208-3.

[6] H. Aziz, M. Brill, V. Conitzer, E. Elkind, R. Freeman, and T. Walsh. Justified representation in approval-based committee voting. *Social Choice and Welfare*, 48(2):461–485, 2017.

[7] H. Aziz, E. Elkind, S. Huang, M. Lackner, L. Sánchez-Fernández, and P. Skowron. On the complexity of extended and proportional justified representation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pages 902–909, 2018.

[8] H. Aziz, B. E. Lee, and N. Talmon. Proportionally representative participatory budgeting: Axioms and algorithms. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 23–31, 2018.

[9] H. Aziz, A. Bogomolnaia, and H. Moulin. Fair mixing: The case of dichotomous preferences. *ACM Transactions on Economics and Computation*, 8(4):1–27, 2020. doi: 10.1145/3417738.

[10] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv:2204.05862*, 2022.

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

[12] P. Barnett, R. Freedman, J. Svegliato, and S. Russell. Active reward learning from multiple teachers. *arXiv:2303.00894*, 2023.

[13] C. Baumler, A. Sotnikova, and H. Daumé III. Which examples should be multiply annotated? Active learning when annotators may disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371, 2023.

[14] E. Biyik and D. Sadigh. Batch active preference-based learning of reward functions. In *Conference on Robot Learning*, pages 519–528, 2018.

[15] A. Bobu, A. Peng, P. Agrawal, J. Shah, and A. D. Dragan. Aligning robot and human representations. *arXiv:2302.01928*, 2023.

[16] J. S. Borg, W. Sinnott-Armstrong, and V. Conitzer. *Moral AI: And How We Get There*. Random House, 2024.

[17] F. Brandl, F. Brandt, D. Peters, and C. Stricker. Distribution rules under dichotomous preferences: two out of three ain't bad. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pages 158–179, 2021. doi: 10.1145/3465456.3467653.

[18] M. Brill and J. Peters. Robust and verifiable proportionality axioms for multiwinner voting. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 301–301, 2023.

[19] M. Brill, S. Forster, M. Lackner, J. Maly, and J. Peters. Proportionality in approval-based participatory budgeting. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, pages 5524–5531, 2023.

[20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[21] L. Bulteau, N. Hazon, R. Page, A. Rosenfeld, and N. Talmon. Justified representation for perpetual voting. *IEEE Access*, 9:96598–96612, 2021.

[22] S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, T. T. Wang, S. Marks, C.-R. Segerie, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=bx24KpJ4Eb.

[23] N. Chandak, S. Goel, and D. Peters. Proportional aggregation of preferences for sequential decision making. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)*, pages 9573–9581, 2024.

[24] X. Chen, B. Fain, L. Lyu, and K. Munagala. Proportionally fair clustering. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 1032–1041, 2019.

[25] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[26] V. Conitzer, R. Freeman, and N. Shah. Fair public decision making. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 629–646, 2017.

[27] V. Conitzer, R. Freedman, J. Heitzig, W. H. Holliday, B. M. Jacobs, N. Lambert, M. Mossé, E. Pacuit, S. Russell, H. Schoelkopf, E. Tewolde, and W. S. Zwicker. Social choice should guide AI alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://openreview.net/forum?id=w1d9DOGymR.

[28] O. Daniels-Koch and R. Freedman. The expertise problem: Learning from specialized feedback. *arXiv:2211.06519*, 2022.

[29] A. M. Davani, M. Díaz, and V. Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

[30] S. Ebadian, A. Kahng, D. Peters, and N. Shah. Optimized distortion and proportional fairness in voting. *ACM Transactions on Economics and Computation*, 12(1):1–39, 2024. doi: 10.1145/3640760.

[31] B. Fain, A. Goel, and K. Munagala. The core of the participatory budgeting problem. In *Proceedings of the 12th International Conference on Web and Internet Economics (WINE)*, pages 384–399, 2016. doi: 10.1007/978-3-662-54110-4_27.

[32] M. Feffer, H. Heidari, and Z. C. Lipton. Moral machine or tyranny of the majority? In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, pages 5974–5982, 2023.

[33] R. Freedman, J. S. Borg, W. Sinnott-Armstrong, J. P. Dickerson, and V. Conitzer. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, 283:103261, 2020.

[34] P. Gölz, D. Peters, and A. D. Procaccia. In this apportionment lottery, the house always wins. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 562–562, 2022.

[35] M. L. Gordon, K. Zhou, K. Patel, T. Hashimoto, and M. S. Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[36] M. L. Gordon, M. S. Lam, J. S. Park, K. Patel, J. Hancock, T. Hashimoto, and M. S. Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[37] D. Halpern, G. Kehne, A. D. Procaccia, J. Tucker-Foltz, and M. Wüthrich. Representation with incomplete votes. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, pages 5657–5664, 2023.

[38] A. Hylland. Allotment methods: Procedures for proportional distribution of indivisible entities. *Norwegian School of Management (Handelshøyskolen BI)*, Working Paper 1990/11, 1978. URL https://www.sv.uio.no/econ/personer/vit/aanundh/upubliserte-artikler-og-notater/Allotment-Methods%5B1%5D.pdf.

[39] S. Janson. Phragmén's and Thiele's election methods. *arXiv:1611.08826*, 2016.

[40] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language

generation. *ACM Computing Surveys*, 55(12), 2023. doi: 10.1145/3571730.

[41] A. Kahng, M. K. Lee, R. Noothigattu, A. D. Procaccia, and C.-A. Psomas. Statistical foundations of virtual democracy. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 3173–3182, 2019.

[42] L. Kellerhals and J. Peters. Proportional fairness in clustering: A social choice perspective. *arXiv:2310.18162*, 2023.

[43] H. R. Kirk, B. Vidgen, P. Röttger, and S. A. Hale. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv:2303.05453*, 2023.

[44] B. Kluiving, A. de Vries, P. Vrijbergen, A. Boixel, and U. Endriss. Analysing irresolute multiwinner voting rules with approval ballots via SAT solving. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, pages 131–138. 2020.

[45] D. Kumar, P. G. Kelley, S. Consolvo, J. Mason, E. Bursztein, Z. Durumeric, K. Thomas, and M. Bailey. Designing toxic content classification for a diversity of perspectives. In *SOUPS@ USENIX Security Symposium*, pages 299–318, 2021.

[46] M. Lackner. Perpetual voting: Fairness in long-term decision making. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2103–2110, 2020.

[47] M. Lackner and J. Maly. Proportional decisions in perpetual voting. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, pages 5722–5729, 2023.

[48] M. Lackner and P. Skowron. Consistent approval-based multi-winner rules. *Journal of Economic Theory*, 192:105173, 2021.

[49] M. Lackner and P. Skowron. *Multi-Winner Voting with Approval Preferences*. SpringerBriefs in Intelligent Systems. Springer, 2023. doi: 10.1007/978-3-031-09016-5.

[50] P. Lederer, D. Peters, and T. Wąs. The Squared Kemeny rule for averaging rankings. *arXiv:2404.08474*, 2024.

[51] M. K. Lee, D. Kusbit, A. Kahng, J. T. Kim, X. Yuan, A. Chan, D. See, R. Noothigattu, S. Lee, A. Psomas, and A. D. Procaccia. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction (HCI)*, 3, 2019.

[52] B. Li, L. Li, A. Sun, C. Wang, and Y. Wang. Approximate group fairness for clustering. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 6381–6391, 2021.

[53] E. Micha and N. Shah. Proportionally fair clustering revisited. In *Proceedings of the 47th International Colloquium on Automata, Languages, and Programming (ICALP)*, 2020. doi: 10.4230/LIPIcs.ICALP.2020.85.

[54] F. Mohsin, L. Luo, W. Ma, I. Kang, Z. Zhao, A. Liu, R. Vaish, and L. Xia. Making group decisions from natural language-based preferences. In *Proceedings of the 8th International Workshop on Computational Social Choice (COMSOC)*, 2021.

[55] R. Noothigattu, S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, and A. D. Procaccia. A voting-based system for ethical decision making. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pages 1587–1594, 2018. doi: 10.1609/aaai.v32i1.11512.

[56] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[57] A. Peng, A. Netanyahu, M. K. Ho, T. Shu, A. Bobu, J. Shah, and P. Agrawal. Diagnosis, feedback, adaptation: A human-in-the-loop framework for test-time policy adaptation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 27630–27641, 2023.

[58] D. Peters. Proportionality and strategyproofness in multiwinner elections. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1549–1557, 2018.

[59] D. Peters and P. Skowron. Proportionality and the limits of welfarism. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 793–794, 2020. Full version arXiv:1911.11747.

[60] D. Peters, G. Pierczyński, N. Shah, and P. Skowron. Market-based explanations of collective decisions. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, pages 5656–5663, 2021.

[61] D. Peters, G. Pierczyński, and P. Skowron. Proportional participatory budgeting with additive utilities. In *Advances in Neural Information Processing Systems*, volume 34, pages 12726–12737, 2021.

[62] E. Phragmén. Sur une méthode nouvelle pour réaliser, dans les élections, la représentation proportionnelle des partis. *Öfversigt af Kongliga Vetenskaps-Akademiens Förhandlingar*, 51(3):133–137, 1894.

[63] V. Prabhakaran, A. M. Davani, and M. Diaz. On releasing annotator-level labels and information in datasets. *arXiv:2110.05699*, 2021.

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

[65] S. Rey and J. Maly. The (computational) social choice take on indivisible participatory budgeting. *arXiv:2303.00621*, 2023.

[66] L. Sánchez-Fernández, E. Elkind, M. Lackner, N. Fernández, J. A. Fisteus, P. Basanta Val, and P. Skowron. Proportional justified representation. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, pages 670–676, 2017.

[67] P. Skowron, M. Lackner, M. Brill, D. Peters, and E. Elkind. Proportional rankings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 409–415, 2017.

[68] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

[69] T. N. Thiele. Om flerfold valg. *Oversigt over det Kongelige Danske Videnskabernes Selskabs Fordhandlinger*, 1895.

[70] N. Tideman. The single transferable vote. *Journal of Economic Perspectives*, 9(1):27–38, 1995. doi: 10.1257/jep.9.1.27.