Representative Social Choice: From Learning Theory to AI Alignment

Tianyi Qiu
Center for Human-Compatible AI
University of California, Berkeley
qiutianyi.qty@gmail.com

Abstract

Social choice theory is the study of preference aggregation across a population, used both in mechanism design for human agents and in the democratic alignment of language models. In this study, we propose the *representative social choice* framework for the modeling of democratic representation in collective decisions, where the number of issues and individuals are too large for mechanisms to consider all preferences directly. These scenarios are widespread in real-world decision-making processes, such as jury trials, indirect elections, legislation processes, corporate governance, and, more recently, language model alignment. In representative social choice, the population is *represented* by a finite sample of individual-issue pairs, based on which social choice decisions are made. We show that many of the deepest questions in representative social choice can be naturally formulated as statistical learning problems, and prove the generalization properties of social choice mechanisms using the theory of statistical machine learning. We further formulate axioms for representative social choice, and prove Arrow-like impossibility theorems with new combinatorial tools of analysis. Our framework introduces the representative approach to social choice, and opens up research directions at the intersection of social choice, learning theory, and AI alignment.

1 Introduction

Social choice theory is a field of study that deals with the aggregation of individual preferences to form a collective decision. It has been applied in domains such as economics [Feldman and Serrano, 2006], political science [Miller, 1983, Coleman and Ferejohn, 1986], and computer science [Conitzer et al., 2024], to name a few. In these applications, the goal is to design mechanisms that aggregate individual preferences in a way that satisfies certain desirable properties, most especially fairness.

However, existing theoretical models in social choice theory tend to be simplistic and rely on relatively strong assumptions. Two such assumptions are (1) *independent, single-issue choices*, and (2) *complete information on all preferences of all individuals*. In practice, these assumptions are often violated. In common large-scale elections, candidates may have policies that are correlated across a huge number of different issues, and it is also infeasible to collect preferences of all voters on all issues, due to the large number of issues and the large number of voters involved.

These problems are not merely practical details that can be ignored. They are fundamental to the theory of social choice itself, since these complexities are exactly what give rise to democratic *representation* — the idea that individuals can delegate their decision-making power to a small number of representatives who can make decisions on their behalf, when there are too many issues and too many individuals to consider all preferences directly. The introduction of representation leads to fundamental questions for social choice theory that are not well-understood, such as the problem of *generalization* — how can we ensure that the decisions made by the representatives are representative

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

of the population's preferences, when the representatives are only chosen based on a small number of individuals' opinions on a small number of issues?

In this paper, we propose a new framework for social choice theory that models these complexities, which we call *representative social choice*. As we will show in the following sections, many of the deepest questions in representative social choice can be formulated as statistical learning problems, despite the seemingly different natures of the two fields. This connection allows us to leverage the rich theory of statistical machine learning to formulate axioms and mechanisms for representative social choice, and to analyze their properties.

Applications Variants of the representative social choice model can be applied in the modeling of all *collective decision-making processes involving representation and delegation*, which include most real-world decision-making processes. Below are some examples.

- *Jury Trials*: Citizens delegate their legal decision-making power to a randomly selected jury, which makes decisions on their behalf.
- *Legislation Processes*: Legislature body members, which can be viewed as representative samples of the citizen population, delegates the population's legal decicions to a set of laws that they decide on. We have a multi-issue setting, since the laws cover a wide range of issues. The space of possible collective preference profiles is the space of profiles implementable by laws.
- *Corporate Governance*: Shareholders elect a board of directors to make decisions on their behalf. We have a multi-issue setting, since the board decides on a wide range of issues during its service.
- *AI Alignment*: Frontier AI systems, including large language models (LLMs), undergo the *alignment* process during training, where they are trained to make decisions that are aligned with human values [Bai et al., 2022a,b], using preference datasets sampled from humans. Alignment can be viewed as a form of representative social choice, where the LLM is trained to make decisions that are representative of the human population's preferences, the latter represented by individual-issue pairs sampled from human evaluators (*i.e.*, the preference dataset). The space of possible collective preference profiles is the space of profiles that can be actualized as LLM policies — a *feature space* that has been the subject of much research in the field of statistical learning theory [Vapnik, 1999].

Related Work Social choice theory has had a long history [Satterthwaite, 1975, Young, 1975, Nisan and Ronen, 1999], with more recent research studying its intersection with machine learning [Fish et al., 2023, Parkes and Procaccia, 2013], and its applications in AI alignment [Conitzer et al., 2024, Köpf et al., 2024, Klingefjord et al., 2024, Huang et al., 2024, Prasad, 2018, Mishra, 2023, Ge et al., 2024]. Due to space constraints, we defer a detailed discussion of related work to Appendix A.

2 Problem Settings

In this section, we present the formal definitions of the representative social choice problem.

Issues We consider a discrete (but possibly infinite) set of N-ary issues $\mathcal{I}$, where each issue $i \in \mathcal{I}$ (*e.g.*, in a given state or province, which construction project to launch this year?) comes with N outcomes $[N] = \{1, 2, \cdots, N\}$. Each individual's preference profile can therefore be represented as a mapping from $\mathcal{I}$ to $\mathrm{LO}(N)$, where $\mathrm{LO}(N)$ is the set of linear orders over $[N]$.

We define a *saliency distribution* $\mathcal{D}_{\mathcal{I}}$ with full support over $\mathcal{I}$, which represents the importance of different issues, and decides the probability of each issue being sampled in the representation process. If there are a finite number of equally important issues, then $\mathcal{D}_{\mathcal{I}}$ is the uniform distribution over $\mathcal{I}$.

Population We consider a possibly infinite population, represented by a distribution $\mathcal{D}_{\mathcal{P}} \in \Delta[\mathcal{P}]$, where $\Delta[\mathcal{P}]$ is the space of probability distributions over the support set $\mathcal{P} \subseteq \mathrm{LO}(N)^{\mathcal{I}}$.[1] For any preference profile $C \in \mathcal{P}$, denote with $\mathcal{D}_{\mathcal{P}}(C)$ the probability (mass or density) that a random individual in the population has preference profile C over the outcomes of all issues.

Often, we only need to consider the marginal distribution of $\mathcal{D}_{\mathcal{P}}$ — the mapping $\mathcal{M} : \mathcal{I} \to \Delta[\mathrm{LO}(N)]$.[2] For any issue $i \in \mathcal{I}$, $\mathcal{M}(i)$ is the distribution of preferences over the N outcomes of

[1] In this paper, we denote with A^B the space of mappings from B to A.

[2] $\Delta[\mathrm{LO}(N)]$ is the space of probability distributions over $\mathrm{LO}(N)$

issue i in the population. For preference ordering $o \in \mathrm{LO}(N)$, we denote with $\mathcal{M}(i)_o$ the probability that a random individual in the population has preference ordering o over outcomes of issue i.

Outcomes The result of a decision-making process is a preference profile $C : \mathcal{I} \to \mathrm{LO}(N)$, which represents the aggregated preference of the population generated by some mechanism.

The mechanism is *representational* if the decision is made based on a finite collection of individual-issue pairs $\mathcal{S} = \{(o_1, i_1), (o_2, i_2), \ldots, (o_{|S|}, i_{|S|})\}$, with $i_k \in \mathcal{I}$ sampled from $\mathcal{D}_{\mathcal{I}}$, and $o_k \sim \mathcal{M}(i_k)$ is an individual's preference over the outcomes of issue i_k, sampled from the population distribution.

However, not all preference profiles are allowed. As a key feature of representative social choice, the mechanism is only allowed to output preference profiles from a limited *candidate space* $\mathcal{C} \subseteq \mathrm{LO}(N)^{\mathcal{I}}$, which represents the space of possible preference profiles that can be generated by the mechanism (*e.g.*, mutually compatible combinations of per-state construction projects in the national policy case, or language model policies in the AI alignment case). A mechanism is thus a function $f : (\mathrm{LO}(N) \times \mathcal{I})^* \to \mathcal{C}$ that maps the sample collection $\mathcal{S}$ to a preference profile $C \in \mathcal{C}$.

Binary Setting as Special Case In the binary ($N = 2$) case of representative social choice, each issue is binary (Yes/No), such as in the preference annotations of language model alignment [Bai et al., 2022a]. The main difference in the binary case is the reduction in complexity, allowing for analysis of a well-defined majority vote mechanism, as we will see later.

3 Key Results in Representative Social Choice

In this section, we present the key results from our analysis of representative social choice. Due to space constraints, we leave the details to Appendix B and C. Instead, this section selectively presents key results of our analysis, along with pointers to the detailed derivations in the appendices.

3.1 Generalization Bounds (Appendix B.2)

Generalization bounds are essential in representative social choice. They ensure that decisions made by the mechanism based on a finite sample of preferences are representative of the overall population. This brings us to the concept of *generalization error* — the gap between the performance of a mechanism on a sample versus the whole population.

Theorem (Binary Generalization Bound, Thm. 1)**.** *Let $\mathcal{C}$ be a candidate space with VC dimension $\mathrm{VC}(\mathcal{C})$, and let $\epsilon > 0$ be a desired generalization error. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the sample utility and population utility of* any *preference profile $C \in \mathcal{C}$ are ϵ-close, i.e.,*

$$\Pr\left[\left|\frac{1}{|\mathcal{S}|}\sum_{k=1}^{|\mathcal{S}|}\mathbf{1}_{C(i_k)=o_k} - \mathrm{E}_{i\sim\mathcal{D}_{\mathcal{I}}}[\mathcal{M}(i)_{C(i)}]\right| \le \epsilon, \quad \forall C \in \mathcal{C}\right] \ge 1 - \delta \tag{1}$$

as long as we have the following, for some constant $c > 0$:

$$|\mathcal{S}| \ge \frac{c}{\epsilon^2}\mathrm{VC}(\mathcal{C})\left(\log \mathrm{VC}(\mathcal{C}) + \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right) \tag{2}$$

See Theorem 2 for a generalization of this bound to non-binary settings.

In essence, the bound tells us that the larger the sample size, the better the mechanism's decision reflects the population's true preferences, with the sample size needing to grow in proportion to the complexity of the candidate space, as measured by the VC dimension [Vapnik, 1999] — for instance, when election candidates are allowed to tailor their messaging in a fine-grained manner, the population needs to watch more debates to find broadly aligned candidates, or else candidates could easily *overfit* their message to a few flagship issues.

3.2 Majority Vote and Scoring Mechanisms (Appendix B.3, C.2)

In the binary setting, one of the simplest and most effective mechanisms for preference aggregation is the majority vote. This mechanism selects the outcome that receives the majority of votes for each issue based on a finite sample of individual-issue pairs.

Corollary (Majority Vote Approximately Maximizes Population Utility, Cor. B.3)**.** *Under* (2)*, for any error $\epsilon > 0$ and confidence requirement $\delta > 0$, the majority vote mechanism f_{maj} has*

$$\Pr\left[U_{\mathcal{D}_{\mathcal{I}},\mathcal{M}}(f_{\mathrm{maj}}(\mathcal{S})) \geq -2\epsilon + \max_{C\in\mathcal{C}} U_{\mathcal{D}_{\mathcal{I}},\mathcal{M}}(C)\right] \geq 1-\delta \tag{3}$$

where the population utility

$$U_{\mathcal{D}_{\mathcal{I}},\mathcal{M}}(C) \coloneqq \mathrm{E}_{i\sim\mathcal{D}_{\mathcal{I}}}[\mathcal{M}(i)_{C(i)}], \tag{4}$$

This corollary shows that the majority vote mechanism works well under binary settings, where it approximately maximizes the population's utility. In settings where issues are no longer binary, majority vote becomes no longer well-defined. Instead, we turn to scoring mechanisms, which assign scores to preference profiles and select the profile with the highest average score.

Corollary (Scoring Mechanisms Approximately Maximize Population Score, Cor. C.2)**.** *When $|\mathcal{I}|$ is finite, for any scoring rule s with value bounded by constants, the scoring mechanism f_s satisfies*

$$\Pr\left[U^s_{\mathcal{D}_{\mathcal{I}},\mathcal{M}}(f_s(\mathcal{S})) \geq -2\epsilon + \max_{C\in\mathcal{C}} U^s_{\mathcal{D}_{\mathcal{I}},\mathcal{M}}(C)\right] \geq 1-\delta \tag{5}$$

as long as we have the following, for some constant $c > 0$:

$$|\mathcal{S}| \geq \frac{c}{\epsilon^2}\left(|\mathcal{I}|N\log N + \log\frac{1}{\delta}\right) \tag{6}$$

In multi-outcome cases, scoring mechanisms provide a more generally applicable approach than majority vote. Like majority vote, scoring mechanisms exhibit good generalization properties as long as the sample size is large enough and in proportion to the complexity of the candidate space.

3.3 Representative Impossibilities (Appendix C.3, C.5)

While the mechanisms we have discussed so far have desirable properties, they cannot satisfy all the axioms we might want from a social choice mechanism simultaneously. This brings us to impossibility theorems that generalize Arrow's famous result in classical social choice theory.

Theorem (Weak Representative Impossibility, Thm. 3)**.** *When $N \geq 3, \mathcal{C} = \mathrm{LO}(N)^{\mathcal{I}}$, no representational mechanism simultaneously satisfies probabilistic Pareto efficiency (PPE), weak probabilistic independence of irrelevant alternatives (W-PIIA), and weak probabilistic convergence (W-PC).*

Theorem (Strong Representative Impossibility, Thm. 4)**.** *For any $\mathcal{C}$, when there is at least one cyclically privileged issue (Definition C.10), no representational mechanism simultaneously satisfies PPE, strong probabilistic independence of irrelevant alternatives (S-PIIA), and strong probabilistic convergence (S-PC). This cyclicity condition is both sufficient and necessary for impossibility.*

These results show that, in representative social choice, we must make trade-offs between different desirable properties, such as fairness, utility maximization, and convergence. The two theorems differ in the strength of the axioms they consider, with the strong impossibility theorem takes into account interdependence between issues by lifting the constraint on the candidate space $\mathcal{C}$.

4 Conclusion

In this paper, we have formulated the problem of representative social choice, where a mechanism aggregates the preferences of a population based on a finite sample of individual-issue pairs. We have derived results that reflect both optimistic and pessimistic aspects of representative social choice.

Implications for AI Alignment Representative social choice can be used to model the alignment of AI systems to diverse human preferences. Generalization analysis of social choice mechanisms naturally apply to alignment mechanisms, while representative impossibility results highlight trade-offs between the alignment objectives of fairness and utility maximization. These insights can guide the development of more robust alignment strategies that manage these trade-offs explicitly.

Limitations and Future Directions We focused on the generalization properties of representational mechanisms without studying other important properties, such as incentive compatibility and computational tractability. Future research could explore them and their interactions with generalization.

References

Kenneth J Arrow. *Social choice and individual values*, volume 12. Yale university press, 2012.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12: 149–198, 2000.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer school on machine learning*, pages 169–207. Springer, 2003.

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

Jules Coleman and John Ferejohn. Democracy and social choice. *Ethics*, 97(1):6–25, 1986.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. Social choice for ai alignment: Dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*, 2024.

Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pages 12004–12019. PMLR, 2022.

Allan M Feldman and Roberto Serrano. *Welfare economics and social choice theory*. Springer Science & Business Media, 2006.

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.

Peter C Fishburn. *The theory of social choice*. Princeton University Press, 2015.

Luise Ge, Daniel Halpern, Evi Micha, Ariel D Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. Axioms for ai alignment from human feedback. *arXiv preprint arXiv:2405.14758*, 2024.

Allan Gibbard. Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society*, pages 587–601, 1973.

Martin Hellman and Josef Raviv. Probability of error, equivocation, and the chernoff bound. *IEEE Transactions on Information Theory*, 16(4):368–372, 1970.

Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective constitutional ai: Aligning a language model with public input. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1395–1417, 2024.

Eric Kauderer-Abrams. Quantifying translation-invariance in convolutional neural networks. *arXiv preprint arXiv:1801.01450*, 2017.

Oliver Klingefjord, Ryan Lowe, and Joe Edelman. What are human values, and how do we align ai to them? *arXiv preprint arXiv:2404.10636*, 2024.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models, like humans, show content effects on reasoning tasks. *PNAS nexus*, 3(7), 2024.

Dominique Lepelley, Patrick Pierron, and Fabrice Valognes. Scoring rules, condorcet efficiency and social homogeneity. *Theory and Decision*, 49(2):175–196, 2000.

Kehan Li, Runyi Yu, Zhennan Wang, Li Yuan, Guoli Song, and Jie Chen. Locality guidance for improving vision transformers on tiny datasets. In *European Conference on Computer Vision*, pages 110–127. Springer, 2022.

Christian List. The theory of judgment aggregation: an introductory review. *Synthese*, 187(1): 179–207, 2012.

Nicholas R Miller. Pluralism and social choice. *American Political Science Review*, 77(3):734–747, 1983.

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

Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. *Advances in neural information processing systems*, 21, 2008.

Noam Nisan and Amir Ronen. Algorithmic mechanism design. In *Proceedings of the thirty-first annual ACM Symposium on Theory of Computing*, pages 129–140, 1999.

David Parkes and Ariel Procaccia. Dynamic social choice with evolving preferences. In *Proceedings of the AAAI conference on artificial intelligence*, volume 27, pages 767–773, 2013.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Mahendra Prasad. Social choice and the value alignment problem. *Artificial intelligence safety and security*, pages 291–314, 2018.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

Mark Allen Satterthwaite. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 10(2): 187–217, 1975.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020.

Alan D Taylor. *Social choice and the mathematics of manipulation*. Cambridge University Press, 2005.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Han Wang, Erfan Miahi, Martha White, Marlos C Machado, Zaheer Abbas, Raksha Kumaraswamy, Vincent Liu, and Adam White. Investigating the properties of neural network representations in reinforcement learning. *Artificial Intelligence*, 330:104100, 2024.

H Peyton Young. Social choice scoring functions. *SIAM Journal on Applied Mathematics*, 28(4): 824–838, 1975.

Appendices

Table of Contents

Overall Structure of the Appendices

Related Work (Appendix A) We first give a brief overview of the related work in social choice theory, machine learning methods in social choice theory, and applications of social choice in AI alignment. They serve as a background for the study of representative social choice.

Binary Representative Social Choice (Appendix B) We start with a simpler case of representative social choice, where an infinitely large population hold preferences over a possibly infinite number of *binary* issues — issues that can be resolved by a simple Yes/No vote — and a collection of individual-issue pairs are randomly drawn as samples, from which a collective preference profile[3] is constructed to represent the entire population. This setting is meant for analyzing the problem of *generalization* in representative social choice, and we examine the majority vote mechanism under this binary representative setting as a case study.

General Representative Social Choice (Appendix C) We then extend our framework to the general case of representative social choice, where the issues are no longer binary and can have any finite number of outcomes. In this case, we introduce a more general class of mechanisms, the *scoring mechanisms*, which assign scores to candidate profiles based on individual-issue pairs, and output the candidate profile with the highest average score. We show that the generalization properties of scoring mechanisms can be analyzed using the theory of statistical learning. On the pessimistic side, however, we present Arrow-like impossibility theorems for representative social choice. To this end, we introduce new combinatorial tools, *privileged orderings* and the *privilege graph*, to analyze the structure of the candidate space and the interdependence between issues and between outcomes.

We consider our contribution in Appendix B, C.1, C.2, C.3 to be primarily conceptual and stage-setting, establishing the representative framework with techniques well-known in the fields of social choice theory and statistical learning theory. Appendix C.4 and C.5, when establishing the conceptually important strong impossibility theorem, additionally introduce new combinatorial tools of analysis, which we believe to be of independent interest.

[3] In this paper, we abuse the term *preference profile* to mean either a collection of preference ordering for different *individuals* in a population, or a collection of preference ordering for different *issues*.

A Related Work

Social Choice Theory Social choice theory studies the aggregation of individual preferences to form a collective decision. The field was founded by Arrow [2012], who proved the famous *Arrow's impossibility theorem*, stating that no social choice mechanism can satisfy a set of desirable properties, including *unrestricted domain*, *non-dictatorship*, *Pareto efficiency*, and *independence of irrelevant alternatives*. Since then, many extensions of Arrow's theorem have been proposed, including the *Gibbard-Satterthwaite theorem* [Gibbard, 1973, Satterthwaite, 1975] which introduces strategy-proofness to the analysis. Beyond impossibility results, social choice theory also studies concrete mechanisms for preference aggregation, such as *voting rules* [Taylor, 2005], *scoring rules* [Young, 1975], and *judgment aggregation* [List, 2012], while featuring intersections with other fields such as mechanism design [Nisan and Ronen, 1999] and computational social choice [Brandt et al., 2016]. in this paper, we extend the study of social choice theory to the representative setting, where the number of issues and individuals is too large to consider all preferences directly.

Machine Learning Methods in Social Choice Theory Various machine learning methods have been applied to social choice theory to address limitations of over-simplification in certain social choice models. For instance, *generative social choice* studies the problem of handling open-ended outcomes in social choice theory in theoretically sound ways [Fish et al., 2023], and *dynamic social choice* studies the problem of handling evolving preferences in social choice theory, using theoretical models from reinforcement learning [Parkes and Procaccia, 2013]. in this paper, we apply the theory of statistical learning to the study of representative social choice, and show that many of the deepest questions in representative social choice can be naturally formulated as statistical learning problems.

Applications of Social Choice in AI Alignment Social choice theory has been applied to the study of AI alignment, where the goal is to design AI systems that make decisions that are aligned with human values. Current approaches to AI alignment involves the aggregation of human preferences, and social choice-based algorithms [Köpf et al., 2024, Klingefjord et al., 2024], experimental studies [Huang et al., 2024], conceptual frameworks [Prasad, 2018, Mishra, 2023, Conitzer et al., 2024], and axiomatic frameworks [Ge et al., 2024] have been proposed to address the problem. in this paper, we extend the study of social choice theory to the representative setting, which can be applied to the AI alignment setting, where the human preferences are represented by a collection of individual-issue pairs sampled from human evaluators, and the AI system is trained to make decisions that are representative of the human population's preferences.

B Binary Representative Social Choice

In this section, we consider the case of representative social choice where the issues are binary. We will show that this setting can be naturally formulated as a statistical learning problem, and we will analyze the generalization properties of the majority vote mechanism under this setting.

A real-world example of binary representative social choice is the case of jury trials, where a randomly selected jury makes decisions on behalf of the entire population, and the outcome of the trial is a binary decision (guilt or innocence).

B.1 Problem Settings

Here we present the formal definitions of the binary representative social choice problem.

Issues We consider a discrete (but possibly infinite) set of binary issues $\mathcal{I}$. Each issue $i \in \mathcal{I}$ can be resolved by a simple Yes/No vote, and therefore each individual's preference profile can be represented as a mapping from $\mathcal{I}$ to $\mathrm{LO}(2)$, where $\mathrm{LO}(2)$ is the set of linear orders over a set of two outcomes (Yes/No).

Furthermore, we define a *saliency distribution* $\mathcal{D}_{\mathcal{I}}$ with full support over $\mathcal{I}$, which represents the importance of different issues, and decides the probability of each issue being sampled in the representative process. If there are a finite number of equally important issues, then $\mathcal{D}_{\mathcal{I}}$ is the uniform distribution over $\mathcal{I}$; if there are a few important issues and many unimportant ones, then $\mathcal{D}_{\mathcal{I}}$ is a distribution that assigns high probability to the important issues and low probability to the unimportant ones, including in cases with infinitely many issues. The distribution $\mathcal{D}_{\mathcal{I}}$ is assumed to be known to the mechanism.

Population We consider a possibly infinite population, represented by a distribution $\mathcal{D}_{\mathcal{P}} \in \Delta[\mathcal{P}]$, where $\Delta[\mathcal{P}]$ is the space of probability distributions over the support set $\mathcal{P} \subseteq \mathrm{LO}(2)^{\mathcal{I}}$.[4][5] For any preference profile $C \in \mathcal{P}$, denote with $\mathcal{D}_{\mathcal{P}}(C)$ the probability[6] that a randomly selected individual in the population has preference profile C over the two outcomes of all issues.

Often, we only need to consider the marginal distribution of $\mathcal{D}_{\mathcal{P}}$ — the mapping $\mathcal{M} : \mathcal{I} \to \Delta[\mathrm{LO}(2)]$, where $\Delta[\mathrm{LO}(2)]$ is the space of probability distributions over $\mathrm{LO}(2)$. For any issue $i \in \mathcal{I}$, $\mathcal{M}(i)$ is the distribution of preferences over the two outcomes of issue i in the population. For any preference ordering $o \in \mathrm{LO}(2)$, we denote with $\mathcal{M}(i)_o$ the probability that a randomly selected individual in the population has preference ordering o over the two outcomes of issue i.

Outcomes The result of a decision-making process is a preference profile $C : \mathcal{I} \to \mathrm{LO}(2)$, which represents the aggregated preference of the population generated by some mechanism.

The mechanism is said to be *representational* if the decision is made based on a finite collection of individual-issue pairs $\mathcal{S} = \{(o_1, i_1), (o_2, i_2), \ldots, (o_{|S|}, i_{|S|})\}$, where $i_k \in \mathcal{I}$ is an issue sampled from $\mathcal{D}_{\mathcal{I}}$, and $o_k \sim \mathcal{M}(i_k)$ is an individual's preference over the two outcomes of issue i_k, sampled from the population distribution.

However, not all preference profiles are allowed. As a key feature of representative social choice, the mechanism is only allowed to output preference profiles from a limited *candidate space* $\mathcal{C} \subseteq \mathrm{LO}(2)^{\mathcal{I}}$, which represents the space of possible preference profiles that can be generated by the mechanism (*e.g.*, presidential candidates in the election case, or language model policies in the AI alignment case). Without such a candidate space, multi-issue settings become degrenerate, as different issues become mutually independent and can be decided separately. Finally, a mechanism is a function $f : (\mathrm{LO}(2) \times \mathcal{I})^* \to \mathcal{C}$ that maps the sample collection $\mathcal{S}$ to a preference profile $C \in \mathcal{C}$.

[4] Here, we define $\mathcal{P}$ instead of directly using $\mathrm{LO}(2)^{\mathcal{I}}$, because when $|\mathcal{I}|$ is infinite, $\mathrm{LO}(2)^{\mathcal{I}}$ is uncountably infinite, and a distribution on it will be hard to specify. Instead, $\mathcal{P}$ can allow for, for instance, distributions over parameterizations.

[5] In this paper, we denote with A^B the space of mappings from B to A.

[6] Probability mass or probability density, depending whether $\mathcal{P}$ is discrete or continuous.

B.2 Binary Generalization Bound

When deciding on the aggregation result, the mechanism only has access to a finite collection of individual-issue pairs, and therefore the decision is made based on a finite sample of all population-issue pairs. This raises the question of *generalization* — how can we ensure that the decision made by the mechanism is representative of the population's preferences, when the mechanism is only optimized for a small number of individual opinions on individual issues?

To answer this question, we can leverage the theory of statistical learning, which studies the generalization properties of learning/optimization algorithms based on finite samples. However, the reliability of generalization depends on the complexity of the candidate space $\mathcal{C}$ — the more flexibile the candidate space, the more likely a profile can be picked by the mechanism that specifically fits the sample (*overfitting*) rather than the broader population. To characterize complexity, we can use the concept of *Vapnik-Chervonenkis (VC) dimension* [Vapnik, 1999], which measures the capacity of a hypothesis space to fit arbitrary finite samples.

Definition B.1 (Vapnik-Chervonenkis Dimension [Vapnik, 1999])**.** *Given any issue space $\mathcal{I}$, we consider candidate profiles mapping $\mathcal{I}$ to $\mathrm{LO}(2)$, and a candidate space $\mathcal{C} \subseteq \mathrm{LO}(2)^{\mathcal{I}}$. The VC dimension of $\mathcal{C}$ is the cardinality of the largest finite set of issues $I \subseteq \mathcal{I}$ such that for any binary function $o \in \mathrm{LO}(2)^I$, there exists a preference profile $c \in \mathcal{C}$ such that $c(i) = o(i)$ for all $i \in I$. If I can be arbitrarily large, then the VC dimension is infinite. Here we assume that the VC dimension is nonzero.*

With the VC dimension as a measure of complexity, we can now introduce the sample complexity theorem. For any preference profile C as the aggregation outcome, the theorem gives an upper bound on the difference between the *sample utility* $\frac{1}{|\mathcal{S}|}\sum_k \mathbf{1}_{C(i_k)=o_k}$ (the goodness of the aggregated profile, evaluated on the selected samples) and the *population utility* $\mathrm{E}_{i\sim\mathcal{D}_\mathcal{I}}[\mathcal{M}(i)_{C(i)}]$ (the unknown utility of the aggregated profile for the entire population), as a function of the sample size $|\mathcal{S}|$ and the VC dimension of the candidate space $\mathcal{C}$. Such a difference is called the *generalization error*.

Theorem 1 (Binary Generalization Bound)**.** *Let $\mathcal{C}$ be a candidate space with VC dimension $\mathrm{VC}(\mathcal{C})$, and let $\epsilon > 0$ be a desired generalization error. Then, for any $\delta > 0$, with probability at least $1-\delta$, the sample utility and population utility of* any *preference profile $C \in \mathcal{C}$ are ϵ-close, i.e.,*

$$\Pr\left[\left|\frac{1}{|\mathcal{S}|}\sum_{k=1}^{|\mathcal{S}|}\mathbf{1}_{C(i_k)=o_k} - \mathrm{E}_{i\sim\mathcal{D}_\mathcal{I}}[\mathcal{M}(i)_{C(i)}]\right| \le \epsilon, \quad \forall C \in \mathcal{C}\right] \ge 1-\delta \tag{7}$$

as long as we have the following, for some constant $c > 0$:

$$|\mathcal{S}| \ge \frac{c}{\epsilon^2}\mathrm{VC}(\mathcal{C})\left(\log \mathrm{VC}(\mathcal{C}) + \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right) \tag{8}$$

Proof. First consider the case where population has size 1, and the resulting population $\mathcal{M}$ always maps an issue to a one-point distribution (*i.e.*, a deterministic preference).

In this case, $\mathcal{M}(i)_{C(i)} \in \{0,1\}$ represents whether the aggregated profile C is correct on issue i. The population utility $\mathrm{E}_{i\sim\mathcal{D}_\mathcal{I}}[\mathcal{M}(i)_{C(i)}]$ can thus be formulated as the *population error* in statistical learning settings, and the result in Theorem 3.1 follows directly from the VC generalization error bound in statistical learning theory [Vapnik, 1999].

To generalize the result to the case where the population has arbitrary cardinality, we can construct the following reduction to the deterministic case. We make the following definitions:

$$\tilde{\mathcal{I}} \coloneqq \mathcal{I} \times \mathrm{LO}(2) \tag{9}$$

$$\tilde{\mathcal{D}}_{\tilde{\mathcal{I}}}(i,b) \coloneqq \mathcal{D}_\mathcal{I}(i)\cdot\mathcal{M}(i)_b \tag{10}$$

$$\tilde{\mathcal{M}}(i,b)_{b'} \coloneqq \mathbf{1}_{b=b'} \tag{11}$$

$$\tilde{\mathcal{C}} \coloneqq \{\tilde{C} : (i,b) \mapsto C(i) \mid C \in \mathcal{C}\} \tag{12}$$

$$\tilde{\mathcal{S}} \coloneqq \{(o_1,(i_1,o_1)),(o_2,(i_2,o_2)),\ldots,(o_{|\mathcal{S}|},(i_{|\mathcal{S}|},o_{|\mathcal{S}|}))\} \tag{13}$$

where $b, b' \in \mathrm{LO}(2)$.

In other words, we duplicate each issue i into two issues $(i,0)$ and $(i,1)$, and for each issue i, we construct a population $\tilde{\mathcal{M}}(i,b)$ that always maps the issue to a one-point distribution, where the point is b. We then construct a saliency distribution $\tilde{\mathcal{D}}_{\tilde{\mathcal{I}}}$ that accounts for the original population probabilities over preferences. It can be verified that

$$\mathrm{E}_{(i,b)\sim\tilde{\mathcal{D}}_{\tilde{\mathcal{I}}}}[\tilde{\mathcal{M}}(i,b)_{\tilde{C}(i,b)}] = \mathrm{E}_{i\sim\mathcal{D}_{\mathcal{I}}}[\mathcal{M}(i)_{C(i)}] \tag{14}$$

$$\frac{1}{|\tilde{\mathcal{S}}|}\sum_{k=1}^{|\tilde{\mathcal{S}}|}\mathbf{1}_{\tilde{C}((i_k,o_k))=o_k} = \frac{1}{|\mathcal{S}|}\sum_{k=1}^{|\mathcal{S}|}\mathbf{1}_{C(i_k)=o_k} \tag{15}$$

$$\mathrm{VC}(\tilde{\mathcal{C}}) = \mathrm{VC}(\mathcal{C}) \tag{16}$$

Since Theorem 3.1 holds for the deterministic case $(\tilde{\mathcal{I}}, \tilde{\mathcal{D}}_{\tilde{\mathcal{I}}}, \tilde{\mathcal{M}}, \tilde{\mathcal{C}})$, it also holds for the original case $(\mathcal{I}, \mathcal{D}_{\mathcal{I}}, \mathcal{M}, \mathcal{C})$. □

Intuitively, the sample complexity theorem states that the sample utility of any preference profile in the candidate space $\mathcal{C}$ approximates the population utility, as long as the sample size is sufficiently large. The sample size required for this guarantee depends on the VC dimension of the candidate space, the desired generalization error ϵ, and the desired confidence level δ.

Note that here we cannot directly utilize the tail inequalities [Hellman and Raviv, 1970] to estimate the generalization error, because when the profile C is picked to maximize sample utility, the sample utility ceases to be an unbiased estimator of the population utility.

Theorem 3.1 will be central in Section B.3, where we analyze the generalization properties of the majority vote mechanism under the binary representative setting.

B.3 Case Study: Majority Vote in the Binary Case

In this section, we consider the *majority vote* mechanism under the binary representative setting. It generalizes the well-known majority vote mechanism in social choice theory, where the population directly votes on a set of binary issues, to the representative setting, where the population is represented by a collection of individual-issue pairs.

Definition B.2 (Majority Vote Mechanism)**.** *The majority vote mechanism is a representational mechanism f_{maj} that outputs the preference profile C that maximizes the sample utility,* i.e.*,*

$$f_{\mathrm{maj}}(\mathcal{S}) = \arg\max_{C\in\mathcal{C}} \frac{1}{|\mathcal{S}|}\sum_{k=1}^{|\mathcal{S}|}\mathbf{1}_{C(i_k)=o_k}. \tag{17}$$

The majority vote mechanism can be viewed as a voting process where each individual-issue pair in the sample collection $\mathcal{S}$ casts a vote for the outcome that the individual prefers, based on the individual's preference over the two outcomes of the issue. The candidate profile that receives the most votes is then selected as the aggregated preference profile. When there is only one issue and candidate space $\mathcal{C} = \mathrm{LO}(2)$, the majority vote mechanism reduces to the standard majority vote mechanism in social choice theory.

From Theorem 3.1, we know that the majority vote mechanism has good generalization properties when the VC dimension of the candidate space $\mathcal{C}$ is small, resulting in the following corollary.

Corollary B.3 (Majority Vote Approximately Maximizes Population Utility)**.** *For any error requirement $\epsilon > 0$ and confidence requirement $\delta > 0$, the majority vote mechanism f_{maj} satisfies*

$$\Pr\left[U_{\mathcal{D}_{\mathcal{I}},\mathcal{M}}(f_{\mathrm{maj}}(\mathcal{S})) \geq -2\epsilon + \max_{C\in\mathcal{C}} U_{\mathcal{D}_{\mathcal{I}},\mathcal{M}}(C)\right] \geq 1-\delta \tag{18}$$

where the population utility

$$U_{\mathcal{D}_{\mathcal{I}},\mathcal{M}}(C) \coloneqq \mathrm{E}_{i\sim\mathcal{D}_{\mathcal{I}}}[\mathcal{M}(i)_{C(i)}], \tag{19}$$

as long as the sample size $|\mathcal{S}|$ satisfies

$$|\mathcal{S}| \geq \frac{c}{\epsilon^2}\mathrm{VC}(\mathcal{C})\left(\log \mathrm{VC}(\mathcal{C}) + \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right) \tag{20}$$

Proof. From Theorem 3.1, with probability $1-\delta$, we have

$$U_{\mathcal{D}_\mathcal{I},\mathcal{M}}(f_{\text{maj}}(\mathcal{S})) \geq -\epsilon + \frac{1}{|\mathcal{S}|}\sum_{k=1}^{|\mathcal{S}|} \mathbf{1}_{f_{\text{maj}}(\mathcal{S})(i_k)=o_k} \tag{21}$$

$$= \max_{C\in\mathcal{C}} \left(-\epsilon + \frac{1}{|\mathcal{S}|}\sum_{k=1}^{|\mathcal{S}|} \mathbf{1}_{f_{\text{maj}}C(i_k)=o_k} \right) \tag{22}$$

$$\geq -2\epsilon + \max_{C\in\mathcal{C}} U_{\mathcal{D}_\mathcal{I},\mathcal{M}}(C) \tag{23}$$

□

Corollary 3.2 shows that the majority vote mechanism approximately maximizes the population utility, as long as the sample size is sufficiently large and the VC dimension of the candidate space is small.

Furthermore, we can verify that the majority vote mechanism satisfies the following formal axioms, analogous to the classical *Pareto efficiency* and *non-dictatorship* axioms in social choice theory [Fishburn, 2015].

Axiom: Probabilistic Pareto Efficiency (PPE), Binary Case

Consider profiles $C, C' \in \mathcal{C}$ such that for the only issue $i \in \mathcal{I}$ that they disagree on, we have $1 \succ_C 0, 0 \succ_{C'} 1$ and the population is unanimous on $1 \succ 0$.[a] For any such C, C' and a sufficiently large sample size $|\mathcal{S}|$, with probability at least $1 - e^{\alpha|\mathcal{S}|}$,[b] $f_{\text{maj}}(\mathcal{S}) \neq C'$. Likewise when $0 \succ_C 1, 1 \succ_{C'} 0$ and the population is unanimous on $0 \succ 1$.

[a] In other words, $\mathcal{M}(i)_{1\succ 0} = 1$

[b] ... where $\alpha > 0$ is an arbitrary constant dependent only on C, C', and $|\mathcal{S}|$ denotes the number of samples *that fall into issue* i. From now on, we will abbreviate this expression to $1 - e^{\Omega(|\mathcal{S}|)}$.

Remark B.4. *Introduction of the candidate space $\mathcal{C}$ leads to interdependence between issues, which is not present in the classical social choice setting. As a result, we could not simply require that the mechanism output $1 \succ 0$ when the population is unanimous on $1 \succ 0$ as in the classical setting — what if the population is unanimous on $1 \succ 0$ for one issue, but unanimous on $0 \succ 1$ for another issue, and the two issues are strongly correlated (*e.g.*, every candidate profile agrees on the two issues)? Cases like these, while less extreme, are widespread in the real world. The way we state the PPE axiom avoids this problem by comparing between two candidate profiles C and C' that keep the same preference on all issues except one, a way to ensure "all else being equal".*

The probability bound $1 - e^{\Omega(|\mathcal{S}|)}$ is the convergence rate guaranteed by Hoeffding's inequality for independent samples. Intuitively speaking, this is the probability that you can correctly tell a majority from a minority in the population by looking at a large sample $\mathcal{S}$.

Axiom: Probabilistic Non-Dictatorship (PND), Binary Case

For any issue $i \in \mathcal{I}$, for any subpopulation $\mathcal{D'}_\mathcal{P}$ that occupies a probability mass $|\mathcal{D'}_\mathcal{P}| < 0.5$ in the whole population, at least one of the following is true w.r.t. issue i:

- When $\mathcal{D'}_\mathcal{P}$ is unanimous on $0 \succ 1$, there exists a preference specification $\mathcal{D}_\mathcal{P}$ of the whole population for which $f_{\text{maj}}(\mathcal{S})(i) = (1 \succ 0)$ with probability $1 - e^{\Omega(|\mathcal{S}|)}$.
- When $\mathcal{D'}_\mathcal{P}$ is unanimous on $1 \succ 0$, there exists a preference specification $\mathcal{D}_\mathcal{P}$ of the whole population for which $f_{\text{maj}}(\mathcal{S})(i) = (0 \succ 1)$ with probability $1 - e^{\Omega(|\mathcal{S}|)}$.

Remark B.5. *Here, the requirement that* at least one *(as opposed to* both*) condition is fulfilled is, again, a consequence of the candidate space $\mathcal{C}$. In the extreme case, if all candidate profiles agree that $0 \succ 1$ for issue i, then at most one of the two conditions can be met.*

The non-dictatorship axiom above is stronger than the classical version, in the sense that not only individuals, but also coalitions, cannot dictate the aggregation result.

In fact, our problem setting (Appendix B.1) treats individuals as interchangeable, thereby implicitly forcing *anonymity* of the mechanism — a stronger property than PND. As a result, representational mechanisms always satisfy PND, a fact that is formalized in Lemma C.5 for the more general, non-binary case.

Note that the classical axiom of independence of irrelevant alternatives (IIA) [Fishburn, 2015] is not applicable to the binary case, since irrelevant alternatives only exist when there are more than two outcomes for each issue.

C General Representative Social Choice

Having formulated and analyzed the binary case of representative social choice, we now extend our analysis to the general case where there can be an arbitrary number of outcomes for each issue. We will first present the formulation, show a generalization bound similar to that in the binary case, and present Arrow-like impossibility theorems.

C.1 Problem Settings

Here we present the full formal definition of the representative social choice problem.

Issues We consider a discrete (but possibly infinite) set of N-ary issues $\mathcal{I}$. Each individual's preference profile can be represented as a mapping from $\mathcal{I}$ to $\mathrm{LO}(N)$, where $\mathrm{LO}(N)$ is the set of linear orders over $[N] = \{1, 2, \cdots, N\}$. We then define the *saliency distribution* $\mathcal{D}_{\mathcal{I}}$ with full support over $\mathcal{I}$, representing the importance of different issues.

Population We consider a possibly infinite population, represented by a distribution $\mathcal{D}_{\mathcal{P}} \in \Delta[\mathcal{P}]$, where $\Delta[\mathcal{P}]$ is the space of probability distributions over the support set $\mathcal{P} \subseteq \mathrm{LO}(N)^{\mathcal{I}}$.

We then define the marginal distribution $\mathcal{M} : \mathcal{I} \to \Delta[\mathrm{LO}(N)]$. For any issue $i \in \mathcal{I}$, $\mathcal{M}(i)$ is the distribution of preferences over the N outcomes of issue i in the population. For any preference ordering $o \in \mathrm{LO}(N)$, we denote with $\mathcal{M}(i)_o$ the probability that a randomly selected individual in the population has preference ordering o over outcomes of issue i.

Outcomes The result of a decision-making process is a preference profile $C : \mathcal{I} \to \mathrm{LO}(N)$, which represents the aggregated preference of the population generated by some mechanism.

The mechanism is said to be *representational* if the decision is made based on a finite collection of individual-issue pairs $\mathcal{S} = \{(o_1, i_1), (o_2, i_2), \ldots, (o_{|S|}, i_{|S|})\}$, where $i_k \in \mathcal{I}$ is an issue sampled from $\mathcal{D}_{\mathcal{I}}$, and $o_k \sim \mathcal{M}(i_k)$ is an individual's preference over the two outcomes of issue i_k, sampled from the population distribution.

The social choice mechanism is only allowed to output preference profiles from a limited *candidate space* $\mathcal{C} \subseteq \mathrm{LO}(N)^{\mathcal{I}}$, which represents the space of possible preference profiles that can be generated by the mechanism. A mechanism is thus a function $f : (\mathrm{LO}(N) \times \mathcal{I})^* \to \mathcal{C}$ that maps the sample collection $\mathcal{S}$ to a preference profile $C \in \mathcal{C}$.

C.2 Scoring Rules and Generalization Errors

Generalizing the majority vote mechanism to the general case meets challenges, as the majority vote is over candidate profiles as opposed to outcomes, and when $N > 2$, the way each individual-issue pair votes is no longer well-defined. Instead, we can consider a more general class of mechanisms called *scoring mechanisms* [Lepelley et al., 2000], which for every individual-issue pair, assign a score to each candidate profile, and then output the candidate profile with the highest average score.

As an example, the scoring rule could be an arbitrary distance measure that measures the degree of alignment between the sampled individual's preference and the candidate profile's preference. The mechanism then outputs the candidate profile that maximizes the average sample alignment score.

Definition C.1. *A scoring mechanism is defined by a* scoring rule $s : \mathrm{LO}(N) \times \mathrm{LO}(N) \to \mathbb{R}$*, which assigns a score to each pair of preference orderings over the N outcomes. The scoring mechanism then outputs the candidate profile that maximizes the average score over all individual-issue pairs in the sample collection $\mathcal{S}$,* i.e.*,*

$$f_s(\mathcal{S}) = \arg\max_{C \in \mathcal{C}} \frac{1}{|\mathcal{S}|} \sum_{k=1}^{|\mathcal{S}|} s(o_k, C(i_k)). \tag{24}$$

For scoring mechanisms, we can similarly define the sample score $\frac{1}{|\mathcal{S}|} \sum_{k=1}^{|\mathcal{S}|} s(o_k, C(i_k))$ and the population score $U^s_{\mathcal{D}_{\mathcal{I}}, \mathcal{M}} := \mathrm{E}_{i \sim \mathcal{D}_{\mathcal{I}}, o \sim \mathcal{M}(i)}[s(o, C(i))]$, and have the following guarantees on generalization, analogous to Corollary 3.2.

Theorem 2 (Generalization Bound for Scoring Mechanisms)**.** *For any scoring rule s with value bounded by constants, the scoring mechanism f_s satisfies*

$$\Pr\left[U^s_{\mathcal{D}_\mathcal{I},\mathcal{M}}(f_s(\mathcal{S})) \geq -2\epsilon + \max_{C\in\mathcal{C}} U^s_{\mathcal{D}_\mathcal{I},\mathcal{M}}(C)\right] \geq 1-\delta \tag{25}$$

as long as the sample size $|\mathcal{S}|$ satisfies

$$|\mathcal{S}| \geq \frac{c}{\epsilon^2}\log\frac{1}{\delta} \tag{26}$$

and

$$\hat{\mathrm{R}}_{|\mathcal{S}|}(\bar{\mathcal{C}}) \leq \frac{\epsilon}{c} \tag{27}$$

where the function class $\bar{\mathcal{C}}$ is defined as

$$\bar{\mathcal{C}} \coloneqq \{(o,i) \mapsto s(o, C(i)) \mid C \in \mathcal{C}\} \tag{28}$$

and $\hat{\mathrm{R}}_{|\mathcal{S}|}(\bar{\mathcal{C}})$ is the empirical Rademacher complexity *of $\bar{\mathcal{C}}$ with respect to the sample collection $\mathcal{S}$ [Mohri and Rostamizadeh, 2008] — a generalization of the VC dimension to real-valued (as opposed to binary) functions.*

The result follows directly from the Rademacher complexity generalization error bound in statistical learning theory.

Corollary C.2 (Scoring Mechanisms Approximately Maximize Population Score)**.** *When $|\mathcal{I}|$ is finite, for any scoring rule s with value bounded by constants, the scoring mechanism f_s satisfies*

$$\Pr\left[U^s_{\mathcal{D}_\mathcal{I},\mathcal{M}}(f_s(\mathcal{S})) \geq -2\epsilon + \max_{C\in\mathcal{C}} U^s_{\mathcal{D}_\mathcal{I},\mathcal{M}}(C)\right] \geq 1-\delta \tag{29}$$

as long as we have the following, for some constant $c > 0$:

$$|\mathcal{S}| \geq \frac{c}{\epsilon^2}\left(|\mathcal{I}|N\log N + \log\frac{1}{\delta}\right) \tag{30}$$

Proof. Proof of the corollary is done by bounding the empirical Rademacher complexity of the function class $\bar{\mathcal{C}}$ using Massart's Lemma [Bousquet et al., 2003], which gives us

$$\hat{\mathrm{R}}_{|\mathcal{S}|}(\bar{\mathcal{C}}) \leq c\sqrt{\frac{\log|\bar{\mathcal{C}}|}{|\mathcal{S}|}}$$

for some constant $c > 0$.

Given that the cardinality of $\bar{\mathcal{C}}$ is $(N!)^{|\mathcal{I}|}$, the Corollary is not hard to verify by plugging the bound into Theorem 2. □

Rademacher complexity could be seen a generalization of the VC dimension to real-valued functions, and is a measure of the capacity of a function class to fit arbitrary finite samples. Intuitively, Corollary C.2 states that the sample score of any candidate profile in the candidate space $\mathcal{C}$ approximates the population score, as long as the sample size exceeds the ability of the candidate space to fit arbitrary finite samples — *i.e.*, when the mechanism is forced to *generalize*.

It's worth noting that Corollary C.2 is agnostic towards the correlation structure among issues, and as a result, the sample complexity grows approximately linearly with the number of issues $|\mathcal{I}|$ and the number of outcomes N. When given a issue correlation structure, the sample complexity can potentially be reduced using Theorem 2.

C.3 Weak Representative Impossibility

In this section, we present certain axioms that ideal social choice mechanisms should satisfy, and then present Arrow-like impossibility theorems that show that no mechanism can satisfy all these axioms simultaneously. A weaker version of the impossibility theorem — which we present in this section — is a simple generalization of the classical Arrow's impossibility theorem, while a stronger version shall be derived in Section C.5.

We first introduce the necessary notations, and then restate the axioms from the binary case, adapting them to our general setting.

Definition C.3 (Operation of a Permutation on an Ordering)**.** *For a set S, a linear order $o \in \mathrm{LO}(S)$, and a permutation $\sigma \in \mathfrak{S}_S$, we define the operation $o \odot \sigma$ as the ordering obtained by applying the permutation σ to the elements of o. Specifically, for any $s_1, s_2 \in S$, we have $s_1 \succ_{o\odot\sigma} s_2$ if and only if $\sigma^{-1}(s_1) \succ_o \sigma^{-1}(s_2)$.*

For a subset $T \subseteq S$ and $\sigma \in \mathfrak{S}_T$, we similarly define $o \odot \sigma := o \odot \sigma|_S$, where $\sigma|_S$ is the extension of σ to the whole set S, mapping elements outside T to themselves.

Finally, for a full profile $C \in \mathrm{LO}(N)^{\mathcal{I}}$, we define $C \odot_i \sigma$ as the profile obtained by applying the permutation σ to $C(i)$, while keeping C unchanged for all other issues.

Remark C.4. *Operation of a permutation on an ordering aims to capture "local changes" to a preference profile, where the permutation only affects a subset of the outcomes. They will be used in the definition of the axioms below.*

One common case of local change is swapping two outcomes, for which we denote with $\sigma_{(c_1,c_2)}$ the 2-element permutation (transposition) that swaps c_1 and c_2.

Axiom: Probabilistic Pareto Efficiency (PPE)

Consider profiles $C, C' \in \mathcal{C}$ such that for the only issue $i \in \mathcal{I}$ that they disagree on, there exists $c, c' \in [N]$ such that $c \succ_C c', C' = C \odot_i \sigma_{(c,c')}$[a] and the population is unanimous on $c \succ c'$.[b] For any such C, C', with probability $1 - e^{\Omega(|\mathcal{S}|)}$, $f(\mathcal{S}) \neq C'$.

[a] The second condition means that the order between c, c' is the only disagreement they have on issue i.

[b] In other words, $\mathcal{M}(i)_o = 0$ for all $o \in \mathrm{LO}(N)$ that prefers c' over c.

Axiom: Probabilistic Non-Dictatorship (PND)

For any issue $i \in \mathcal{I}$ and $c, c' \in [N]$, for any subpopulation $\mathcal{D'}_{\mathcal{P}}$ that occupies a probability mass $|\mathcal{D'}_{\mathcal{P}}| < 0.5$ in the whole population, at least one of the following is true w.r.t. issue i:

- When $\mathcal{D'}_{\mathcal{P}}$ is unanimous on $c \succ c'$, there exists a preference specification $\mathcal{D}_{\mathcal{P}}$ of the whole population for which $c' \succ_{f(\mathcal{S})} c$ with probability $1 - e^{\Omega(|\mathcal{S}|)}$.
- When $\mathcal{D'}_{\mathcal{P}}$ is unanimous on $c' \succ c$, there exists a preference specification $\mathcal{D}_{\mathcal{P}}$ of the whole population for which $c \succ_{f(\mathcal{S})} c'$ with probability $1 - e^{\Omega(|\mathcal{S}|)}$.

In fact, *all* representational mechanisms satisfy PND, since our problem setting (Appendix C.1) treats individuals as interchangeable and homogeneous, therefore implicitly forcing *anonymity* of the mechanism, which inturn implies PND. We formalize this fact in the following lemma, which will turn out useful in the proof of the later, strong version of the impossibility theorem.

Lemma C.5 (Probabilistic Non-Dictatorship for All Representational Mechanisms)**.** *For any $(\mathcal{I}, \mathcal{D}_{\mathcal{I}}, \mathcal{C})$ and any representational mechanism f, PND is satisfied.*

Proof. Assume otherwise, that under some mechanism f, a subpopulation $\mathcal{D'}_{\mathcal{P}}$ occupying a probability mass strictly less than 0.5 can dictate the aggregation result in both directions, by being unanimous on either $c' \succ c$ or $c \succ c'$.

Take another subpopulation $\mathcal{D''}_{\mathcal{P}}$ that's disjoint with $\mathcal{D'}_{\mathcal{P}}$ and has the same probability mass. Let $\mathcal{D'}_{\mathcal{P}}$ be unanimous on $c' \succ c$, and $\mathcal{D''}_{\mathcal{P}}$ be unanimous on $c \succ c'$. Since populations of the same probability mass are undisguishable to the mechanism, $\mathcal{D''}_{\mathcal{P}}$ must also dictate the aggregation result in both directions, leading to a contradiction. □

As a result, we shall remove PND from the explicit statements of the impossibility theorems, but it should be understood that PND is still implicitly satisfied.

Finally, we define a weak version of the independence of irrelevant alternatives (IIA) axiom, as well as a new axiom specific to the representative setting.

Axiom: Weak Probabilistic Independence of Irrelevant Alternatives (W-PIIA)

When $\mathcal{C} = \mathrm{LO}(N)^{\mathcal{I}}$, for two populations $\mathcal{D}_{\mathcal{P}}, \mathcal{D}'_{\mathcal{P}}$ that differ only in the preference over a single issue $i \in \mathcal{I}$ satisfying $\mathcal{M}(i) \mid_{\mathrm{LO}(\{c,c'\})} = \mathcal{M}'(i) \mid_{\mathrm{LO}(\{c,c'\})}$ (where c, c' are any two elements of $[N]$),[a] with probability $1 - e^{\Omega(|\mathcal{S}|)}$, we have $f(\mathcal{S}) \mid_{\mathrm{LO}(\{c,c'\})} = f(\mathcal{S}') \mid_{\mathrm{LO}(\{c,c'\})}$.

[a] The condition here means that the population distribution over the relative preference between c, c' are the same in the two populations.

Remark C.6. *$\mathcal{C} = \mathrm{LO}(N)^{\mathcal{I}}$ implies independence both between issues and between outcomes of the same issue. As a result, we won't need to worry about,* e.g.*, the population's insistance on $c_1 \succ c_2$ leads to the side effect of $c_3 \succ c_4$ due to all profiles in $\mathcal{C}$ ranking c_1 and c_3 close to each other and c_2 and c_4 close to each other. Such cross-outcome or cross-issue dependencies may contribute positively or negatively to the fulfillment of axioms, making $\mathcal{C} = \mathrm{LO}(N)^{\mathcal{I}}$ the easiest candidate space to analyze. We will aim to overcome these difficulties in Section C.4 and C.5.*

Axiom: Weak Probabilistic Convergence (W-PC)

When $\mathcal{C} = \mathrm{LO}(N)^{\mathcal{I}}$, for a population $\mathcal{D}_{\mathcal{P}}$ that's non-uniform on a pair of outcomes c, c' of issue $i \in \mathcal{I}$,[a] there exists a partial preference ordering $o \in \mathrm{LO}(\{c, c'\})$ such that with probability $1 - e^{\Omega(|\mathcal{S}|)}$, $f(\mathcal{S})(i) \mid_{\mathrm{LO}(\{c,c'\})} = o$.

[a] *i.e.*, $\mathcal{M}(i) \mid_{\mathrm{LO}(\{c,c'\})} (c \succ c') \neq 0.5$. We are only concerned with the marginal distribution.

Remark C.7. *Intuitively speaking, W-PC requires that the mechanism not be torn between two outcomes when the population is not torn between them. This is to rule out "indecisive" mechanisms that are unable to make a decision with high probability (or requires too many samples to make that decision) when the population is clear on the preference. This need arises only in the representative setting, where the mechanism is no longer deterministic.*

It's worth noting that $1 - e^{\Omega(|\mathcal{S}|)}$ is the convergence rate guaranteed by Hoeffding's inequality for independent samples, which, intuitively speaking, asks that the mechanism not be qualitatively slower in convergence than the majority vote mechanism. And again, $\mathcal{C} = \mathrm{LO}(N)^{\mathcal{I}}$ removes the complexity of cross-outcome dependencies, which simplifies the statement of the axiom.

We now present the weak representative impossibility theorem.

Theorem 3 (Weak Representative Impossibility)**.** *When $N \geq 3$, for any $(\mathcal{I}, \mathcal{D}_{\mathcal{I}}, \mathcal{C} = \mathrm{LO}(N)^{\mathcal{I}})$, no representational mechanism simultaneously satisfies PPE, W-PIIA, and W-PC for all $\mathcal{D}_{\mathcal{P}}$.*

Proof. The proof proceeds by a simple reduction to Arrow's impossibility theorem. Given the number of outcomes $N \geq 3$, any instance of the original voting problem in Arrow's theorem with N outcomes and an odd number n of voters can be reduced to the weak representative setting.

Specifically, let n be the number of voters in the original problem. One could simulate this by dividing the population into n disjoint subpopulations, each with a probability mass of $1/n$. Each subpopulation is then unanimous on their respective preference profile. Since n is odd, there can never be a pair of outcomes on which there is a tie in the population's preference. By W-PC, at the limit $|\mathcal{S}| \to +\infty$, the mechanism must converge upon a deterministic preference profile with probability 1, allowing us to treat the problem as a deterministic one, as in Arrow's theorem.

Meanwhile, the problem that multiple issues exist can be resolved by the independence between issues due to the candidate space $\mathcal{C} = \mathrm{LO}(N)^{\mathcal{I}}$, which allows us to examine each issue in isolation. The reduction is thus complete.

Since Arrow's theorem shows that for any problem instance with $N \geq 3$, no deterministic social choice mechanism can satisfy Pareto efficiency, independence of irrelevant alternatives, and non-dictatorship simultaneously, we are able to reduce at least one such hard instance to every instance $(\mathcal{I}, \mathcal{D}_{\mathcal{I}}, \mathcal{C} = \mathrm{LO}(N)^{\mathcal{I}})$ of the weak representative setting, and the theorem follows. □

The theorem is "weak" in the sense that it focuses on the $\mathcal{C} = \mathrm{LO}(N)^{\mathcal{I}}$ case, where the candidate space is the simplest, and no interdependence between issues or outcomes exist. In the next section, we will present a stronger version of the impossibility theorem that overcome this limitation.

C.4 Privileged Orderings and Privilege Graph

Before we present the strong representative impossibility theorem, we need tools to represent and analyze the structure of the candidate space $\mathcal{C}$. To this end, we introduce the concept of *privileged orderings*, which are partial orderings that are preferred over all other alternative orderings in the candidate space.

Definition C.8 (Privileged Ordering)**.** *For an issue $i \in \mathcal{I}$ and a subset of outcomes $T = \{c_1, c_2, \cdots, c_k\} \subseteq [N]$ $(k \geq 2)$, we call $o \in \mathrm{LO}(T) : c_1 \succ c_2 \succ \cdots \succ c_k$ a* privileged ordering *if for any extension $C \in \mathrm{LO}(N)^{\mathcal{I}}$ of o[7] and permutation $\sigma \in \mathfrak{S}_T$ of T, we have $C \odot_i \sigma \in \mathcal{C} \implies C \in \mathcal{C}$.*

Remark C.9. *Intuitively, a privileged ordering o is a partial ordering that is preferred in the candidate space $\mathcal{C}$ over all other alternative partial orderings. The definition requires that any full preference profile in $\mathcal{C}$ that disagrees with o must have a counterpart in $\mathcal{C}$ that agrees with o while keeping the rest of the profile unchanged. These privileged orderings are in fact surprisingly common in practice, as will be showcased in Example C.11 and C.12.*

Definition C.10 (Privilege Graph)**.** *For an issue $i \in \mathcal{I}$, we define the* privilege graph *G_i as a directed graph with vertices as the outcomes in $[N]$, and an edge from u to v iff there exists a privileged ordering $u \succ v$. We call i* cyclically privileged *if its privilege graph contains a simple directed cycle of length at least 3.*[8]

Example C.11 (Privileged Orderings in the Real World)**.** *Consider the following examples of privileged orderings and privilege graphs. We use the legal setting for the examples for simplicity, but it should be understood that privileged orderings are also widespread in other real-world settings.*

Specifically, consider three persons $\mathcal{I} = \{A, B, C\}$ on trial before the jury, after each being accused of a crime. A is charged of burglary ($5k, first trial), B also of burglary ($50k, second trial), and C of fraud. The possible outcomes for each defendant ($N = 3$) include aqcuittal ($\mathbf{a}$), community service ($\mathbf{c}$), and imprisonment ($\mathbf{i}$). The jury ranks the outcomes for each defendant in order of recommendation. In this hypothetical case, the following factors may lead to privileged orderings:

- ***Monotonicity and fairness constraints.*** *The jury may consider it unfair to punish A harder than B given the difference in the amount of burglary. As a result, locally changing the recommended outcome of A from $\mathbf{c}$ to $\mathbf{i}$ may violate the monotonicity constraint (namely when the outcome of B is $\mathbf{c}$), but changing from $\mathbf{i}$ to $\mathbf{c}$ will not. This line of reasoning results in privileged orderings $\mathbf{c} \succ \mathbf{i}$ and (analogously) $\mathbf{a} \succ \mathbf{c}$ for A, and $\mathbf{i} \succ \mathbf{c}$ and $\mathbf{c} \succ \mathbf{a}$ for B.*
- ***Local independence.*** *While the crimes of A and B are similar and therefore correlated in the jury's judgment, the crime of C is unrelated to the other two. As a result, the candidate space $\mathcal{C}$ may be the Cartesian product of the candidate space $\mathcal{C}_{(A,B)}$ for (A, B) and $\mathcal{C}_C$ for C. When $\mathcal{C}_C = \mathrm{LO}(\{\mathbf{a}, \mathbf{c}, \mathbf{i}\})$ and $\mathcal{C} = \mathcal{C}_{(A,B)} \times \mathrm{LO}(\{\mathbf{a}, \mathbf{c}, \mathbf{i}\})$, it can be verified that C's privilege graph G_C is a complete digraph, and every ordering is privileged for C.*
- ***Default outcomes.*** *Assume that B's sentence in the first trial was $\mathbf{c}$, and the jury is inclined to keep the sentence unchanged since no new substantial evidence is presented. This may lead to the privileged orderings $\mathbf{c} \succ \mathbf{i}$ and $\mathbf{c} \succ \mathbf{a}$ for B, since $\mathbf{c}$, being the default outcome, is a plausible substitute for any other outcome.*

*These considerations are generalizable. Locally independent decisions, and decisions with a default outcome, are common in the real world; and many decisions (*e.g.*, those involving numerical balancing of costs and benefits), according to common sense, should be monotonic as well. There are other reasons for privileged orderings too, and the three examples above are only for illustration.*

Example C.12 (Privileged Orderings in AI and AI Alignment)**.** *Outside of the human case, privileged orderings are also natural consequences of* inductive biases *in machine learning models [Baxter, 2000], including deep neural networks. Examples of this include:*

[7] *i.e.*, C must agree with o on issue i that $c_1 \succ c_2 \succ \cdots \succ c_k$, but can differ on the remaining outcomes of i, as well as on other issues.

[8] Note that for our purposes, we don't consider graphs containing only 2-cycles as cyclic.

- ***Language models***. *Language models are known to display a wide range of human behavioral tendencies [Perez et al., 2022, Santurkar et al., 2023, Lampinen et al., 2024], including ones discussed in Example C.11. Tendencies including consistent treatment of similar decisions, default outcomes, monotonicity, and many more, can all lead to privileged orderings. In the finetuning-based process of alignment, these pretrained tendencies serve as a form of inductive bias that sets limits on the possible outcomes of alignment training.*
- ***Reinforcement learning agents***. *In reinforcement learning, the reward function is a key component that guides the learning process. The reward function has been demonstrated to induce agents to learn deeply ingrained hiearchies or equivalences between outcomes [Di Langosco et al., 2022, Wang et al., 2024]. Similar to the language model case, these prior tendencies set limits on the possible outcomes of later learning, and can lead to privileged orderings.*
- ***Vision models***. *Vision models have been shown to exhibit spatial locality [Li et al., 2022], translation invariance [Kauderer-Abrams, 2017], simplicity bias [Shah et al., 2020], and other inductive biases. These biases, which are often necessary for the models to learn effectively, operate by setting limits on the possible outcomes of learning, leading to privileged orderings.*

While these tendencies in AI systems are often probabilistic and less clear-cut than in the human case, modeling them with privileged orderings and privilege graphs can still be a useful abstraction for understanding and approximating their behaviors.

Next, we characterize some important properties of privileged orderings and privilege graphs.

Lemma C.13 (Transitivity of Privileged Graph)**.** *For an issue $i \in \mathcal{I}$ and $u, v, w \in [N]$, if $u \succ v$ and $v \succ w$ are both privileged, then $u \succ w$ is privileged.*

Proof. For any extension $C \in \mathrm{LO}(N)^{\mathcal{I}}$ of $u \succ w$, we need to show that $C' \coloneqq C \odot_i \sigma_{(u,w)} \in \mathcal{C} \implies C \in \mathcal{C}$. Assuming $C' \in \mathcal{C}$, we consider ordering between u, v, w in $C(i)$.

When $u \succ_C w \succ_C v$, we have $v \succ_{C'} w \succ_{C'} u$. Since $u \succ v$ is privileged, $C'' \coloneqq C' \odot_i \sigma_{(u,v)} \in \mathcal{C}$, and we have $u \succ_{C''} w \succ_{C''} v$. Since $v \succ w$ is privileged, $C''' \coloneqq C'' \odot_i \sigma_{(v,w)} \in \mathcal{C}$. It can be verified that $u \succ_{C'''} v \succ_{C'''} w$ and $C''' = C$. The case $v \succ_C u \succ_C w$ is analogous.

When $u \succ_C v \succ_C w$, we have $w \succ_{C'} v \succ_{C'} u$. Since $v \succ w$ is privileged, $C'' \coloneqq C' \odot_i \sigma_{(v,w)} \in \mathcal{C}$. Now $v \succ_{C''} w \succ_{C''} u$, reducing to the first case. □

Lemma C.14 (Closure of Privileged Orderings Under Concatenation)**.** *Let $c_1, c_2, \cdots, c_{k+m-1} \in [N]$ $(k, m \geq 2)$ be distinct outcomes in an issue i. If $o : c_1 \succ c_2 \succ \cdots \succ c_k$ and $o' : c_k \succ c_{k+1} \succ c_{k+2} \succ \cdots \succ c_{k+m-1}$ are both privileged orderings, then $o \oplus o' : c_1 \succ c_2 \succ \cdots \succ c_{k+m-1}$ is also a privileged ordering.*

Proof. For any extension $C \in \mathrm{LO}(N)^{\mathcal{I}}$ of $o \oplus o'$ and permutation $\sigma \in \mathfrak{S}_{\{c_1, c_2, \cdots, c_{k+m-1}\}}$, we need to show that $C' \coloneqq C \odot_i \sigma \in \mathcal{C} \implies C \in \mathcal{C}$.

Similar to the proof of Lemma C.13, there are two transformations that we are allowed to apply to C' to obtain C: first, we can apply some permutation $\sigma_o \in \mathfrak{S}_{\{c_1, c_2, \cdots, c_k\}}$ on C to sort $\{c_1, c_2, \cdots, c_k\}$ in this order if it isn't already sorted; second, we can apply some permutation $\sigma_{o'} \in \mathfrak{S}_{\{c_k, c_{k+1}, \cdots, c_{k+m-1}\}}$ on C to sort $\{c_k, c_{k+1}, \cdots, c_{k+m-1}\}$ in this order if it isn't already. Since o and o' are privileged, these two transformations preserve membership in $\mathcal{C}$.

We repeated apply these two transformations in arbitrary order, until neither of them can be applied — *i.e.*, until both $\{c_1, c_2, \cdots, c_k\}$ and $\{c_k, c_{k+1}, \cdots, c_{k+m-1}\}$ are sorted in the correct order. Since $(c_1 \succ c_2 \succ \cdots \succ c_k) \wedge (c_k \succ c_{k+1} \succ c_{k+2} \succ \cdots \succ c_{k+m-1}) \implies c_1 \succ c_2 \succ \cdots \succ c_{k+m-1}$, when such a process terminates, we have $C \in \mathcal{C}$.

We still need to show that the process indeed terminates. We define a *potential function* Φ such that for any $E \in \mathrm{LO}(N)^{\mathcal{I}}$, $\Phi(E) = \mathrm{Inv}_{c_1 \succ \cdots \succ c_{k+m-1}}(E(i)) = \sum_{1 \leq x < y \leq k+m-1} \mathbf{1}_{c_y \succ_{E(i)} c_x}$, the number of inversions in $E(i)$ with respect to the ordering $o \oplus o' : c_1 \succ \cdots \succ c_{k+m-1}$.

We show that the transformation $\odot_i \sigma_o$ strictly decreases $\mathrm{Inv}_{o \oplus o'}$. We first decompose $\odot_i \sigma_o$ into a series of transpositions $\odot_i \sigma_{(c_{j_1}, c_{j_2})}$ such that $j_1 < j_2$ and $c_{j_2} \succ c_{j_1}$ in the pre-transposition ordering. It can be verified that this transposition removes the inversion (c_{j_1}, c_{j_2}) without introducing new

inversions. Therefore the transformation $\odot_i \sigma_o$ strictly decreases Φ, and likewise for $\odot_i \sigma_{o'}$. Since the initial $\Phi(C')$ is finite and Φ is always non-negative, the process must terminate. □

Corollary C.15. *For any issue $i \in \mathcal{I}$ and $o : c_1 \succ c_2 \succ \cdots \succ c_k$ $(k \geq 2)$ containing distinct outcomes in i, o is a privileged ordering if the path $c_1 \to c_2 \to \cdots \to c_k$ exists in G_i.*

The privilege graph G_i captures all binary privileged orderings in issue i, and will be the primary way in which we represent the structure of the candidate space $\mathcal{C}$.

It's worth noting that properties stronger than Lemma C.13 and C.14 often do not hold, *e.g.*, a subsequence of a privileged ordering is not necessarily privileged. Also, the condition in Corollary C.15 is a sufficient but not necessary condition for a privileged ordering.

Finally, we review a few graph theoretical concepts that will be useful in the proof of the strong representative impossibility theorem.

Definition C.16 (Strongly Connected Component)**.** *For a directed graph G, a* strongly connected component *(SCC) is a maximal subgraph of G (including those of size 1) in which there is a directed path between every ordered pair of vertices. The SCCs of G form a partition of the vertex set of G.*

Definition C.17 (Condensation)**.** *The* condensation $\mathrm{cond}(G)$ *of a directed graph G is a directed acyclic graph (DAG) where each vertex represents an SCC of G, and there is an edge from SCC A to SCC B iff there is an edge from a vertex in A to a vertex in B in G.*

Remark C.18. *Intuitively, SCC is the unit of bidirectional connectivity in a directed graph, and condensation is a way to simplify a directed graph into a DAG by contracting SCCs into single vertices. There is a directed path from u to v in G iff there is a directed path from the SCC containing u to the SCC containing v in $\mathrm{cond}(G)$. It's worth noting that, given the transitivity of the privilege graph, any SCC in the privilege graph is a complete digraph.*

Both SCC and condensation can be computed in linear time in the number of vertices and edges of the graph, using algorithms such as Tarjan's algorithm and Kosaraju's algorithm.

With these preparations, we are now able to state and prove the strong representative impossibility.

C.5 Strong Representative Impossibility

We first present the version of independence of irrelevant alternatives that will be used in the strong representative impossibility theorem, as well as the strong version of the probabilistic convergence axiom. They imply their weak counterparts while being applicable to arbitrary candidate spaces $\mathcal{C}$ and thus more general.

Axiom: Strong Probabilistic Independence of Irrelevant Alternatives (S-PIIA)

For arbitrary $\mathcal{C}$ and $c, c' \in [N]$ such that $c \succ c'$ and $c' \succ c$ are both privileged orderings in some issue $i \in \mathcal{I}$, for two populations $\mathcal{D}_\mathcal{P}, \mathcal{D}'_\mathcal{P}$ satisfying $\mathcal{M}(i) \mid_{\mathrm{LO}(\{c,c'\})} = \mathcal{M}'(i) \mid_{\mathrm{LO}(\{c,c'\})}$,[a] with probability $1 - e^{\Omega(|\mathcal{S}|)}$, we have $f(\mathcal{S}) \mid_{\mathrm{LO}(\{c,c'\})} = f(\mathcal{S}') \mid_{\mathrm{LO}(\{c,c'\})}$.

[a] The condition here means that the population distribution over the relative preference between c, c' are the same in the two populations.

Axiom: Strong Probabilistic Convergence (S-PC)

For arbitrary $\mathcal{C}$ and $c, c' \in [N]$ such that $c \succ c'$ and $c' \succ c$ are both privileged orderings in some issue $i \in \mathcal{I}$, for a population $\mathcal{D}_\mathcal{P}$ that's non-uniform on $\{c, c'\}$,[a] there exists a partial preference ordering $o \in \mathrm{LO}(\{c, c'\})$ such that with probability $1 - e^{\Omega(|\mathcal{S}|)}$, $f(\mathcal{S})(i) \mid_{\mathrm{LO}(\{c,c'\})} = o$.

[a] *i.e.*, $\mathcal{M}(i) \mid_{\mathrm{LO}(\{c,c'\})} (c \succ c') \neq 0.5$. We are only concerned with the marginal distribution.

Remark C.19. *S-PIIA and S-PC are generalizations of W-PIIA and W-PC, respectively, to arbitrary candidate spaces $\mathcal{C}$. This is achieved by limiting comparisons to pairs that are privileged orderings in both directions, ensuring that they are locally independent from other outcomes and issues.*

Before we present the strong representative impossibility theorem, we need to establish a few corollaries of the axioms we have defined.

Definition C.20 (Decisiveness)**.** *Given any representational mechanism f, for a subpopulation $\mathcal{D'}_{\mathcal{P}}$ with probability mass $0 < |\mathcal{D'}_{\mathcal{P}}| \leq 1$ in the larger population $\mathcal{D}_{\mathcal{P}}$, we say that it is* decisive over *$c \succ c'$ in issue i if when $\mathcal{D'}_{\mathcal{P}}$ is unanimous on $c \succ c'$, whatever the preference specification $\mathcal{D}_{\mathcal{P}}$ of the whole population is, $c \succ_{f(\mathcal{S})} c'$ with probability $1 - e^{\Omega(|\mathcal{S}|)}$. We say that it is* weakly decisive *if when $\mathcal{D'}_{\mathcal{P}}$ is unanimous on $c \succ c'$ and the rest of the population is unanimous on $c' \succ c$, $c \succ_{f(\mathcal{S})} c'$ with probability $1 - e^{\Omega(|\mathcal{S}|)}$.*

Remark C.21. *Given the guaranteed anonymity of our problem setting, all subpopulations of the same probability mass are entirely interchangeable, and thus the decisiveness of a subpopulation over a specific preference only depends on the subpopulation's mass.*

Lemma C.22 (Field Expansion Lemma)**.** *Given any representational mechanism f satisfying PPE and S-PIIA, consider outcomes u, v, w of issue $i \in \mathcal{I}$ such that all 6 pairwise orderings among them are privileged, and a subpopulation $\mathcal{D'}_{\mathcal{P}}$ that is weakly decisive over $u \succ v$. We then have that $\mathcal{D'}_{\mathcal{P}}$ is decisive over $u \succ w$.*

Note that by analogy, it will also be decisive over $w \succ v$. With repeated application and transitivity, it follows that $\mathcal{D'}_{\mathcal{P}}$ is decisive over all 6 pairwise orderings among u, v, w.

Proof. Assume that $\mathcal{D'}_{\mathcal{P}}$ is unanimous on $u \succ w$. By S-PIIA, we can assume that $\mathcal{D'}_{\mathcal{P}}$ is further unanimous on $u \succ v \succ w$, while the rest of the population is unanimous on $v \succ u$ and $v \succ w$. This does not affect $f(\mathcal{S})\mid_{\mathrm{LO}(\{u,w\})}$.

By weak decisiveness, we have $u \succ_{f(\mathcal{S})} v$ with probability $1 - e^{\Omega(|\mathcal{S}|)}$. Due to the privilegedness of $v \succ w$, PPE can be applied to every extension of the opposite ordering $w \succ v$ (where the extension serve as C'), guaranteeing that $v \succ_{f(\mathcal{S})} w$ with probability $1 - e^{\Omega(|\mathcal{S}|)}$. Taken together, we have $u \succ_{f(\mathcal{S})} w$ with probability $1 - e^{\Omega(|\mathcal{S}|)}$, and thus $\mathcal{D'}_{\mathcal{P}}$ is decisive over $u \succ w$. The rest follows. □

This finally leads us to the strong representative impossibility theorem.

Theorem 4 (Strong Representative Impossibility)**.** *For any $(\mathcal{I}, \mathcal{D}_{\mathcal{I}}, \mathcal{C})$, when there is at least one cyclically privileged issue,[9] there exists no representational mechanism that simultaneously satisfies PPE, S-PIIA, and S-PC for all $\mathcal{D}_{\mathcal{P}}$.*

Meanwhile, given any mapping ϕ from each issue i to a privilege graph $\phi(i)$ without cyclic privileges, there exist a candidate space $\mathcal{C}$ whose privilege graph G_i (for each i) is $\phi(i)$ or its supergraph, and a representational mechanism f over $(\mathcal{I}, \mathcal{D}_{\mathcal{I}}, \mathcal{C})$ satisfying PPE, S-PIIA, and S-PC for all $\mathcal{D}_{\mathcal{P}}$.

Proof. We first show that issues with cyclic privilege graphs imply the impossibility of a mechanism satisfying the axioms. We assume the existence of such an f and show that it leads to a contradiction.

For any cyclic privileged issue i, by Lemma C.13, we know that any directed simple cycle in G_i induces a complete digraph on the outcomes in the cycle. We arbitrarily pick outcomes u, v, w from the cycle. Arbitrarily partition the population into equal portions $\mathcal{D}^{\mathrm{a}}{}_{\mathcal{P}}$ and $\mathcal{D}^{\mathrm{b}}{}_{\mathcal{P}}$ of probablity mass $\frac{2}{3}$ and $\frac{1}{3}$ respectively. Let $\mathcal{D}^{\mathrm{a}}{}_{\mathcal{P}}$ be unanimous on $u \succ v$ and $\mathcal{D}^{\mathrm{b}}{}_{\mathcal{P}}$ be unanimous on $v \succ u$. By S-PC, the representational mechanism f must converge upon $u \succ v$ or $v \succ u$ with probability $1 - e^{\Omega(|\mathcal{S}|)}$ in the full population.

Assume that f converges upon $u \succ v$, in which case $\mathcal{D}^{\mathrm{a}}{}_{\mathcal{P}}$ is weakly decisive over $u \succ v$. By the field expansion lemma, $\mathcal{D}^{\mathrm{a}}{}_{\mathcal{P}}$ is decisive over $u \succ v$. We further partition $\mathcal{D}^{\mathrm{a}}{}_{\mathcal{P}}$ into $1:2$ portions $\mathcal{D}^{\mathrm{a},1}{}_{\mathcal{P}}$ and $\mathcal{D}^{\mathrm{a},2}{}_{\mathcal{P}}$. Then, we let:

- $\mathcal{D}^{\mathrm{a},1}{}_{\mathcal{P}}$ (probability mass $\frac{2}{9}$) be unanimous on $u \succ v \succ w$,
- $\mathcal{D}^{\mathrm{a},2}{}_{\mathcal{P}}$ (probability mass $\frac{4}{9}$) be unanimous on $w \succ u \succ v$, and
- $\mathcal{D}^{\mathrm{b}}{}_{\mathcal{P}}$ (probability mass $\frac{1}{3}$) be unanimous on $v \succ w \succ u$.

[9] *i.e.*, issues whose privilege graphs contain simple cycles of length *at least* 3.

Since $\mathcal{D}^{\mathrm{a}}{}_{\mathcal{P}}$ is decisive, we have $u \succ_{f(\mathcal{S})} v$ with probability $1 - e^{\Omega(|\mathcal{S}|)}$. Since the probability masses of all subpopulations have denominators 3 or 9 (both odd numbers), there can be no ties, and f must converge due to S-PC. Given that $u \succ_{f(\mathcal{S})} v$, with probability $1 - e^{\Omega(|\mathcal{S}|)}$, we have either $u \succ_{f(\mathcal{S})} w$ or $w \succ_{f(\mathcal{S})} v$. In the former case, $\mathcal{D}^{\mathrm{a},1}{}_{\mathcal{P}}$ is weakly decisive over $u \succ w$, and in the latter case, $\mathcal{D}^{\mathrm{a},2}{}_{\mathcal{P}}$ is weakly decisive over $w \succ v$. In either case, the subpopulation is decisive by the field expansion lemma, but its probability mass ($\frac{2}{9}$ or $\frac{4}{9}$) is smaller than 0.5, contradicting PND (Lemma C.5).

We can similarly show that f converging upon $v \succ u$ leads to a contradiction. Therefore, no such f can exist. Note that in the proof above, we have been misusing $|\mathcal{S}|$ to denote the number of samples that falls into issue i, which is not a problem since i's probability mass $\Pr_{\mathcal{D}_{\mathcal{I}}}[i] > 0$ is a positive constant (we can pick the i with the largest probability mass, which makes the mass only dependent on $\mathcal{D}_{\mathcal{I}}$ itself), and the asymptotic behavior of the convergence probability is the same.

Then, assuming that all issues (except possibly those with zero probability in $\mathcal{D}_{\mathcal{I}}$) have specific non-cyclic privilege graphs, we construct a candidate space $\mathcal{C}$ consistent with the privilege graphs, and a mechanism f that satisfies the axioms for all $\mathcal{D}_{\mathcal{P}}$.

First consider the case where $|\mathcal{I}| = 1, \mathcal{I} = \{i\}$. Since $\phi(i)$ is transitive (Lemma C.13) but doesn't have simple cycles of length at least 3, it follows that $\phi(i)$ doesn't have any SCC containing 3 or more vertices. Fix an arbitrary topological order $(g_1, \cdots, g_m)$ of vertices in the DAG $\mathrm{cond}(\phi(i))$, where $\{g_j\}_{j=1}^m$ constitute a partition of $[N]$, and $|\{g_j\}| \in \{1, 2\}$ for each j; let $g_{k_1}, \cdots, g_{k_l}$ be the SCCs of size 2. Now consider translating $(g_1, \cdots, g_m)$ into an ordering in $\mathrm{LO}(N)$, where the vertices in each SCC are ordered arbitrarily, and vertices in different SCCs are ordered according to the topological order $(g_1, \cdots, g_m)$. There are 2^l ways to perform such translation, and we define $\mathcal{C} \subset \mathrm{LO}(N)$ to be the set of these 2^l orderings. It can be verified that the G_i resulting from $\mathcal{C}$ contains $\phi(i)$ as a subgraph.

We then define the mechanism f. For each $g_{k_j} = \{u_j, v_j\}$ $(1 \le j \le l)$, it examines if a majority in the sampels prefer u_j over v_j. If yes, it outputs the ordering that places u_j above v_j; otherwise, it outputs the ordering that places v_j above u_j. Ties are broken arbitrarily. Let us now verify that f satisfies PPE, S-PIIA, and S-PC for all $\mathcal{D}_{\mathcal{P}}$.

- PPE: Relative orderings between outcomes in different SCCs are fixed by the topological order, and therefore the outcomes c, c' in the PPE axiom must be in the same SCC. f resolves this preference using a majority vote, which, by Hoeffding's inequality, has a convergence probability of $1 - e^{\Omega(|\mathcal{S}|)}$.
- S-PIIA: It can be verified that if $c \succ c'$ and $c' \succ c$ are both privileged orderings in i, then c, c' must be in the same SCC. f resolves this preference using a majority vote that considers only the relative preference between c and c', and thus satisfies S-PIIA.
- S-PC: Again, c, c' must be in the same SCC. f resolves this preference using a majority vote, which, by Hoeffding's inequality, has a convergence probability of $1 - e^{\Omega(|\mathcal{S}|)}$ when the population is non-uniform on c, c'.

In the general case where $|\mathcal{I}| > 1$, we can apply the above construction to each issue independently, and define $\mathcal{C}$ to be the Cartesian product of the constructed candidate spaces. The mechanism f is then defined to apply the constructed mechanism for each issue independently. It can be verified that f satisfies PPE, S-PIIA, and S-PC for all $\mathcal{D}_{\mathcal{P}}$. □

Theorem 4 is a generalization of Arrow's theorem and Theorem 3 to arbitrary candidate spaces $\mathcal{C}$. It shows that the cyclicity is both necessary and sufficient for the impossibility of a representative mechanism satisfying the axioms. However, the "necessary" part is not as strong as one would hope, since it constructs counterexample for at least one $\mathcal{C}$ associated with each non-cyclic privilege graph, instead of showing that the impossibility holds for all $\mathcal{C}$ associated with the privilege graph. It is an open question whether such a stronger necessity holds, and shall be the subject of future research.

The $N \ge 3$ condition ("at least 3 outcomes") in Theorem 3 (as well as in the original Arrow's theorem) is replaced by the cyclicity condition here ("at least 3 outcomes in a cycle, in the privilege graph"). Intuitively speaking, the cyclicity condition is a more general condition that identifies a "core structure" in the candidate space $\mathcal{C}$ that is incompatible with the axioms.

The cyclicity condition can be checked computationally in linear time (in the number of vertices and edges) for any privilege graph, since it is equivalent to the existence of SCCs with at least 3 vertices.