MAPPING SOCIAL CHOICE THEORY TO RLHF

Jessica Dai* and Eve Fleisig*
Department of Computer Science
University of California, Berkeley
{jessicadai,efleisig}@berkeley.edu

ABSTRACT

Recent work on the limitations of using reinforcement learning from human feedback (RLHF) to incorporate human preferences into model behavior often raises *social choice theory* as a reference point. Social choice theory's analysis of settings such as voting mechanisms provides technical infrastructure that can inform how to aggregate human preferences amid disagreement. We analyze the problem settings of social choice and RLHF, and identify differences between them that prevent well-known technical results in social choice from immediately applying to RLHF. We then redefine canonical desiderata from social choice theory for the RLHF context and discuss how they may serve as analytical tools for open problems in RLHF. Finally, we contextualize the role of social choice in the broader political theory literature on democracy and collective decision making.

1 INTRODUCTION AND RELATED WORK

Reinforcement learning from human feedback (RLHF) has recently emerged as a key technique for incorporating human values into AI models.[1] The central problem setting of RLHF, in which people provide preferences over options that are then used to determine global behavior, shares key similarities with scenarios studied under social choice theory (SCT).[2] Recent work has discussed some of the ways in which SCT can serve as a reference for analysis of RLHF, including direct application of social choice axioms to RLHF (Mishra, 2023) and indicating that RLHF implicitly optimizes for the Borda count (Siththaranjan et al., 2023). Open problems in RLHF identified in surveys by Casper et al. (2023) and Lambert et al. (2023), among others, include the difficulty of selecting evaluators, accounting for disagreement among evaluators and cognitive biases, non-representative sampled prompts, challenges of developing a single reward function for diverse users, measuring downstream impacts, and assumptions regarding the ease of quantifying or aggregating complex individual preferences. Both Casper et al. (2023) and Lambert et al. (2023) raise *social choice* as one potential way to analyze some of the problems they identify (see also Appendix A). However, differences in the RLHF setting mean that established technical results in SCT require adjustment to be applicable to RLHF. We outline core differences and similarities between the RLHF and SCT problem settings (§2). We then propose SCT-style axioms for RLHF given the differences in problem settings and discuss open problems raised by these axioms (§3). Finally, we discuss implications of our analysis for conceptualization of human preference models in the context of RLHF (§4).

2 LEARNING PROBLEMS: PREFERENCE MODELING AND SOCIAL CHOICE

We first outline each problem setting and their core differences (see Table 1). We notate *alternatives*, the options that voters or evaluators can choose, as $a \in \mathcal{A}$, with $\mathcal{A} \subseteq \mathbb{R}^d$ as the space of all possible alternatives. Each alternative a should be thought of as a *prompt-response pair*, not just a response.[3] Humans—often called evaluators (preference modeling) or voters (social choice)—are indexed with

[1] RLHF is the dominant paradigm for incorporating human preferences into language models, but this paper discusses learning human preferences from pairwise or k-wise comparisons of outputs more generally. Our analyses are not restricted only to using RL as the method of incorporating these preferences into model behavior.

[2] See Brandt et al. (2016) for a textbook treatment of SCT for a computational audience.

[3] We choose to consider alternatives at the level of prompt-response pairs, rather than individual responses per prompt. This is because even though RLHF annotations are pairwise comparisons for responses to the same prompt, the high-level goal is to learn a global reward function that can score any prompt-response pair.

	Social Choice	**RLHF/ Preference Modeling**
Space of alternatives	Fixed and finite (e.g., candidates in an election)	Evaluators see only samples from a structured but infinite space (e.g. all possible prompts and completions)
Inputs to process	Fixed set of voters; often assume full information over alternatives (e.g. each voter submits a ballot ranking all candidates)	Set of evaluators (not always representative of general population) give pairwise comparisons over subset of alternatives; no guarantee over which evaluators give feedback on which alternatives
Goal of process	A *single winner* (e.g. an elected official) or a set of top-k winners	A *reward function* that measures the quality of any alternative, even if unseen (in the election example, a way to quantify not only the goodness of all the current candidates, but also of any new candidate)
Evaluation	Preferences of fixed set of voters over fixed alternatives evaluated based on *axioms* (analyze properties of voting rule) or *distortion* (measure utility of outcome)	*Generalization*, as measured by accuracy of the reward model or utility/regret of a policy using the reward model, with respect to all possible prompts/completions

Table 1: Summary of major differences in problem settings of social choice and preference modeling for LLMs.

i, with n total evaluators. The set of actual preferences (votes) is denoted $\{\pi\}_{i\in[n]}$, with π_i, a set of pairwise comparisons, as the preferences from voter i. For $a, a' \in \mathcal{A}$, $a \succ a'$ denotes "a is preferred to a'."

Human Preference Modeling The goal of human preference modeling[4] is to learn a reward model $r : \mathcal{A} \to \mathbb{R}$ that quantifies the desirability of a particular alternative a. The standard data collection method is via pairwise comparisons of text completions for a given prompt, and the standard noise model assumed is a Bradley-Terry model, where the likelihood of *observing* an instance of a being preferred to a' is proportional to how much "better" a is than a', i.e. $p(a \succ a') = \frac{\exp(r(a))}{\exp(r(a))+\exp(r(a'))}$.

Social Choice In the standard social choice setting, given a set of alternatives and a set of voters, we receive a ballot from each voter i quantifying the voter's preferences over all $\mathcal{A}$ alternatives. Commonly, this takes the form of *rankings*, where π_i gives voter i's rankings of all alternatives. A *voting rule* aggregates all n ballots to produce a single election winner. Voting rules can be evaluated *axiomatically*, i.e. with respect to whether they satisfy particular axioms (properties of the aggregation process), or in terms of *distortion*, i.e. with respect to the (aggregate) benefit derived from selecting the winner over other alternatives.

Properties and goals of the learning problem Common assumptions of the social choice problem setting include: that alternatives are *fixed* and finite, i.e. that $\mathcal{A}$ contains the only options that could possibly be considered; that we have full information from each voter about each alternative; and that voters and their preferences are fixed (see also Appendix A). By contrast, in RLHF, the space of alternatives is structured but infinite. The evaluators, who may not be representative of the full population, give pairwise or k-wise preferences over subsets of alternatives, often from a handpicked set of prompts. Each evaluator only sees a small subset of alternatives, and each alternative is only seen by a small subset of evaluators.

In SCT, common goals of the voting process include selecting a single winner (e.g. an elected official) or a set of top-k winners. By contrast, the goal of RLHF is to produce a reward model that can score the quality of new alternatives. That is, alternatives are not only ranked but assigned real-valued rewards, and the assignment of these rewards must generalize to unseen alternatives.

[4]See Lambert et al. (2023); Casper et al. (2023) for detailed surveys, and citations therein for mathematical derivations of the RLHF objective.

2.1 Defining the (Preference Modeling) Voting Rule

We propose a reformulation of the social choice problem that lets us interpolate between both settings.[5]

Consideration 1: Parameterization of alternatives and preferences. First, we assume that alternatives are parameterized in a d-dimensional space as $a \in \mathbb{R}^d$, with $\mathcal{A} \subseteq \mathbb{R}^d$. Recall that we let a represent a prompt and its completion; accordingly, we continue to assume that all alternatives are commensurable, i.e. that there is a reasonable and well-defined comparison across alternatives. We also parameterize the preferences of each voter i over text features as $\theta_i \in \mathbb{R}^d$ and model voter i's reward ascribed to a particular piece of text as $r_{\theta_i}(x) = \langle \theta_i, x \rangle$. In our model, voters have fixed preferences over features, and rewards for each piece of text are *scaled* by those preferences; see also Appendix A.

Consideration 2: Voters from a population. Instead of considering n fixed voters (evaluators), as in social choice, we assume all voters' preferences are drawn from some underlying population $\theta_i \sim \mathcal{V}$.[6] This allows us to model both the scenario in which there exists some shared societal norm from which individual preferences are drawn, and more complicated distributions (e.g. *mixtures* of preference models; Zhao et al. (2016; 2018); Liu & Moitra (2018) discuss mixtures of Bradley-Terry and random utility models).

With these considerations in mind, we present the preference modeling voting rule. Though the definition can simply be interpreted as a reward model, our statement is intended to emphasize the interplay between how the voting rule aggregates *seen* preferences and how it evaluates *unseen* alternatives.

Definition 2.1 (Preference Modeling Voting Rule)**.** A *preference modeling voting rule* $f : \mathbb{R}^d \to \mathbb{R}$ is a function that maps some (parameterized) alternative $a \in \mathbb{R}^d$ to some real-valued score, which represents the population's assessment of the quality of a.

3 Evaluating the (Preference Modeling) Voting Rule

How can social choice inform preference model evaluation? We argue that preference modeling can be divided into two subproblems: that of generalizing a particular set of preferences to new outputs; and that of deciding how to aggregate preferences over outputs. RLHF research has focused almost exclusively on the generalization problem. This perspective, though reasonable under the assumption that evaluators do not meaningfully disagree on their preferences, does not account for meaningful and widespread disagreement between evaluators, as evidenced by work such as Aroyo et al. (2023). When evaluators disagree, two perspectives from SCT can help to analyze problems that arise in RLHF: axiomatic approaches and distortion. We discuss the tenets and potential contributions of these three perspectives–generalization, axiomatic approaches, and distortion–in this section.

3.1 Perspective 1: Generalization

Following standard approaches in statistics and machine learning theory, recent work (Zhu et al., 2023; Wang et al., 2020) studies the problem of estimating $\hat{\theta}$ under Bradley-Terry models where $r_\theta(x) \propto \langle \theta, x \rangle$. These approaches take the alternatives to be fixed and predetermined, and all randomness is due to Bradley-Terry. Even under the goal of estimating $\hat{\theta}$ well (or minimizing the regret of a policy that would use an estimated $\hat{\theta}$, as analyzed in Zhu et al., 2023), natural extensions could consider the complication of psychological factors that bias observed preferences, such as preferences for "sycophantic" text or long outputs (Perez et al., 2023; Singhal et al., 2023); modeling more diverse evaluators; or analyzing the set of prompts to be annotated.

3.2 Perspective 2: Axiomatic Characterizations

A core tenet of social choice is that voting rules can be *axiomatically* analyzed; i.e., absent the notion of some "ground truth," there are particular principles that the final output should follow. This permits

[5] Siththaranjan et al. (2023) give an interpretation of the maximum-likelihood estimator (MLE) under the Bradley-Terry model as an implementation of Borda count, a standard voting rule from the social choice literature, though without explicitly making the modeling decisions we propose here. In future work, we hope to give a deeper analysis of this proposition based on the discussion in Sec. 3.

[6] Some work in SCT, e.g. the line of work in Dey & Bhattacharyya (2015), also considers a population of voters; most, however, assumes fixed voters.

a finer-grained understanding of how aggregations handle individual pieces of input. However, to apply axiomatic analysis to the preference modeling setting, we must still consider "generalization" in a sense not considered by standard SCT approaches.

On a technical level, axioms for the preference modeling setting must satisfy two criteria: (1) they must apply to scores, rather than single winners; and (2) they must distinguish between relationships that apply to the full space of alternatives $\mathcal{A}$ and all voters, and those that apply only to alternatives and ballots seen at train time. These properties mean that some canonical SCT axioms may be wholly inapplicable, while others may require careful reformulation to apply in the RLHF setting. We highlight examples of each of these cases below.

Example: unanimity, consistency, and Condorcet consistency. In the single-winner setting, an alternative $a \in \mathcal{A}$ is a *Condorcet winner* on a particular set of ballots if, for the majority of voters, $a \succ a'$ for all $a' \neq a$ and $a' \in \mathcal{A}$. A *Condorcet-consistent* voting rule for the single-winner setting always returns the Condorcet winner, if it exists. *Consistency* is satisfied if when, for every partition of voters, the voting rule selects the same winner, it holds that the voting rule over all the voters also selects that winner. *Unanimity* is satisfied when, if every individual voter expresses the preference $a \succ b$, then the voting rule also selects a over b.

Recall that, for RLHF, we are interested not in the final winner but in the scores that f assigns to arbitrary (unseen) alternatives, and that in this setting, arbitrarily small preferences may matter less than differences in reward greater than some margin ε. Also recall that voters are sampled from an underlying population. To account for these differences, we propose definitions for unanimity, consistency, and Condorcet consistency for the RLHF setting as follows:

Definition 3.1 (Unanimity for preference modeling)**.** First, define (a, ε)-unanimity as follows. For a fixed a (a potentially unseen alternative) and a population of (potentially unseen) voters $\mathcal{V}$, let $\mathcal{A}'_a \subseteq \mathcal{A}$ be the set of alternatives such that for all voters, $\langle \theta_i, (a - a') \rangle > \varepsilon$ for all $a' \neq a$ and $a' \in \mathcal{A}'_a$. Then a (preference modeling) voting rule f is (a, ε)-unanimous when $f(a) - f(a') > \varepsilon$ for all $a' \in \mathcal{A}'_a$, and ε-*unanimous* when it is (a, ε)-unanimous for all $a \in \mathcal{A}$. If f is ε-unanimous for $\varepsilon = 0$, then we say f is unanimous.

Definition 3.2 (Condorcet consistency for preference modeling)**.** We define (a, ε)-Condorcet consistency as follows. For a fixed a and voter population $\mathcal{V}$, let $\mathcal{A}'_a \subseteq \mathcal{A}$ be the set of alternatives such that $\mathbb{E}_{\theta_i \sim \mathcal{V}}[\langle \theta_i, (a - a') \rangle] > \varepsilon$ for all $a' \neq a$ and $a' \in \mathcal{A}'_a$. Then a (preference modeling) voting rule f is (a, ε)-Condorcet consistent when $f(a) - f(a') > \varepsilon$ for all $a' \in \mathcal{A}'_a$, and ε-*Condorcet consistent* when it is (a, ε)-Condorcet consistent for all $a \in \mathcal{A}$. If f is ε-Condorcet consistent for $\varepsilon = 0$, then we say f is Condorcet consistent.

Definition 3.3 (Consistency for preference modeling)**.** We define (a, ε)-consistency as follows. For a fixed a, let $\mathcal{A}'_a \subseteq \mathcal{A}$ be the set of alternatives such that for any sufficiently-large subset of voters $\mathcal{V}' \subseteq \mathcal{V}$, $f_{\mathcal{V}'}(a) - f_{\mathcal{V}'}(a') > \varepsilon$ for all $a' \in \mathcal{A}'_a$. Then a (preference modeling) voting rule f is (a, ε)-consistent when $f_{\mathcal{V}}(a) - f_{\mathcal{V}}(a') > \varepsilon$ for all $a' \in \mathcal{A}'_a$, and ε-*consistent* when it is (a, ε)-consistent for all $a \in \mathcal{A}$. If f is ε-consistent for $\varepsilon = 0$, then we say f is consistent.

Note the distinction between 3.1, 3.2, and 3.3 lies primarily in *what slice of voters* is being examined; the expectation in 3.2 captures the idea of majority preference, while 3.3 is concerned with agreement across subsets of voters. The fact that these axioms can be expressed in terms of which types of *agreement* are important to respect suggests that there is room for more explicit consideration of which types of disagreement are important in preference modeling.

Other axioms of (single-winner) social choice may not apply. *Resolvability* (that the voting rule produces no ties) or *majority* (that any alternative ranked first by a majority of voters should win) have no clear translation to the preference model setting, since choosing a single winner is no longer the main objective. *Strategyproofness*—the robustness of the final outcome to strategic behavior by individual voters—is a less pressing concern when the number of alternatives is very large, and when the final outcome is a scoring rule rather than a single winner.[7] Despite the seeming inapplicability of some "classic" SCT axioms, we argue that it is still worthwhile to consider what new axioms for preference modeling might look like: axioms were developed to establish more fundamental desiderata for democratic processes; moreover, they are designed to apply in cases that lack an obvious "ground truth" optimum.

[7]Even in the context of binary classification, Hardt et al. (2023) require *coordinated* action of around 10% of evaluators to substantially affect the output of the learned model.

3.3 Perspective 3: Distortion

An alternative approach to evaluating voting rules in social choice is through *distortion* (see Anshelevich et al. (2021) for a recent survey). At a high level, distortion is a quantitative notion of *suboptimality* with respect to some information that the voting rule may not be able to access (e.g. the "hidden context" of Siththaranjan et al. (2023). In an extension of the standard setting, voters have utilities over alternatives, and submit ballots—rankings—that are consistent with those true utilities. [8] The *distortion* of a voting rule is the worst-case difference in utility between the optimal election winner and the winner chosen by the voting rule, where the worst-case is taken over all possible utility functions that would have still been consistent with the observed ballots.

As with axiomatic characterizations, it is nontrivial to redefine distortion directly for RLHF. In SCT, it is reasonable to analyze "worst-case" (hidden) utilities due to the finiteness of both the alternatives and the voter, the assumption of no noise in reporting utilities, and because ballots tend to contain information about voters' preferences over the entire space of alternatives. For RLHF, therefore, we might consider an approach that captures the conceptual insight of the distortion metric rather than attempting to transliterate it directly. For example, we might consider some per-feature weight $w \in \mathbb{R}^d$ where w_j scales the importance of feature j. This could represent phenomena like cognitive bias in labeling, such that annotations are received according to a proxy reward function $r_{\theta,w}(a) = \sum_{j\in[d]} \theta_j w_j a_j$, but where true utilities depend only on θ and not w. On the other hand, this could also reflect the scenario where annotations are given with respect to $r_\theta = \langle \theta, a \rangle$ but utility for how accurately a particular feature θ_i is learned varies by feature (scaled by w_i)—for example, someone may prefer both longer text completions and text that avoids gendered stereotypes, but would care more about the final θ capturing the latter. To operationalize the "worst-case" in conjunction with the randomness in voting assumed by Bradley-Terry-Luce, we could work with expected welfare (where the expectation includes randomness in voting); assume some relationship between θ and w; and take the worst-case over possible θ and w that would be consistent with the observed votes. Interestingly, for the statistical inference problem of estimating θ, this is a fundamentally distinct perspective from the standard RLHF approach of maximum likelihood; instead, it is more analogous to estimating a θ that is robust to "worst case" realizations of hidden context.

4 Discussion

SCT provides a rich infrastructure for finer-grained discussion of many difficulties in using RLHF to represent human preferences by providing theoretical underpinnings for sources of these problems. These include conceptions of the prompt space, including how prompts for RLHF are chosen and what guarantees on representation may be lacking if the sample is handpicked in an unusual way; and of the evaluator space, including consequences of using an unrepresentative or small set of evaluators. In addition, they may help to conceptualize evaluator behavior: though potentially "misaligned" or "strategic" evaluators are often discussed (Casper et al., 2023), SCT provides a framework for analyzing the perhaps more common issues of evaluators dealing with poor incentives, incomplete directions, and cognitive biases (Singhal et al., 2023; Huang et al., 2023). Careful engagement is necessary to combine insights from these communities in a way that is both rigorous and fruitful. One naive interpretation of impossibility results such as Arrow's Theorem is that a version of democracy that relies on direct input from the public is simply untenable. However, a key aspect of SCT is that the axioms or properties of a voting mechanism that are actually desirable depend greatly on the context of the decision. By focusing on the properties that matter in the RLHF setting, and adapting SCT formulations to differences in the RLHF problem, we reach a space of problems that are both critical and tractable.

Acknowledgments

Many thanks to Bailey Flanigan, Paul Gölz, Nika Haghtalab, Cassidy Laidlaw, and Anand Siththaranjan for their helpful discussions and feedback.

[8] These utilities are hidden and perhaps unknown even to the voter, so they cannot be elicited directly.

REFERENCES

Elliot Anshelevich, Onkar Bhardwaj, Edith Elkind, John Postl, and Piotr Skowron. Approximating optimal social choice under metric preferences. *Artificial Intelligence*, 264:27–51, 2018. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2018.07.006. URL https://www.sciencedirect.com/science/article/pii/S0004370218304569.

Elliot Anshelevich, Aris Filos-Ratsikas, Nisarg Shah, and Alexandros A Voudouris. Distortion in social choice problems: The first 15 years and beyond. *arXiv preprint arXiv:2103.00911*, 2021.

Lora Aroyo, Alex S. Taylor, Mark Diaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. Dices dataset: Diversity in conversational ai evaluation for safety, 2023.

Gerdus Benade, Ariel D. Procaccia, and Mingda Qiao. Low-distortion social welfare functions. In *AAAI Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:53078427.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Palash Dey and Arnab Bhattacharyya. Sample complexity for winner prediction in elections. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1421–1430, 2015.

Sara Fish, Paul Gölz, David C Parkes, Ariel D Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. Generative social choice. *arXiv preprint arXiv:2309.01291*, 2023.

Bailey Flanigan, Daniel Halpern, and Alexandros Psomas. Smoothed analysis of social choice revisited. In *International Conference on Web and Internet Economics*, pp. 290–309. Springer, 2023.

Daniel Halpern, Gregory Kehne, Ariel D Procaccia, Jamie Tucker-Foltz, and Manuel Wüthrich. Representation with incomplete votes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5657–5664, 2023.

Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünner, and Tijana Zrnic. Algorithmic collective action in machine learning. *arXiv preprint arXiv:2302.04262*, 2023.

Olivia Huang, Eve Fleisig, and Dan Klein. Incorporating worker perspectives into mturk annotation practices for nlp. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1010–1028, 2023.

Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. The history and risks of reinforcement learning and human feedback. *arXiv e-prints*, pp. arXiv–2310, 2023.

Allen Liu and Ankur Moitra. Efficiently learning mixtures of mallows models. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 627–638. IEEE, 2018.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*, 2023.

David Miller. Deliberative democracy and social choice. *Political studies*, 40(1_suppl):54–67, 1992.

Abhilash Mishra. Ai alignment and social choice: Fundamental limitations and policy implications. *arXiv preprint arXiv:2310.16048*, 2023.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847.

Ariel D Procaccia and Jeffrey S Rosenschein. The distortion of cardinal preferences in voting. In *International Workshop on Cooperative Information Agents*, pp. 317–331. Springer, 2006.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf. *ArXiv*, abs/2312.08358, 2023. URL https://api.semanticscholar.org/CorpusID:266191810.

Piotr Skowron and Edith Elkind. Social choice under metric preferences: Scoring rules and stv. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Jingyan Wang, Nihar Shah, and R Ravi. Stretching the effectiveness of mle from accuracy to bias for pairwise comparisons. In *International Conference on Artificial Intelligence and Statistics*, pp. 66–76. PMLR, 2020.

Nils Wilde, Erdem Biyik, Dorsa Sadigh, and Stephen L Smith. Learning reward functions from scale feedback. In *Conference on Robot Learning*, pp. 353–362. PMLR, 2022.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *arXiv preprint arXiv:2306.01693*, 2023.

Zhibing Zhao, Peter Piech, and Lirong Xia. Learning mixtures of plackett-luce models. In *International Conference on Machine Learning*, pp. 2906–2914. PMLR, 2016.

Zhibing Zhao, Tristan Villamil, and Lirong Xia. Learning mixtures of random utility models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Banghua Zhu, Jiantao Jiao, and Michael I. Jordan. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons, 2023.

A Additional Related Work

Social choice Instead of applying social choice to AI, Fish et al. (2023) show that applying AI to augment democratic processes can improve canonical social choice results. In line with our discussion of application-specific reworkings of social choice axioms, Flanigan et al. (2023) uses smoothed analysis for relaxation of worst-case social choice axioms, which helps to distinguish voting rules that rarely satisfy axioms from those that often do.

Halpern et al. (2023) discuss relaxed assumptions on fixed preferences and full information. Fish et al. (2023) discusses relaxations of assumptions regarding fixed, finite, and commensurable preferences. The literature on *metric preferences* (see e.g. Skowron & Elkind (2017); Anshelevich et al. (2018)) also parameterizes alternatives in Euclidean space; however, voters' preferences over them are determined by their *distance* from each alternative.

In single-winner elections, the distortion of Borda count is unbounded (Procaccia & Rosenschein, 2006); however, numerical experiments from Benade et al. (2019) suggest that Borda count may be near-optimal in a setting where the voting rule outputs a ranking over alternatives rather than a single winner.

RLHF Related empirical RLHF work has raised alternative ways to collect human preferences, such as measuring the degree of preference (Wilde et al., 2022) or soliciting fine-grained preferences that compare alternatives along multiple dimensions (Wu et al., 2023). Alternative optimization methods to RLHF have also been proposed (e.g. Rafailov et al. (2023); Liu et al. (2023)).

Political theory For democratic theorists, social choice results over the last few decades have provided both challenges and clarity to normative discussions of what makes collective decision-making "legitimate." For example, *types* of choices have different normative consequences (e.g., choosing the best way to explain a math problem or the best way to discuss a conspiracy theory), which links to discussion of what decisions are best suited to different voting rules; for instance, Miller (1992) argue that Borda count often works well, but key normative decisions might demand majoritarianism. In the context of LMs, this distinction also helps to distinguish between *preferences* that can be personalized and universal *judgments* that must affect all users.