Policy Aggregation

Parand A. Alamdari[1, 2], Soroush Ebadian[1], Ariel D. Procaccia[3]
[1]University of Toronto, [2]Vector Institute, [3]Harvard University
{parand, soroush}@cs.toronto.edu, arielpro@seas.harvard.edu

Abstract

We consider the challenge of AI value alignment with multiple individuals that have different reward functions and optimal policies in an underlying Markov decision process. We formalize this problem as one of *policy aggregation*, where the goal is to identify a desirable collective policy. We argue that an approach informed by social choice theory is especially suitable. Our key insight is that social choice methods can be reinterpreted by identifying ordinal preferences with volumes of subsets of the *state-action occupancy polytope*. Building on this insight, we demonstrate that a variety of methods — including approval voting, Borda count, the proportional veto core, and quantile fairness — can be practically applied to policy aggregation.

1 Introduction

Early discussion of AI value alignment had often focused on learning desirable behavior from an individual teacher, for example, through inverse reinforcement learning [22, 1]. But, in recent years, the conversation has shifted towards aligning AI models with large groups of people or even entire societies. This shift is exemplified at a policy level by OpenAI's "democratic inputs to AI" program [31] and Meta's citizens' assembly on AI governance [6], and at a technical level by the ubiquity of reinforcement learning from human feedback [25] as a method for fine-tuning LLMs.

We formalize the challenge of value alignment with multiple individuals as a problem that we view as fundamental — *policy aggregation*. Our starting point is the common assumption that the environment can be represented as a *Markov decision process (MDP)*. While the states, actions and transition functions are shared by all agents, their reward functions — which incorporate values, priorities or subjective beliefs — may be different. In particular, each agent has its own optimal policy in the underlying MDP. *How should we aggregate the individual policies into a desirable collective policy?*

A naïve answer is to define a new reward function that is the sum of the agents' reward functions and compute an optimal policy for this aggregate reward function; such a policy would guarantee maximum *utilitarian social welfare*. This approach has a major shortcoming, however, in that it is sensitive to affine transformations of rewards, so, for example, if we doubled one of the reward functions, the aggregate optimal policy may change. This is an issue because each agent's individual optimal policy is invariant to (positive) affine transformations of rewards, so while it is possible to recover a reward function that induces an agent's optimal policy by observing their actions over time, it is impossible to distinguish between reward functions that are affine transformations of each other. More broadly, economists and moral philosophers have long been skeptical about *interpersonal comparisons of utility* [14] due to the lack of universal scale — an issue that is especially pertinent in our context. Therefore, aggregation methods that are invariant to affine transformations are strongly preferred.

Our approach. To develop such aggregation methods, we look to social choice theory, which typically deals with the aggregation of *ordinal* preferences. To take a canonical example, suppose agents report *rankings* over m alternatives. Under the *Borda count* rule, each voter gives $m - k$

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

points to the alternative they rank in the k'th position, and the alternative with most points is selected. The voting approach can be directly applied to our setting. For each agent, it is (in theory) possible to compute the value of every possible (deterministic) policy, and rank them all by value. Then, any standard voting rule, such as Borda count, can be used to aggregate the rankings over policies and single out a desirable policy. The caveat, of course, is that this method is patently impractical, because the number of policies is exponential in the number of states of the MDP.

The main insight underlying our approach is that ordinal preferences over policies have a much more practical volumetric interpretation in the *state-action occupancy polytope* $\mathcal{O}$. Roughly speaking, a point in the state-action occupancy polytope represents a (stochastic) policy through the frequency it is expected to visit different state-action pairs. If a policy is preferred by an agent to a subset of policies $\mathcal{O}'$, its "rank" is the volume of $\mathcal{O}'$ as a fraction of the volume of $\mathcal{O}$. The "score" of a policy under Borda count, for example, can be interpreted as the sum of these "ranks" over all agents.

Our results. We investigate two classes of rules from social choice theory, voting rules and those that guarantee a notion of fairness. By mapping ordinal preferences to the state-action occupancy polytope, we adapt the different rules to the policy aggregation problem.

The former class is examined in Section 4; we focus on volumetric interpretations of α-*approval* (including the ubiquitous *plurality* rule, which is the special case of $\alpha = 1$) and the aforementioned Borda count. Computation is a challenge for these voting rules, and indeed we establish several computational hardness results. To overcome this obstacle, we implement voting rules for policy aggregation through mixed integer linear programming, which leads to practical solutions.

We examine the latter class in Appendix A. As a warm-up we start from the notion of *proportional veto core*; it follows from recent work by Chaudhury et al. [5] that a volumetric interpretation of this notion is nonempty and can be computed efficiently. We then turn to *quantile fairness*, which was recently introduced by Babichenko et al. [3]; we prove that the volumetric interpretation of this notion yields guarantees that are far better than those known for the original, discrete setting, and we design a computationally efficient algorithm to optimize those guarantees.

Our experiments in Appendix E evaluate the policies returned by different rules based on fairness, and illustrate the advantage of our approach over rules that optimize measures of social welfare.

Related work. Noothigattu et al. [23] consider a setting related to ours, in that different agents have different reward functions and different policies that must be aggregated. However, they assume that the agents' reward functions are noisy perturbations of a ground-truth reward function, and the goal is to learn an optimal policy according to the ground-truth rewards.

There is a body of work on multi-objective reinforcement learning (MORL) and planning that uses a scalarization function to reduce the problem to a single-objective one [27, 16] or that seeks a set of policies approximating the problem's Pareto frontier [29]. A closely related line of work focuses on fairness in sequential decision making, often taking the scalarization approach by maximizing a (cardinal) social welfare function [24, 28, 20, 9, 18]. Alamdari et al. [2] study this problem in a non-Markovian setting, where fairness depends on the history of actions over time. A shortcoming of these solutions is that they are not invariant to affine transformations of rewards.

Concurrent work explores reinforcement learning from human feedback (RLHF) from a social choice perspective, where the reward model is based on pairwise human preferences, often constructed using the Bradley-Terry model [4]. Zhong et al. [32] consider the maximum Nash and egalitarian welfare solutions, and Gokul et al. [12] propose a method based on maximal lotteries due to Fishburn [10].

2 Preliminaries

For $t \in \mathbb{N}$, let $[t] = \{1, 2, \ldots, t\}$. For a closed set S, let $\Delta(S)$ denote the probability simplex over the set S. We denote the dot product of two vectors as $\langle x, y\rangle = \sum_{i=1}^{d} x_i \cdot y_i$ for $x, y \in \mathbb{R}^d$. A *halfspace* in $\mathbb{R}^d$ determined by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ is the set of points satisfying $\{x \in \mathbb{R}^d \mid \langle x, w\rangle \leqslant b\}$. A *polytope* $\mathcal{O} \subseteq \mathbb{R}^d$ is the intersection of a finite number of halfspaces, i.e., a convex subset of the d-dimensional space $\mathbb{R}^d$ determined by a set of linear constraints $\{x \mid Ax \leqslant b\}$ where $A \in \mathbb{R}^{k \cdot d}$ is a matrix of coefficients of k linear inequalities and $b \in \mathbb{R}^k$.

A multi-objective Markov decision process (MOMDP) is a tuple defined as $M = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, R_1, \ldots, R_n\rangle$ for the average-reward case and $M = \langle \mathcal{S}, d_{\text{init}}, \mathcal{A}, \mathcal{P}, R_1, \ldots, R_n, \gamma\rangle$ for the discounted-reward

case, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, and $\mathcal{P} : (\mathcal{S} \times \mathcal{A}) \to \Delta(\mathcal{S})$ is the transition probability distribution. $\mathcal{P}(s_t, a_t, s_{t+1})$ is the probability of transitioning to state s_{t+1} by taking action a_t in s_t. For $i \in [n]$, $R_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function of the ith agent, the initial state is sampled from $d_{\text{init}} \in \Delta(\mathcal{S})$, and $\gamma \in (0, 1]$ is the discount factor.

A *(Markovian) policy* $\pi(a|s)$ is a probability distribution over the actions $a \in \mathcal{A}$ given the state $s \in \mathcal{S}$. A policy is *deterministic* if at each state s one action is selected with probability of 1, and otherwise it is *stochastic*. The expected *average* return of agent i for a policy π and the expected *discounted* return of agent i for a policy π are defined over an *infinite time horizon* as

$$J_i^{\text{avg}}(\pi) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_{\pi,\mathcal{P}} \left[\sum\nolimits_{t\in[T]} R_i(s_t, a_t) \right], J_i^{\gamma}(\pi) = (1-\gamma)\, \mathbb{E}_{\pi,\mathcal{P}} \left[\sum\nolimits_{t=1}^{\infty} \gamma^t R_i(s_t, a_t)|_{s_1\sim d_{\text{init}}} \right].$$

Definition 1 (state-action occupancy measure)**.** *Let $\mathcal{P}_\pi^t$ be the probability measure over states at time t under policy π. The state-action occupancy measure for state s and action a is defined as*

$$d_\pi^{\text{avg}}(s, a) = \lim_{T\to\infty} \frac{1}{T}\ \mathbb{E} \left[\sum\nolimits_{t\in[T]} \mathcal{P}_\pi^t(s)\pi(a|s) \right],\ d_\pi^{\gamma}(s, a) = (1-\gamma)\, \mathbb{E} \left[\sum\nolimits_{t=1}^{\infty} \gamma^t \mathcal{P}_\pi^t(s)\pi(a|s) \right].$$

For both the average and discounted cases, we can rewrite the expected return as the dot product of the state-action occupancy measures and rewards, that is, $J_i(\pi) = \sum_{(s,a)} d_\pi(s, a) \cdot R_i(s, a) = \langle d_\pi, R_i \rangle$. In addition, the policy can be calculated given the occupancy measure as $\pi(a|s) = d_\pi(s, a)/(\sum_a d_\pi(s, a))$ if $\sum_a d_\pi(s, a) > 0$, and $\pi(a|s) = 1/|\mathcal{A}|$ otherwise.

Definition 2 (state-action occupancy polytope [26, 30])**.** *For a MOMDP M in the average-reward case, the space of valid state-action occupancies is the polytope*

$$\mathcal{O}^{\text{avg}} = \left\{ d_\pi^{\text{avg}} \mid d_\pi^{\text{avg}} \geqslant 0, \sum_{s,a} d_\pi^{\text{avg}}(s, a) = 1, \sum_a d_\pi^{\text{avg}}(s, a) = \sum_{s',a'} \mathcal{P}(s', a', s) d_\pi^{\text{avg}}(s', a'), \forall s \in \mathcal{S} \right\}.$$

3 Occupancy Polytope as the Space of Alternatives

In a MOMDP M, each agent i incorporates their values and preferences into their respective reward function R_i. Agent i prefers π over π' if and only if π achieves higher expected return, $J_i(\pi) > J_i(\pi')$, and is indifferent between two policies π and π' if and only if $J_i(\pi) = J_i(\pi')$. As discussed before, given a state-action occupancy measure d_π in the state-action occupancy polytope $\mathcal{O}$, we can recover the corresponding policy π. Therefore, we can interpret $\mathcal{O}$ as the domain of all possible alternatives over which the n agents have heterogeneous *weak* preferences (with ties). Agent i prefers d_π to $d_{\pi'}$ in $\mathcal{O}$ if and only if they prefer π to π'. We study the policy aggregation problem through this lens; specifically, we design or adapt voting mechanisms where the (continuous) space of alternatives is $\mathcal{O}$ and agents have weak preferences over them determined by their reward functions R_i.

Affine transformation. A particular benefit of this interpretation, as mentioned before, is that all positive affine transformations of the reward functions, i.e., $aR_i + b$ for all $a \in \mathbb{R}_{\geqslant 0}$ and $b \in \mathbb{R}$, yield the same weak ordering over the polytope. Hence, we can assume w.l.o.g. that $J_i(\pi) \in [0, 1]$.

Volumetric definitions. A major difference between voting over a continuous space of alternatives and the classical voting setting is that the domain of alternatives is infinite and not all voting mechanisms can be directly applied to the policy aggregation problem. In particular, various voting rules require reasoning about the rank of an alternative or the size of some subset of alternatives. For instance, the Borda count of an alternative c over a finite set of alternatives is defined as the number (or fraction) of candidates ranked below c. In the continuous setting, we use the *measure* or *volume* of a subset of alternatives to refer to their size. For a measurable subset $\mathcal{O}' \subseteq \mathcal{O}$, let $\text{vol}(\mathcal{O}')$ denote its measure. The ratio $\text{vol}(\mathcal{O}')/\text{vol}(\mathcal{O})$ is the fraction of alternatives that lie in $\mathcal{O}'$. We can also define the *expected return distribution* of an agent over $\mathcal{O}$ as a random variable that maps a policy to its expected return, i.e., one that maps $d_\pi \in \mathcal{O}$ to $\langle d_\pi, R_i \rangle$. The pdf and cdf of this r.v. is defined below.

Definition 3 (expected return distribution)**.** *For a MOMDP M and $v \in \mathbb{R}$, the* expected return distribution *of agent $i \in [n]$ is defined as*

$$f_i(v) = \frac{1}{\text{vol}(\mathcal{O})} \int_{x\in\mathcal{O}} \delta(v - \langle x, R_i \rangle)\, dx, \quad F_i(v) = \int_{x=-\infty}^{v} f_i(x)\, dx = \frac{\text{vol}(\mathcal{O} \cap \{\langle x, R_i \rangle \leqslant v\})}{\text{vol}(\mathcal{O})},$$

where f_i and F_i are the pdf and cdf of the expected return distribution and $\delta(\cdot)$ is the Dirac delta function indicating $v = \langle x, R_i \rangle$.

It follows from the Brunn–Minkowski inequality [13] that f_i is increasing up to its mode $\mathrm{mode}(f_i) \in \arg\max_{v \in \mathbb{R}} f_i(v)$ and decreasing afterwards. Since f_i (pdf) is the derivative of F_i (cdf), the unimodality of f_i implies that F_i is a convex function in $(-\infty, \mathrm{mode}(f_i)]$ and concave in $[\mathrm{mode}(f_i), \infty)$.

In our algorithms, we use a subroutine that measures the volume of a polytope, which we denote by $\text{vol-comp}(\{Ax \leqslant b\})$. Dyer et al. [8] designed a fully polynomial time randomized approximation scheme (FPRAS) for computing the volume of polytopes.

4 Policy Aggregation with Voting Rules

In this section, we adapt existing voting rules from the discrete setting to policy aggregation and discuss their computational complexity. We defer the proofs to Appendix D.

Plurality. The plurality winner is the policy that achieves the maximum number of plurality votes or "approvals," where agent i approves a policy π if it achieves their maximum expected return $J_i(\pi) = \max_{\pi'} J_i(\pi') = 1$. Hence the plurality winner is a policy in $\arg\max_\pi \sum_{i \in [n]} \mathbb{I}[J_i(\pi) = 1]$. This formulation does not require the volumetric interpretation. However, in contrast to the discrete setting where one can easily count the approvals for all candidates, we show that solving this problem in the context of policy aggregation is not only $\mathbf{NP}$-hard, but hard to approximate.

Theorem 1. *For any fixed $\epsilon \in (0, 1)$, there is no polynomial-time $\frac{1}{n^{1/2-\epsilon}}$-approximation algorithm for the maximum plurality score unless $\mathbf{P} = \mathbf{NP}$.*

Nevertheless, we can compute plurality in practice, as we discuss below.

α-approval. We extend the k-approval rule using the volumetric interpretation of the occupancy polytope. For some $\alpha \in [0, 1]$, agents approve a policy π if its return is among their top $1 - \alpha$ fraction of $\mathcal{O}$, i.e., $F_i(J_i(\pi)) \geqslant \alpha$. The α-approval winner is a policy that has the highest number of α-approvals, so it is in $\arg\max_\pi \sum_{i \in [n]} \mathbb{I}[F_i(J_i(\pi)) \geqslant \alpha]$. Plurality is equivalent to 1-approval. There can be infinitely many policies that have the maximum approval score and, to avoid a suboptimal decision, one can return a Pareto optimal solution among the set of α-approval winner policies.

It follows from Theorem 4 and Proposition 5 in App. A that for $\alpha \leqslant 1/e$, there always exists a policy that all agents approve and such policies can be found in polynomial time, assuming access to an oracle for volumetric computations. Therefore, the problem of finding an α-approval winner is "easy" for $\alpha \in (0, 1/e)$. In sharp contrast, for $\alpha = 1$ — namely, plurality — Theorem 1 gives a hardness of approximation. The next theorem shows the hardness of computing α-approval for $\alpha \in (7/8, 1]$.

Theorem 2. *For $\alpha \in (7/8, 1]$, computing a policy with the highest α-approval score is $\mathbf{NP}$-hard.*

Given the above hardness result, to compute the α-approval rule, we turn to *mixed-integer linear programming (MILP)*; Algorithm 3 is presented in Appendix B. Importantly, Algorithm 3 has one binary variable per agent and only n constraints which is key to its practicability.

Borda count. The Borda count rule also has a natural definition in the continuous setting. In the discrete setting, the Borda score of agent i for alternative c is the number of alternatives c' such that $c \succ_i c'$. In the continuous setting, $F_i(J_i(\pi))$ indicates the volume of the occupancy polytope to which agent i prefers π. The Borda count rule then selects a policy among $\arg\max_\pi \sum_{i \in [n]} F_i(J_i(\pi))$.

The computational complexity of the Borda count rule remains an interesting open question, though we make progress on two fronts. The first is based on the observation in Section 3 that F_i is concave in range $[\mathrm{mode}(f_i), \infty)$. We assume that the max Borda count policy π appears in the concave portion of all agents, i.e., $J_i(\pi) \geqslant \mathrm{mode}(f_i)$ for all $i \in [n]$. Then, the problem becomes a maximization of the concave objective $\max \sum_i F_i(\langle d_\pi, R_i \rangle)$ over the convex domain $\{d_\pi \in \mathcal{O} \mid \langle d_\pi, R_i \rangle \geqslant \mathrm{mode}(f_i), \forall i \in [n]\}$. Second, we develop an MILP (see Algorithm 4 in Appendix B) that finds an approximate max Borda count policy and has $1/\epsilon$ binary variables per agent.

5 Discussion

Our approach has some limitations. First, computation is a concern; while our results show that ordinal solutions can be applied practically, further algorithmic advances are needed for scalability beyond thousands of agents, states, and actions. Future work could explore applying our methods to continuous state/action spaces and online reinforcement learning. A second limitation is the possibility of strategic behavior, in the sense that agents could gain from misreporting their preferences.

Acknowledgments

We gratefully acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada CIFAR AI Chairs Program (Vector Institute). This work was also partially supported by the Cooperative AI Foundation; by the National Science Foundation under grants IIS-2147187, IIS-2229881 and CCF-2007080; and by the Office of Naval Research under grants N00014-20-1-2488 and N00014-24-1-2704.

References

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, pages 1–8, 2004.

[2] Parand A. Alamdari, Toryn Q. Klassen, Elliot Creager, and S. A. Mcilraith. Remembering to be fair: Non-Markovian fairness in sequential decision making. In *Proceedings of the 41st International Conference on Machine Learning*, pages 906–920. PMLR, 2024.

[3] Y. Babichenko, M. Feldman, R. Holzman, and V. V. Narayan. Fair division via quantile shares. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC)*, pages 1235–1246, 2024.

[4] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[5] B. R. Chaudhury, A. Murhekar, Z. Yuan, B. Li, R. Mehta, and A. D. Procaccia. Fair federated learning via the proportional veto core. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 42245–42257. PMLR, 2024.

[6] N. Clegg. Bringing people together to inform decision-making on generative AI. Blog post, 2023. URL https://about.fb.com/news/2023/06/generative-ai-community-forum.

[7] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and Y. Halpern. Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *Proceedings of the 3rd Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 525–534, 2020.

[8] M. Dyer, A. Frieze, and R. Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM*, 38(1):1–17, 1991.

[9] Z. Fan, N. Peng, M. Tian, , and B. Fain. Welfare and fairness in multi-objective reinforcement learning. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 1991–1999, 2023.

[10] P. C Fishburn. Probabilistic social choice based on simple voting comparisons. *The Review of Economic Studies*, 51(4):683–692, 1984.

[11] M.R. Garey, D.S. Johnson, and L. Stockmeyer. Some simplified np-complete graph problems. *Theoretical Computer Science*, 1(3):237–267, 1976. ISSN 0304-3975.

[12] S. Gokul, C. Dann, R. Kidambi, Z. S. Wu, and A. Agarwal. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

[13] B. Grünbaum. Partitions of mass-distributions and of convex bodies by hyperplanes. *Pacific Journal of Mathematics*, 10(4):1257–1261, 1960.

[14] P. J. Hammond. Interpersonal comparisons of utility: Why and how they are and should be made. In J. Elster and J. E. Roemer, editors, *Interpersonal Comparisons of Well-Being*, chapter 7, pages 200–254. Cambridge University Press, 1991.

[15] Johan Håstad. Clique is hard to approximate within n 1- ε. *Acta Mathematica*, 182:105–142, 1999.

[16] C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M Zintgraf, R. Dazeley, F. Heintz, et al. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26, 2022.

[17] E. Ianovski and A. Y. Kondratev. Computing the proportional veto core. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[18] P. Ju, A. Ghosh, and N. Shroff. Achieving fairness in multi-agent MDP using reinforcement learning. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2023.

[19] A. Y. Kondratev and E. Ianovski. Veto core consistent preference aggregation. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 1020–1028, 2024.

[20] D. Mandal and J. Gan. Socially fair reinforcement learning. *CoRR*, abs/2208.12584, 2022.

[21] H. Moulin. The proportional veto principle. *The Review of Economic Studies*, 48(3):407–416, 1981.

[22] A. Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pages 663–670, 2000.

[23] R. Noothigattu, T. Yan, and A. D. Procaccia. Inverse reinforcement learning from like-minded teachers. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, pages 9197–9204, 2021.

[24] W. Ogryczak, P. Perny, and P. Weng. A compromise programming approach to multiobjective markov decision processes. *International Journal of Information Technology & Decision Making*, 12(05):1021–1053, 2013.

[25] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[26] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[27] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

[28] U. Siddique, P. Weng, and M. Zimmer. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 8905–8915, 2020.

[29] R. Yang, X. Sun, and K. Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[30] T. Zahavy, B. O'Donoghue, G. Desjardins, and S. Singh. Reward is enough for convex MDPs. *Advances in Neural Information Processing Systems*, 34:25746–25759, 2021.

[31] W. Zaremba, A. Dhar, L. Ahmad, T. Eloundou, S. Santurkar, S. Agarwal, and J. Leung. Democratic inputs to AI. Blog post, 2023. URL https://openai.com/blog/democratic-inputs-to-ai.

[32] H. Zhong, Z. Deng, W. J. Su, Z. S. Wu, and L. Zhang. Provable multi-party reinforcement learning with diverse human feedback. *arXiv preprint arXiv:2403.05006*, 2024.

ALGORITHM 1: Seq. ϵ-Prop. Veto Core [5]

$\mathcal{O}_0 \leftarrow \mathcal{O}, \quad \delta \leftarrow \frac{1}{n} - \frac{\epsilon}{n+1}$
for $i = 1$ *to* n **do**
  Using binary search find $v_i^* \in [0,1]$ s.t.
    $\mathrm{vol}(\mathcal{O}_{i-1} \cap \{\langle x, R_i \rangle \leqslant v_i^*\}) \approx \delta \cdot \mathrm{vol}(\mathcal{O})$
  $\mathcal{O}_i \leftarrow \mathcal{O}_{i-1} \cap \{\langle x, R_i \rangle \geqslant v_i^*\}$
return *a welfare maximizing point* $d_\pi \in \mathcal{O}_n$

ALGORITHM 2: ϵ-Max Quantile Fairness

Procedure `q-quantile-feasible(q)`:
  $\mathcal{O}_q \leftarrow \{x \in \mathcal{O} \mid \langle x, R_i \rangle \geqslant F_i^{-1}(q), i \in [n]\}$
  if $\mathcal{O}_q$ *is a feasible linear program* **then return** *True* **else return** *False*
Using binary search find maximum $q \in [0,1]$ s.t. `q-quantile-feasible()` is feasible
return a welfare maximizing point $d_\pi \in \mathcal{O}_q$

A Fairness in Policy Aggregation

In this section, we utilize the volumetric interpretation of the state-action occupancy polytope to extend fairness notions from social choice to policy aggregation, and we develop algorithms to compute stochastic policies provably satisfying these notions.

A.1 Proportional Veto Core

The proportional veto core was first proposed by Moulin [21] in the classical social choice setting with a finite set of alternatives where agents have full (strict) rankings over the alternatives. For simplicity, suppose the number of alternatives m is a multiple of n. The idea of the proportional veto core is that $x\%$ of the agents should be able to veto $x\%$ of the alternatives. More precisely, for an alternative c to be in the proportional veto core, there should not exist a coalition S that can "block" c using their proportional veto power of $|S|/n$. S blocks c if they can unanimously suggest $m(1 - |S|/n)$ candidates that they prefer to c. For instance, if c is in the proportional veto core, it cannot be the case that a coalition of 60% of the agents unanimously prefer 40% of the alternatives to c.

Chaudhury et al. [5] extended this notion to a continuous domain of alternatives in the federated learning setting. We show that such an extension also applies to policy aggregation.

Definition 4 (proportional veto core)**.** *Let $\epsilon \in (0, 1/n)$. For a coalition of agents $S \subseteq [n]$, let $\mathrm{veto}(S) = |S|/n$ be their veto power. A point $d_\pi \in \mathcal{O}$ is* blocked *by a coalition S if there exists a subset $\mathcal{O}' \subseteq \mathcal{O}$ of measure $\mathrm{vol}(\mathcal{O}') \geqslant 1 - \mathrm{veto}(S) + \epsilon$ such that all agents in S prefer all points in $\mathcal{O}'$ to d_π, i.e., $d_{\pi'} \succ_i d_\pi$ for all $d_{\pi'} \in \mathcal{O}'$ and $i \in S$. A point d_π is in the* ϵ-proportional veto core *if it is not blocked by any coalition.*

A candidate in the proportional veto core satisfies desirable properties that are extensively discussed in prior work [21, 5, 19, 17]. It is worth mentioning that any candidate in the ϵ-proportional veto core, besides the fairness aspect, is also economically efficient as it satisfies *ϵ-Pareto optimality*. An alternative π is ϵ-Pareto optimal if there does not exist another alternative π' such that $J_i(\pi') \geqslant J_i(\pi) + \epsilon$ for all i, with at least one inequality being strict. This holds since the grand coalition $S = [n]$ can veto any ϵ-Pareto dominated alternative.

Moulin [21] proved that the proportional veto core is nonempty in the discrete setting and Chaudhury et al. [5] proved it for the continuous setting. The following result is a corollary of Theorem 1 of Chaudhury et al. [5].

Theorem 3. *Let $\epsilon \in (0, 1/n)$. For a policy aggregation problem, the ϵ-proportional veto core is nonempty. Furthermore, such policies can be found in polynomial time using $O(\log(1/\epsilon))$ many calls per agent to* vol-comp.

We refer the reader to the paper of Chaudhury et al. [5] for the complete proof, and provide a high-level description of Algorithm 1, which finds a point in the proportional veto core. Algorithm 1 iterates over the agents and lets the ith agent "eliminate" roughly $1/n \cdot \mathrm{vol}(\mathcal{O})$ of the remaining space of alternatives. That is, agent i finds the hyperplane $H_i = \{\langle x, R_i \rangle \leqslant v_i^*\}$ such that its intersection with O_{i-1} (the remaining part of $\mathcal{O}$ at the ith iteration) has a volume of approximately $\mathrm{vol}(\mathcal{O})/n$. This value, for each agent, can be found by doing a binary search over v_i^* to a precision of $[v_i^* - \epsilon, v_i^*]$ by $O(\log(1/\epsilon))$ calls to the volume estimating subroutine vol-comp.

A.2 Quantile Fairness

Next, we consider an egalitarian type of fairness based on the occupancy polytope, building on very recent work by Babichenko et al. [3] in the discrete setting; surprisingly, we show that it is possible to obtain stronger guarantees in the continuous setting.

Babichenko et al. [3] focus on the fair allocation of a set of m indivisible items among n agents, where each item must be fully allocated to a single agent. They quantify the extent an allocation A is *fair* to an agent i by the fraction of allocations over which i prefers A (note that the number of all discrete allocations is n^m). In other words, if one randomly samples an allocation, the fairness is measured by the probability that A is preferred to the random allocation. An allocations is *q-quantile fair* for $q \in [0, 1]$ if *all* agents consider this allocation among their top q-quantile allocations. Babichenko et al. [3] aim to find a universal value of q such that for any fair division instance, a q-quantile fair allocation exists. They make an interesting connection between q-quantile fair allocations and the Erdős Matching Conjecture, and show under the assumption that the conjecture holds, $(1/2e)$-quantile fair allocations exist for all instances.[1]

We ask the same question for policy aggregation, and again, a key difference is that our domain of alternatives is continuous. The notion of q-quantile fairness extends well to our setting. Agents assess the fairness of a policy π based on the fraction of the occupancy polytope (i.e., the set of all policies) to which they prefer π, or equivalently, the probability that they prefer the chosen policy to a randomly sampled policy.

Definition 5 (q-quantile fairness)**.** *For a MOMDP M and $q \in [0, 1]$, a policy π is q-quantile fair if for every agent $i \in [n]$, π is among i's top q-fraction of policies,* $\mathrm{vol}(\mathcal{O} \cap \{\langle x, R_i \rangle \leqslant J_i(\pi)\}) \geqslant q \cdot \mathrm{vol}(\mathcal{O})$.

We show that a $(1/e)$-quantile fair policy always exist and that this ratio is tight; note that this bound is twice as good as that of Babichenko et al. [3], and it is unconditional. We prove this by making a connection to an inequality due to Grünbaum [13]. The *centroid* of a polytope P is defined as $\mathrm{centroid}(P) = \frac{\int_{x \in P} x \, dx}{\int_{x \in P} 1 \, dx}$.

Lemma 1 (Grunbaum's Inequality)**.** *Let P be a polytope and w a direction in $\mathbb{R}^d$. For the halfspace $H = \{x \mid \langle w, x - \mathrm{centroid}(P) \rangle \geqslant 0\}$, it holds that*

$$\frac{1}{e} \leqslant \left(\frac{d}{d+1}\right)^d \leqslant \frac{\mathrm{vol}(P \cap H)}{\mathrm{vol}(P)} \leqslant 1 - \left(\frac{d}{d+1}\right)^d \leqslant 1 - \frac{1}{e},$$

Furthermore, this is tight for the n-dimensional simplex.

Theorem 4. *For every MOMDP M, there always exist q-quantile fair policy where $q = (\frac{\ell-1}{\ell})^{\ell-1}$ and $\ell = |\mathcal{S}| \cdot |\mathcal{A}|$. Note that $q \geqslant \frac{1}{e} \approx 36.7\%$. Furthermore, this bound is tight: For any ℓ, there is an instance with a single state and ℓ actions where no q-quantile fair policy exists for any $q > (\frac{\ell-1}{\ell})^{\ell-1}$.*

Proof. First, we show that the centroid of the occupancy polytope $c = \mathrm{centroid}(\mathcal{O})$ is q-quantile fair policy for the aformentioned q. Since $\mathcal{O}$ is a subset of the $(n-1)$-simplex (see Definition 2), $\mathcal{O}$ has a nonzero volume in some lower dimensional space $\ell' \leqslant |\mathcal{S}| \cdot |\mathcal{A}| - 1$. By invoking Grunbaum's inequality (Lemma 1) with w_i being equal to R_i projected to the ℓ'-dimensional subspace for all agents $i \in [n]$, we have that $\mathrm{vol}(\mathcal{O} \cap H_i) \geqslant (\frac{\ell'}{\ell'+1})^{\ell'} \cdot \mathrm{vol}(\mathcal{O})$ where $H_i = \{\langle x - c, w_i \rangle \geqslant 0\} = \{\langle x, R_i \rangle \geqslant J_i(c)\}$. Since $\ell' \leqslant \ell - 1$, we have $(\frac{\ell'}{\ell'+1})^{\ell'} \geqslant (\frac{\ell-1}{\ell})^{\ell-1}$, which completes the proof.

To show tightness, take a MOMDP with a single state $\mathcal{S} = \{s\}$ — hence a constant transition function $\mathcal{P}(\cdot) = s$ — and ℓ actions $\mathcal{A} = \{a_1, \ldots, a_\ell\}$ and ℓ agents. The reward function of agent i is $R_i(s, a_i) = 1$ and 0 otherwise. The state-action occupancy polytope of this MOMDP is the $(\ell-1)$-dimensional simplex $\mathcal{O} = \{\sum_a d_\pi(s, a) = 1, d_\pi \in \mathbb{R}^{1 \times \ell}_{\geqslant 0}\}$. Take any point in $d_\pi \in \mathcal{O}$. There exists at least one agent i that has $J_i(d_\pi) = d_\pi(s, a_i) \leqslant \frac{1}{\ell}$. Take the halfspace $H_i = \{\langle x, R_i \rangle \geqslant \frac{1}{\ell}\}$. Observe that $H_i \cap \mathcal{O}$ is equivalent to a smaller $(\ell-1)$-dimensional simplex $\{\sum_a d_\pi(s, a) = 1 - \frac{1}{\ell}\}$ which has volume of $\mathrm{vol}\left((1 - \frac{1}{\ell})\mathcal{O}\right) = \left(\frac{\ell-1}{\ell}\right)^{\ell-1} \mathrm{vol}(\mathcal{O})$. Therefore, $\frac{\mathrm{vol}(\mathcal{O} \cap H_i)}{\mathrm{vol}(\mathcal{O})} = \left(\frac{\ell-1}{\ell}\right)^{\ell-1}$. $\square$

[1] This result is for arbitrary monotone valuations. For additive valuations, they show $\frac{0.14}{e}$-quantile fair allocations always exist without any assumptions.

ALGORITHM 3: α-Approvals MILP

compute $F_i^{-1}(\alpha)$ for all $i \in [n]$
solve the mixed integer linear program

$$\max \sum_{i\in[n]} a_i$$
$$\text{s.t.} \quad a_i \cdot F_i^{-1}(\alpha) \leqslant \langle d_\pi, R_i \rangle \quad \forall i \in [n]$$
$$d_\pi \in \mathcal{O}, \quad a_i \in \{0,1\} \quad \forall i \in [n]$$

return a Pareto optimal policy subject to max α-approval $\{i \mid a_i = 1\}$

ALGORITHM 4: ϵ-Borda count MILP

compute $F_i^{-1}(k\epsilon)$ for all $i \in [n]$ and $k \in [\frac{1}{\epsilon}]$
solve the mixed integer linear program

$$\max \sum_{i\in[n]} \sum_{k\in[1/\epsilon]} a_{i,k} \cdot (F_i(k\epsilon) - F_i((k-1)\epsilon))$$
$$\text{s.t.} \quad a_{i,k} \cdot k\epsilon \leqslant \langle d_\pi, R_i \rangle \quad \forall i \in [n]$$
$$d_\pi \in \mathcal{O}, \quad a_{i,k} \in \{0,1\}, \quad \forall i \in [n], k \in [1/\epsilon]$$

return a Pareto optimal policy subject to max Borda score

The centroid of the occupancy polytope, as per Theorem 4, attains a worst-case measure of quantile-fairness. However, the centroid policy can be highly suboptimal as it disregards the preferences of the agents involved. For instance, there could exist a 99%-quantile fair policy. To this end, we take an egalitarian approach and aim to find a q-quantile fair policy with the maximum q.

Max quantile fair algorithm. Algorithm 2 searches for the optimal value q^*, for which a q^*-quantile fair policy exists, and gets close to q^* by a binary search. To perform the search, we need a subroutine that checks, for a given value of q, if a q-quantile fair policy exists. Suppose we have the q-quantile expected return $F_i^{-1}(q)$ for all i, that is, the expected return amount v_i such that $F_i(v_i) = q$. Then, the problem of existence of a q-quantile fair policy is equivalent to the feasibility of the linear program $\{x \in \mathcal{O} \mid \langle x, R_i \rangle \geqslant F_i^{-1}(q), i \in [n]\}$, which can be solved in polynomial time. Importantly, after finding a good approximation of q, there can be infinitely many policies that are q-quantile fair and there are various ways to select the final policy after finding the value q. As mentioned earlier, a desirable efficiency property is Pareto optimality; to satisfy it, one can return the policy that maximizes the sum of agents' expected returns among the ones that are q-quantile fair. Finally, to calculate $F_i^{-1}(q)$, we can again use binary search to get δ close to the value v_i for which $F_i(v_i) = q$ using $O(\log 1/\delta)$ calls to vol-comp. The discussion above is summarized below.

Proposition 5. *Assuming an optimal oracle for F_i^{-1}, Algorithm 2 finds a q-quantile fair policy that is ϵ close to the optimal value in polynomial time with $O(\log(1/\epsilon))$ per agent calls to the oracle. A δ-approximation to $F_i^{-1}(q)$ can be computed using $O(\log(1/\delta))$ calls to* vol-comp.

Finally, we make a novel connection between q-quantile fairness and Borda count in Theorem 6. We defer the proof to Appendix D. A corollary of Theorems 4 and 6 is that the policy returned by ϵ-max quantile fair algorithm (Algorithm 2) achieves a $(1/e - \epsilon)$ multiplicative approximation of the maximum Borda score.

Theorem 6. *A q-quantile fair policy is a q-approximation of the maximum Borda score.*

Proof. The Borda count of a policy is defined as $\text{Borda}(\pi) = \sum_{i\in[n]} F_i(J_i(\pi))$. Since $F_i(v) \in [0,1]$, $\max_\pi \sum_{i\in[n]} F_i(J_i(\pi)) \leqslant n$. From Definition 5, for a q-quantile fair policy π_q, we have $F_i(J_i(\pi_q)) \geqslant q$, for all $i \in [n]$. Therefore, $\sum_{i\in[n]} F_i(J_i(\pi_q)) \geqslant qn$ and we have

$$\frac{\text{Borda}(\pi_q)}{\max_\pi \text{Borda}(\pi)} \geqslant \frac{qn}{n} \geqslant q. \qquad \square$$

B Mixed Integer Linear Programs for Voting Rules

α-approvals. Algorithm 3 simply creates n binary variables for each agent i indicating whether i α-approves the policy, i.e., $F_i(J_i(\pi)) \geqslant \alpha$ which is equivalent to $J_i(\pi) \geqslant F_i^{-1}(\alpha)$. To encode the expected return requirement for agent i to approve a policy as a linear constraint, we precompute $F_i^{-1}(\alpha)$. This can be done by a binary search similar to Algorithm 2. Importantly, Algorithm 3 has one binary variable per agent and only n constraints which is key to its practicability.

Borda count. Algorithm 4 is a MILP that finds an approximate max Borda count policy. As a pre-processing step, we estimate F_i for each agent i separately. We measure F_i for the fixed expected return values of $\{\epsilon, 2\epsilon, \ldots, 1-\epsilon, 1\}$. This accounts for $1/\epsilon$ oracle calls to vol-comp per agent. Then, for the MILP, we introduce $1/\epsilon$ binary variables for each agent indicating their ϵ-rounded return

levels, i.e., $a_{i,k} = 1$ iff $\langle d_\pi, R_i \rangle \geqslant k\epsilon$ for $k \in [1/\epsilon]$. The MILP then searches for an occupancy measure $d_\pi \in \mathcal{O}$ with maximum total Borda score among the ϵ-rounded expected return vectors.

Our MILP approximates the Borda score of a return vector r by its ϵ-rounded-down return vector $r_\epsilon = (\lfloor r_i/\epsilon \rfloor \cdot \epsilon \mid i \in [n])$. Therefore, the MILP finds a policy that has a Borda count which is at least as high as the maximum Borda count among ϵ-rounded return vectors. This is *not* necessarily equivalent to a $(1-\epsilon)$ approximation of the max Borda count.

C State-action Occupancy Polytope for Discounted Reward

Definition 6 (State-action Occupancy Polytope for discounted-reward [26, 30])**.** *For a MOMDP M in the discounted-reward case, the space of valid state-action occupancies is the polytope*

$$\mathcal{O}^\gamma = \left\{ d_\pi^\gamma \mid d_\pi^\gamma \geqslant 0, \sum_a d_\pi^\gamma(s,a) = (1-\gamma) d_{\text{init}}(s) + \gamma \sum_{s',a'} \mathcal{P}(s',a',s) d_\pi^\gamma(s',a'), \forall s \in \mathcal{S} \right\}.$$

D Proofs of Section 4

We define a *fully connected MOMDP* as a MOMDP M where for every pair of states $s, s' \in \mathcal{S}$ and any action $a \in \mathcal{A}$, $\mathcal{P}(s,a,s') = \frac{1}{|\mathcal{S}|}$. In all the hardness proofs, we create a fully connected MOMDP. For such MOMDPs, it is not difficult to observe that the state-action occupancy polytope, for both the average and discounted reward, is equivalent to $\{d_\pi \mid \sum_a d_\pi(s,a) = \frac{1}{|\mathcal{S}|}, \forall s \in \mathcal{S}\}$ for both the discounted and average reward case. Furthermore,

$$\pi(a|s) = \frac{d_\pi(s,a)}{\sum_{a' \in \mathcal{A}} d_\pi(s,a')} = |\mathcal{S}| \cdot d_\pi(s,a) \tag{1}$$

D.1 Proof of Theorem 1

Proof of Theorem 1. We show hardness of approximation by a approximation-preserving reduction from the *maximum independent set (MIS)* problem.

Definition 7 (maximum independent set (MIS))**.** *For a graph $G = (V, E)$ with vertex set V and edge set $E \subseteq 2^{\binom{V}{2}}$, the* maximum independent set (MIS) *of G is the maximum subset of vertices $V' \subseteq V$ such that there are no edges between any pair of vertices $v_1, v_2 \in V'$.*

Theorem 7 (Håstad [15])**.** *For any fixed $\epsilon \in (0,1)$, there is no polynomial-time $\frac{1}{n^{1/2-\epsilon}}$-approximation algorithm for the MIS problem unless $\mathbf{P} = \mathbf{NP}$, and no $\frac{1}{n^{1-\epsilon}}$-approximation algorithm unless $\mathbf{ZPP} = \mathbf{NP}$.*

Construction of MOMDP. Let $G = ([n], E)$ be a graph for which we want to find the maximum independent set. Create a fully connected MOMDP M with $|E|$ states $\{s_e\}$, one per edge $e \in E$. There are only two actions $\mathcal{A} = \{a_1, a_2\}$. In state s_e of edge $e = (e_1, e_2)$, performing action a_1 and a_2 correspond to e_1 and e_2 respectively. We create n agents where agent i corresponds to vertex i. The reward function of agent i for state s_e and action a_k for $k \in \{1,2\}$ is defined as

$$R_i(s_e, a_k) = \begin{cases} 1, & \text{if } e_k = i, \\ 0, & \text{o.w.} \end{cases}$$

In other words, the reward functions encode the set of edges incident to vertex i.

Correctness of reduction. A policy π is optimal for agent i iff for all the edges incident to i the action corresponding to i is selected with probability 1. If a policy π is considered optimal by two agents i and i', then $e = (i, i') \notin E$, since at state s_e either $\pi(a_1|s) = 1$ or $\pi(a_2|s) = 1$. Therefore, the set of agents that consider a policy optimal corresponds to an independent set in G. Furthermore, take any independent set $V' \subseteq [n]$. Let π be the policy that for each edge e selects the action corresponding to the vertex in V' and, if no such vertices exist, select one arbitrarily. This policy is well defined since V' is an independent set and there are no edges with both vertices from V'. Policy π is optimal for agents of V' since at each state their favourite action is selected. Thus, we have an equivalence between the maximum independent set of G and the plurality winner policy of M. Therefore, the hardness of approximation follows from Theorem 7. □

D.2 Proof of Theorem 2

Proof of Theorem 2. We reduce the MAX-2SAT problem to finding an α-approval policy for $\alpha \in (7/8, 1]$.

For two Boolean variables x_1 and x_2, we denote the disjunction (i.e., logical or) of two variables by $x_1 \vee x_2$ and the negation of x_1 by $\neg x_1$.

Definition 8 (maximum 2-satisfiability (MAX-2SAT))**.** *Given a set of m Boolean variables $\{x_1, \dots x_m\}$ and a set of n 2-literal disjunction clauses $\{C_1, \dots, C_n\}$, the goal of the* maximum 2-satifiability problem (MAX-2SAT) *is to find the maximum number of clauses that can be satisfied together by an assignment $\phi : \{x_j\}_{j\in[n]} \to \{\text{True}, \text{False}\}$.*

Garey et al. [11] showed that the MAX-2SAT problem is NP-hard.

Construction of MOMDP. For an instance of the MAX-2SAT problem, let $\{C_1, \dots, C_n\}$ be a set of m 2-literal disjunction clauses over n variables $\{x_1, \dots x_m\}$. We create a fully connected MOMDP M with m states — state s_j representing variable x_j. There are only two actions, a_{True} and a_{False} which at state s_j correspond to setting variable x_j to True and False respectively. This way, a policy $\pi(a_{\text{True}}|s_j)$ can be interpreted as the probability of setting the variable x_j to True, subject to $\pi(a_{\text{True}}|s_j) = 1 - \pi(a_{\text{False}}|s_j)$.

Agents and reward function. We introduce some notation before introducing the agents and their reward function. Take a clause $C_i = c_{i,1} \vee c_{i,2}$ where $c_{i,k} \in \{x_j, \neg x_j\}_{j\in[m]}$ for $k \in [2]$. By combining the former relation with the mapping of x_j and $\neg x_j$ to state-action pairs (s_j, a_{True}) and (s_j, a_{False}) respectively, we define $\lambda : \{c_{i,1}, c_{i,2}\} \to \mathcal{S} \times \mathcal{A}$. For every $c_{i,k}$, $\lambda(c_{i,k})$ is the state-action pair that evaluates $c_{i,k}$ to True when selected with probability one. Now, we are ready to introduce the agents. For each clause C_i, we create 3 agents per clause as follows:

- Agent $\text{ag}_{i,1}$ with a reward function that is 1 for $\lambda(c_{i,1})$ and $\lambda(c_{i,2})$ and zero otherwise. An optimal policy of this agent selects the two state-action pairs with probability one, which can be interpreted as $c_{i,1} \leftarrow \text{True}$ and $c_{i,2} \leftarrow \text{True}$ that implies $C_i = \text{True}$.
- Agent $\text{ag}_{i,2}$ with a reward function that is 1 for $\lambda(c_{i,1})$ and $\lambda(\neg c_{i,2})$ — note the negation. This is another assignment of variables, $c_{i,1} \leftarrow \text{True}$ and $c_{i,2} \leftarrow \text{False}$, that implies $C_i \leftarrow \text{True}$.
- Similarly, the reward function of $\text{ag}_{i,3}$ is 1 for $\lambda(\neg c_{i,1})$ and $\lambda(c_{i,2})$ (which again implies $C_i \leftarrow \text{True}$).

For ease of exposition, and by a slight abuse of notation, in the rest of the proof we use π to refer to the occupancy measure. From Equation (1) we have $|\mathcal{S}| \cdot d_\pi(s, a) = \pi(a|s)$ and the α-approval rule is invariant to affine transformation. Therefore, we let $J_i(\pi) = \langle \pi, R \rangle$.

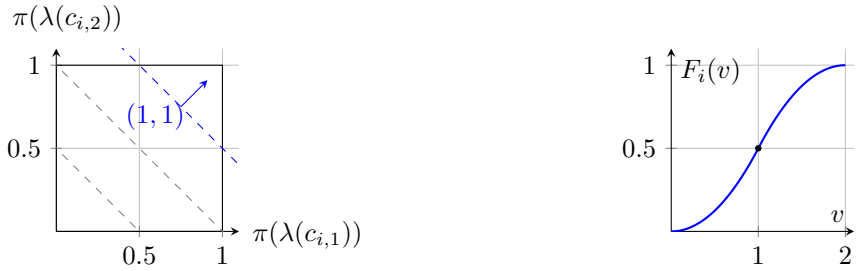
Expected return distribution. The expected return of agent $\text{ag}_{i,1}$ for a policy π is $\pi(\lambda(c_{i,1})) + \pi(\lambda(c_{i,2}))$. For conciseness, let $v_1 = \pi(\lambda(c_{i,1})), v_2 = \pi(\lambda(c_{i,2}))$. Then, the cdf is

$$F_{\text{ag}_{i,1}}(v) = \int_{v_1=0}^{v} \int_{v_2=0}^{v} \mathbb{I}[v_1 + v_2 \leqslant v] \cdot dv_1\, dv_2 = \begin{cases} \frac{v^2}{2} & \text{for } v \in [0, 1], \\ 1 - \frac{(2-v)^2}{2} & \text{for } v \in [1, 2]. \end{cases}$$

See Figure 1 for a visualization. The cdf of all other agents have the same form. For each of the $3n$ agents above, their maximum expected return is 2.

Rounding a policy. Observe that $F_i(3/2) = 7/8$. For agent $\text{ag}_{i,1}$ to α-approve a policy for $\alpha \in (7/8, 1]$, they require a utility (strictly) more than $3/2 = F^{-1}_{\text{ag}_{i,1}}(7/8)$. The fact that $v_1, v_2 \in [0, 1]$ in addition to $v_1 + v_2 > 3/2$, implies that $v_1 > 1/2$ and $v_2 > 1/2$. Therefore, if a policy π is α-approved by agent $\text{ag}_{i,1}$, we have $\pi(\lambda(c_{i,1})) > 1/2$ and $\pi(\lambda(c_{i,2})) > 1/2$. Further, observe that for at most one of the three agents $\{\text{ag}_{i,k}\}_{k\in[3]}$ the condition of $J_{\text{ag}_{i,k}}(\pi) > 3/2$ may hold as every pair of the agent disagree on one literal.

We round such a policy π to an assignment ϕ by letting $\phi(x_j) \leftarrow \text{True}$ if $\pi(a_{\text{True}}|s_j) > \frac{1}{2}$ and $\phi(x_j) \leftarrow \text{False}$ otherwise. If an agent in $\{\text{ag}_{i,k}\}_{k\in[3]}$ α-approves π, then C_i is satisfied by the



(a) The scaled occupancy polytope

(b) The expected return cdf

Figure 1: The effective state-action occupancy polytope of agents and their expected return distribution.

assignment ϕ. Therefore, we have that $\mathrm{OPT}_\alpha \leqslant \mathrm{OPT}_{\mathrm{MAX-2SAT}}$, where OPT_α is the maximum feasible number of α-approvals among all policies and $\mathrm{OPT}_{\mathrm{MAX-2SAT}}$ is the maximum number of clauses that can be satisfied among all assignments.

Next, we show $\mathrm{OPT}_{\mathrm{MAX-2SAT}} \leqslant \mathrm{OPT}_\alpha$ by deriving a policy π that gets $\mathrm{OPT}_{\mathrm{MAX-2SAT}}$ α-approvals based on the optimal assignment for $\mathrm{OPT}_{\mathrm{MAX-2SAT}}$ by simply letting $\pi(a_{\mathrm{True}}|s_j) = 1$ iff $\phi(x_i) = \mathrm{True}$. If ϕ satisfies a clause C_i, then for the agent $\mathrm{ag} \in \{\mathrm{ag}_{i,k}\}_{k\in[3]}$ that matches the literal assignments of ϕ for C_i, we have $J_{\mathrm{ag}}(\pi) = 2$, which is an optimal, i.e., 1-approval, policy for ag. For the other two agents for C_i, $J_{\mathrm{ag}}(\pi) = 1$ which gets a return less than their required utility of $3/2$ for an α-approval. Therefore, we have that $\mathrm{OPT}_{\mathrm{MAX-2SAT}} = \mathrm{OPT}_\alpha$ and the hardness of computation follows from our reduction. $\square$

E Experiments

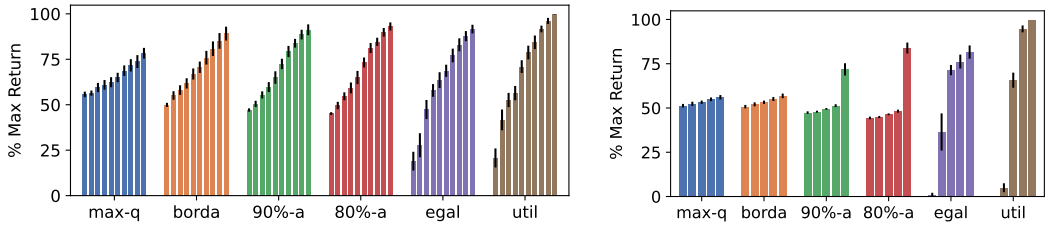
Environment. We adapt the dynamic attention allocation environment introduced by D'Amour et al. [7]. We aim to monitor several sites and prevent potential incidents, but limited resources prevent us from monitoring all sites at all times; this is inspired by applications such as food inspection and pest control [2]. There are $m = 5$ warehouses and each can be in 3 different stages: normal (norm), risky (risk) and incident (inc). There are $|\mathcal{S}| = 3^m$ states containing all possible stages of all warehouses. In each step, we can monitor at most one site, so there are $m + 1$ actions, where action $m + 1$ is no operation and action $i \leqslant m$ is monitoring warehouse i. There are n agents; each agent i has a list ℓ_i of warehouses that they consider valuable and a reward function R_i. In each step t, $R_i(s_t, a_t) = -\mathbb{I}[a_t \leqslant m] - \sum_{j \in \ell_i} \rho_i w_j \cdot \mathbb{I}[s_{t,j} = \mathrm{inc} \wedge a_t \neq j]$, where $w_j \in \{100, 150, \cdots, 250\}$ denotes the penalty of an incident occurring in warehouse j, ρ_i is the scale of penalties for agent i which is sampled from $\{0.25, 0.5, \cdots, n\}$, and -1 is the cost of monitoring. In each step, if we monitor warehouse j, its stage becomes normal. If not, it changes from norm to risk and from risk to inc with probabilities $p_{j,\mathrm{risk}}$ and $p_{j,\mathrm{inc}}$, and stays the same otherwise. Probabilities are sampled i.i.d. uniformly from $[0.5, 0.8]$. The state transitions $\mathcal{P}$ is the product of the warehouses' stage transitions.

Rules. We compare the outcomes of policy aggregation with different rules: *max-quantile, Borda, α-approval* ($\alpha = 0.9, 0.8$), *egalitarian* (maximize minimum return) and *utilitarian* (maximize sum of returns). We sample $5 \cdot 10^5$ random policies based on which we fit a generalized logistic function to estimate the cdf of the expected return distribution F_i (Definition 3) for every agent. The policies for α-approval voting rules are optimized with respect to maximum utilitarian welfare. The egalitarian rule finds a policy that maximizes the expected return of the worst-off agent, then optimizes for the second worst-off agent, and so on. The implementation details of Borda count are in **??**.

Results. In Figure 2, we report the normalized expected return of agents as $\frac{J_i(\pi) - \min_\pi J_i(\pi)}{\max_{\pi'} J_i(\pi') - \min_{\pi''} J_i(\pi'')}$ (sorted from lowest to highest) which are averaged over 10 different environment and agents instances. We observe that the utilitarian and egalitarian rules are sensitive to the different agents' reward scales and tend to perform unfairly. The utilitarian rule achieves the highest utilitarian welfare by almost ignoring one agent. The egalitarian rule achieves higher return for the worst-off agents compared to the utilitarian rule, but still yields an inequitable outcome. The max-quantile rule tends to return the fairest outcomes with similar normalized returns for the agents. The Borda rule, while not a fair rule by design, tends to find fair outcomes which are slightly worse than the max-quantile rule. The α-approval rule with max utilitarian completion tends to the utilitarian rule as $\alpha \to 0$ and to plurality as $\alpha \to 1$. Importantly, although not shown in the plots, the plurality rule ignores almost all agents and performs optimally for a randomly selected agent.

In addition to the fine-grained utility distributions, in Table 1, we report two aggregate measures based on agents' utilities: (i) the Gini index, a statistical measure of dispersion defined as $\frac{\sum_{i \in N} \sum_{j \in N} |J_i(\pi) - J_j(\pi)|}{2n \sum_{i \in N} J_i(\pi)}$ — where a lower Gini index indicates a more equitable distribution, and (ii) the Nash welfare, defined as the geometric mean of agents' utilities $\left(\prod_{i \in N} J_i(\pi)\right)^{1/n}$ — where a higher Nash welfare is preferable. We observe a similar trend as above, where utilitarian and egalitarian rules perform worse across both metrics. For the other four rules, the Nash welfare scores are comparable in both scenarios, with Borda showing slightly better performance. The Gini index, however, highlights a clearer distinction among the rules, with max-quantile performing better.

[2] The code for the experiments is available at https://github.com/praal/policy-aggregation.



(a) 10 agents consider a random subset of warehouses valuable

(b) 5 symmetric agents, one per warehouse

Figure 2: Comparison of policies optimized by different rules in two different scenarios based on the normalized expected return for agents. The bars, grouped by rule, correspond to agents sorted based on their normalized expected return. The error bars show the standard error of the mean.

	5 symmetric agents, one per warehouse		10 agents, random subsets of warehouses	
Rules	**Gini index**	**Nash welfare**	**Gini index**	**Nash welfare**
egalitarian	0.2864 ± 0.0295	0.2208 ± 0.0717	0.2126 ± 0.0209	0.4655 ± 0.0581
utilitarian	0.4392 ± 0.0094	0.0502 ± 0.0174	0.2020 ± 0.0182	0.5736 ± 0.0471
80%-approvals	0.1233 ± 0.0047	0.5186 ± 0.0051	0.1352 ± 0.0037	0.6741 ± 0.0200
90%-approvals	0.0793 ± 0.0056	0.5286 ± 0.0053	0.1257 ± 0.0034	0.6746 ± 0.0211
Borda	0.0225 ± 0.0024	0.5356 ± 0.0062	0.1029 ± 0.0083	0.6801 ± 0.0261
max-quantile	0.0188 ± 0.0022	0.5355 ± 0.0062	0.0625 ± 0.0067	0.6474 ± 0.0232

Table 1: Comparison of policies optimized by different rules in two scenarios based on Gini index and Nash welfare based on their normalized expected return averaged. We report the mean and the standard error.