The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

Moral Machine or Tyranny of the Majority?

Michael Feffer, Hoda Heidari*, Zachary C. Lipton*

Carnegie Mellon University
{mfeffer,hheidari,zlipton}@andrew.cmu.edu

Abstract

With Artificial Intelligence systems increasingly applied in consequential domains, researchers have begun to ask how these systems ought to act in ethically charged situations where even humans lack consensus. In the Moral Machine project, researchers crowdsourced answers to "Trolley Problems" concerning autonomous vehicles. Subsequently, Noothigattu et al. (2018) proposed inferring linear functions that approximate each individual's preferences and aggregating these linear models by averaging parameters across the population. In this paper, we examine this *averaging* mechanism, focusing on fairness concerns in the presence of strategic effects. We investigate a simple setting where the population consists of two groups, with the minority constituting an $\alpha < 0.5$ share of the population. To simplify the analysis, we consider the extreme case in which within-group preferences are homogeneous. Focusing on the fraction of contested cases where the minority group prevails, we make the following observations: (a) even when all parties report their preferences truthfully, the fraction of disputes where the minority prevails is less than proportionate in α; (b) the degree of sub-proportionality grows more severe as the level of disagreement between the groups increases; (c) when parties report preferences strategically, pure strategy equilibria do not always exist; and (d) whenever a pure strategy equilibrium exists, the majority group prevails 100% of the time. These findings raise concerns about stability and fairness of preference vector averaging as a mechanism for aggregating diverging voices. Finally, we discuss alternatives, including randomized dictatorship and median-based mechanisms.

Introduction

Machine learning (ML) has increasingly been employed to automate decisions in consequential domains including autonomous vehicles, healthcare, hiring, finance, and criminal justice. These domains present many ethically charged decisions, where even knowledgeable humans may lack consensus about the right course of action. Consequently, AI researchers have been forced to consider how to resolve normative disputes when competing values come into conflict. The formal study of such ethical quandaries long predates the advent of modern AI systems. For example, philosophers have long debated Trolley Problems (Thomson 1985), which generally take the form of inescapable decisions among normatively undesirable alternatives. Notably, these problems typically lack clear-cut answers, and people's judgments are often sensitive to subtle details in the provided context.

Questions about whose values are represented and what objectives are optimized in ML-based systems have become especially salient in light of documented instances of algorithmic bias in deployed systems (Angwin et al. 2016; Buolamwini and Gebru 2018; Obermeyer et al. 2019). Faced with questions of whose values ought to prevail when a judgment must be made, some AI ethics researchers have advocated *participatory machine learning*, a family of methods for democratizing decisions by incorporating the views of a variety of stakeholders (Lee et al. 2019; Ilvento 2020; Jung et al. 2021). For example, a line of papers on *preference elicitation* tasks stakeholders with choosing among sets of alternatives (Ilvento 2020; Lee et al. 2019; Jung et al. 2021; Hiranandani, Narasimhan, and Koyejo 2020; Hiranandani et al. 2020; Freedman et al. 2020). In many of these studies, the hope is to compute a socially aligned objective function (Lee et al. 2019; Freedman et al. 2020) or fairness metric (Ilvento 2020; Hiranandani et al. 2020; Hiranandani, Narasimhan, and Koyejo 2020).

In a pioneering study, Awad et al. (2018) introduced the Moral Machine, a large-scale crowdsourcing study in which millions of participants from around the world were presented with autonomous driving scenarios in the style of Trolley Problems. Participants were shown images depicting two possible outcomes and asked which alternative they preferred. In one scenario, the first alternative might be to collide with a barrier and sacrifice several young passengers, and the second to pulverize two elderly pedestrians in a crosswalk. Utilizing the Moral Machine dataset, Noothigattu et al. (2018) proposed methods for inferring each participant's preferences. Specifically, they represent each alternative by a fixed-length vector of attributes, $\mathbf{x}$, and each participant's scoring function as the dot product between their *preference vector* $\boldsymbol{\theta}$ and the alternative $\mathbf{x}$. The objective is to infer parameters $\boldsymbol{\theta}$ so that whenever an individual prefers one alternative over another, the preferred alternative receives a higher score. To aggregate these preference vectors across a population, Noothigattu et al. (2018) propose to simply *average* them. Faced with

*These authors contributed equally.

a "moral dilemma", an autonomous vehicle would conceivably featurize each alternative, compute their dot products with the *aggregate preference vector*, and choose the highest scoring option. These approaches have been echoed in a growing body of follow-up studies (Pugnetti and Schläpfer 2018; Wang et al. 2019; Kemmer et al. 2020). However, despite the work's influence, key properties of the proposed mechanism remain under-explored.

In this paper, we analyze the mechanism of averaging preference vectors across a population, focusing on stability, strategic effects, and implications for fairness. While averaging mechanisms have been extensively studied (Hurley and Lior 2002; Renault and Trannoy 2005; Marchese and Montefiori 2011; Renault and Trannoy 2011), Noothigattu et al. (2018)'s approach warrants analysis for several reasons: (i) here, individuals vote on the parameters of a ranking algorithm, rather than directly on the outcomes of interest and (ii) since only the ordering induced by the preference vectors matters, only the direction of the aggregate vector is relevant. We introduce a stylized model in which the population of interest consists of a (disadvantaged) minority group, constituting an $\alpha < 0.5$ share of the population, and an (advantaged) majority group constituting $1 - \alpha$. The goal is to determine how an autonomous vehicle should behave by polling the population via Noothigattu et al. (2018)'s algorithm, where the problem setup is identical to that of the Moral Machine (Awad et al. 2018). Because we only observe which alternative an individual prefers, our analysis focuses on the fraction of cases (among *disputed cases*) where each group prevails.

To clarify the fairness properties of the mechanism, we concentrate our analysis on an stylized setting in which within-group preferences are homogeneous. We emphasize this assumption is only meant for simplifying the analysis and revealing fundamental limitations of the simple preference vector averaging mechanism. The problems we identify here do not simply disappear in more complicated settings with within-group variation in preferences. Moreover, to isolate the role of the aggregation mechanism, we assume participants can directly report their preference vectors. With these assumptions in place, our analysis makes the following key observations: (i) even when preferences are reported truthfully, the fraction of cases where the minority prevails is sub-proportionate (i.e., less than α); (ii) the degree of sub-proportionality grows more severe when the divergence between the two groups' preference vectors is large; (iii) as with most averaging-based approaches, this mechanism is not strategy-proof; (iv) whenever a pure strategy equilibrium exists, the majority group prevails on 100% of cases; and (v) last but not least, while other stable and incentive-compatible mechanisms do exist (e.g., the randomized dictatorship model), they come with other fundamental shortcomings.

Several takeaways flow from our analysis. First, this averaging of preference vectors is qualitatively different from averaging votes directly on outcomes of interest, giving rise to instability and surprising strategic behavior. Second, the degree of compromise among majority and minority demographics, both under truthful and strategic settings, is an important consideration when designing aggregation mechanisms to support participatory ML systems. Finally, our work raises critical questions regarding the limitations of simple voting methods to ensure stakeholders' participation in the design of high-stakes automated decision-making systems. We hope that this work encourages the community to reflect on the importance of addressing normative disagreements among stakeholders—not simply by passing them through an aggregation mechanism, but by effectively giving voice to the disadvantaged communities and facilitating deliberations necessary to reach an acceptable outcome for all stakeholders.

Related Work

Our work draws on the preference elicitation and computational social choice literatures. We build most directly on a line of work consisting of the Moral Machine (Awad et al. 2018) and subsequently proposed procedures for inferring and aggregating preferences (Noothigattu et al. 2018). These studies inspired numerous follow-up articles, such as Pugnetti and Schläpfer (2018), who pose the same questions to Swiss vehicle customers; Wang et al. (2019), who modify the algorithms to support differential privacy; and Kemmer et al. (2020), who evaluate various methods of aggregating crowdsourced judgments, including averaging.

Preference elicitation for participatory ML Lee et al. (2019) helped nonprofit volunteers build ML algorithms via pairwise comparisons and aggregating the resulting preferences using Borda Count voting. Johnston, Blessenohl, and Vayanos (2020) adopted participatory mechanisms to determine how to allocate COVID-19 triage supplies. Freedman et al. (2020) modified parameter weights in their kidney exchange linear programs to help break ties based on inferred participant preferences about who should receive kidney donations. Still other works that learn fairness metrics from user input assume a single participant or a group capable of coming to consensus, and therefore employ no aggregation mechanism (Ilvento 2020; Hiranandani et al. 2020; Hiranandani, Narasimhan, and Koyejo 2020). Note that, per Chamberlin (1985), that using Borda Count instead of averaging does not necessarily alleviate issues related to strategic voting. Similarly, a minimax group regret approach as employed in Johnston, Blessenohl, and Vayanos (2020) may also be vulnerable to subgroup strategic voting.

Computational Social Choice Like us, El-Mhamdi et al. (2021) highlights the general susceptibility of averaging-based methods to strategic voting and discusses alternative median-based mechanisms. While they mention the Moral Machine as an example, they do not analyze the particular mechanism presented in (Noothigattu et al. 2018) or provide any of the insights about fairness and strategic concerns presented in our work. Moulin (1980); Conitzer et al. (2016); Zhang, Cheng, and Conitzer (2019) propose median-based algorithms for voting in a strategy-proof manner in the context of crowdsourcing societal tradeoffs. Brill and Conitzer (2015) address strategic voting, noting that median-based approaches are less susceptible to manipulation. Conitzer, Brill,

and Freeman (2015) describes issues with crowdsourcing societal tradeoffs more generally. Concerning randomized dictatorship models, Gibbard (1973, 1977) introduce randomized solutions as "unattractive" yet strategy-proof approaches. Zeckhauser (1973) similarly states that randomized dictatorship has some favorable characteristics in that it forces voters to report their true preferences (i.e., is strategy-proof) and is *probabilistically linear* (i.e., switching votes from one alternative to another only affects the selection probabilities of those two alternatives in a linear fashion). Instead of proposing other aggregation mechanisms, both Landemore and Page (2015) and Pierson (2017) argue that deliberation and discussion among participants are procedures that, when used in conjunction with voting, can lead to better outcomes than voting as a standalone process.

Our Problem Setup

Consider a population consisting of two groups: A, the majority, and D, a minority constituting $\alpha < 0.5$ fraction of the population. Each group is characterized by a true *preference vector* $\boldsymbol{\theta}_i^* \in \mathbb{R}^d$ that determines the preferences of all members of that group over outcomes/alternatives. (In the autonomous vehicle example, the outcome could be the individuals chosen to be saved in face of an unavoidable accident.) Each alternative is represented by a feature vector $\mathbf{x} \in \mathbb{R}^d$. We assume alternatives are drawn independently from M, a spherically symmetric distribution centered at the origin with radius 1. Members of group i prefer alternative $\mathbf{x} \in \mathbb{R}^d$ to alternative $\mathbf{y} \in \mathbb{R}^d$ whenever $\boldsymbol{\theta}_i^*.\mathbf{x} \geq \boldsymbol{\theta}_i^*.\mathbf{y}$, i.e., whenever $\boldsymbol{\theta}_i^*.(\mathbf{x} - \mathbf{y}) \geq 0$. So the true preference vector $\boldsymbol{\theta}_i^* \in \mathbb{R}^d$ of group $i \in \{A, D\}$ allows us to calculate their rankings over any set of alternatives. See the Appendix of our paper's extended version for a concrete example of such vectors. Throughout and unless otherwise specified, we assume $\boldsymbol{\theta}_A^* \neq \boldsymbol{\theta}_D^*$ (e.g., in the context of the Moral Machine experiment, $\boldsymbol{\theta}_A^*$ may reflect group A's preference for sacrificing pedestrians when an autonomous vehicle is faced with an accident, and $\boldsymbol{\theta}_D^*$ reflects group D's preference for sacrificing passengers).

Note that in enforcing within-group homogeneity of preferences in our model, we do not assert that groups are homogeneous in the real world. Rather, our aim is to elucidate whether this mechanism respects the preferences of minority groups. In other words, the homogeneity assumption allows us to make a clear analytic point rather than state a realistic or normatively desirable situation. Such simplifying assumptions are common in economics and theoretical computer science literature (e.g., Costinot and Kartik (2007); Krishna and Morgan (2012)).

Formulation as a game We consider a two-player normal-form game $G = (S_A, u_A), (S_D, u_D)$ in which groups in the setup described above correspond to players, A and D (so player A and D have true preference vectors $\boldsymbol{\theta}_A^*$ and $\boldsymbol{\theta}_D^*$, respectively). S_i denotes player i's strategy space, and u_i their payoff/utility. More precisely, each player $i \in \{A, D\}$ strategically reports a preference vector $\boldsymbol{\theta}_i \in S_i$ (which may be different from $\boldsymbol{\theta}_i^*$), where S_i consists of all d-dimensional vectors with Euclidean norm equal to 1. The payoff function for player i, $u_i : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$, is a function where $u_i(\boldsymbol{\theta}_A, \boldsymbol{\theta}_D)$ indicates the payoff of player $i \in \{A, D\}$ when players A and D report preference vectors $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_D$, respectively. We assume the payoff function for each player $i \in \{A, D\}$ captures the proximity of their true preference vector $\boldsymbol{\theta}_i^*$ to the aggregate vector, $\boldsymbol{\theta}_C$, obtained by a *simple averaging* mechanism: $\boldsymbol{\theta}_C = (\alpha\boldsymbol{\theta}_D + (1-\alpha)\boldsymbol{\theta}_A)/\|\alpha\boldsymbol{\theta}_D + (1-\alpha)\boldsymbol{\theta}_A\|$.[1]

Each player aims to maximize the fraction of decisions that agree with their true preferences. As shown in Proposition 1, achieving this goal requires that each player $i \in \{A, D\}$ reports a preference vector such that the resulting aggregate vector $\boldsymbol{\theta}_C$ is closest, as measured by cosine similarity, to their true preference vector $\boldsymbol{\theta}_i^*$. To show this formally, we first define the notion of *agreement* between two preference vectors.

Definition 1. *Consider two preference vectors $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j \in \mathbb{R}^d$ and a spherically symmetric distribution M over the space of all alternatives (i.e., $\mathbb{R}^d$). The* level of agreement *between $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, denoted by $\rho(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$, is the probability, over draws of pairs of alternatives from M, such that $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ rank the alternatives in the same order:*

$$\rho(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \mathbb{P}_{\mathbf{x}\sim M, \mathbf{y}\sim M}\left[\operatorname{sign}(\boldsymbol{\theta}_i \cdot (\mathbf{y} - \mathbf{x})) = \operatorname{sign}(\boldsymbol{\theta}_j \cdot (\mathbf{y} - \mathbf{x}))\right].$$

Proposition 1. *Suppose alternatives are sampled i.i.d. from a spherically symmetric distribution M defined over $\mathbb{R}^d$. Then for any two preference vectors $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j \in \mathbb{R}^d$,*

$$\rho(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \frac{\pi - \cos^{-1}(\boldsymbol{\theta}_i \cdot \boldsymbol{\theta}_j)}{\pi}.$$

Proof. Note that the preference of a player i over alternatives $\mathbf{x}$ and $\mathbf{y}$ depends only on the sign of $\boldsymbol{\theta}_i \cdot (\mathbf{y} - \mathbf{x})$. Because $\mathbf{x}$ and $\mathbf{y}$ are drawn independently from a spherically symmetrical distribution M, we can see, by symmetry, that the difference vector, $\mathbf{y} - \mathbf{x}$, can points in any direction from $[0, 2\pi]$ with equal (uniform) probability.

Any preference vector $\boldsymbol{\theta}$ defines two half-spaces over the vector $\mathbf{y} - \mathbf{x} \in \mathbb{R}^d$. By H^+, we denote the half-space in which $\boldsymbol{\theta} \cdot (\mathbf{y} - \mathbf{x}) > 0$ (and thus $\mathbf{y}$ is preferred to $\mathbf{x}$) and by H^-, we denote the half-space in which $\boldsymbol{\theta} \cdot (\mathbf{y} - \mathbf{x}) < 0$ (and thus $\mathbf{x}$ is preferred to $\mathbf{y}$). Because the event where $\mathbf{x} = \mathbf{y}$ has 0 measure, tie-breaking conventions will not impact our analysis.

Note that for any given player, the line separating H^+ from H^- is perpendicular to $\boldsymbol{\theta}$ and passes through the origin. Disagreements among $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ correspond to pairs of alternatives such that either $\mathbf{x} - \mathbf{y} \in \{H_j^- \cap H_i^+\}$ or $\mathbf{x} - \mathbf{y} \in \{H_i^- \cap H_j^+\}$. So we have:

$$\begin{aligned}\rho(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) &= \mathbb{P}_{\mathbf{x}\sim M, \mathbf{y}\sim M}\left[\operatorname{sign}(\boldsymbol{\theta}_i \cdot (\mathbf{y} - \mathbf{x})) = \operatorname{sign}(\boldsymbol{\theta}_j \cdot (\mathbf{y} - \mathbf{x}))\right], \\ &= 1 - \mathbb{P}_{\mathbf{x}\sim M, \mathbf{y}\sim M}\left[\mathbf{x} - \mathbf{y} \in H_j^- \cap H_i^+ \text{ or } \mathbf{x} - \mathbf{y} \in H_j^+ \cap H_i^-\right].\end{aligned}$$

Note that both $\{H_j^- \cap H_i^+\}$ and $\{H_i^- \cap H_j^+\}$ are cones whose vertices lie at the origin and whose vertex angles are each

[1] While only the direction of $\boldsymbol{\theta}_C$ matters, we normalize its length for mathematical convenience.

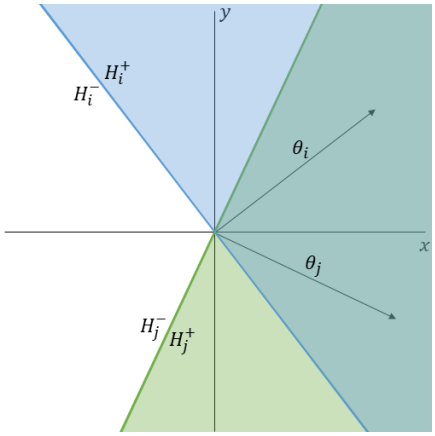

Figure 1: Half-spaces $H_i^+, H_i^-, H_j^+, H_j^-$ and intersections for preference vectors $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$.

given by $\measuredangle(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$. See Figure 1 for an illustration of these half-spaces and their intersections.

Because the alternatives, $\mathbf{x}, \mathbf{y}$ are drawn i.i.d. from a spherically symmetric distribution,

$$\mathbb{P}\left[\mathbf{x}-\mathbf{y} \in \{H_j^- \cap H_i^+\}\right] = \mathbb{P}\left[\mathbf{x}-\mathbf{y} \in \{H_i^- \cap H_j^+\}\right] = \frac{\cos^{-1}(\boldsymbol{\theta}_i \cdot \boldsymbol{\theta}_j)}{2\pi}.$$

Moreover, because these regions of disagreement are disjoint, the total probability of disagreement $\boldsymbol{\theta}_i$, $\boldsymbol{\theta}_j$ is given by the sum of the probabilities of $\mathbf{y}-\mathbf{x}$ lying in either region of disagreement, $\cos^{-1}(\boldsymbol{\theta}_i \cdot \boldsymbol{\theta}_j))/2\pi$. Thus, the level of agreement is given by $\frac{\pi - \cos^{-1}(\boldsymbol{\theta}_i \cdot \boldsymbol{\theta}_j)}{\pi}$. ■

Because the level of agreement of the aggregate decisions with player i is monotonic in the cosine similarity between their true preferences $\boldsymbol{\theta}_i^*$ and the aggregate vector $\boldsymbol{\theta}_C$, we can equivalently take the cosine similarity as the payoff of interest. More precisely, we can define u_A, u_D as follows: for all $\boldsymbol{\theta}_A \in S_A$ and $\boldsymbol{\theta}_D \in S_D$,

$$u_A(\boldsymbol{\theta}_A, \boldsymbol{\theta}_D) = \boldsymbol{\theta}_C \cdot \boldsymbol{\theta}_A^*,$$
$$u_D(\boldsymbol{\theta}_A, \boldsymbol{\theta}_D) = \boldsymbol{\theta}_C \cdot \boldsymbol{\theta}_D^*.$$

We are interested in understanding pure strategy Nash equilibria of the above game. To define pure strategy Nash equilibria precisely, we need to first define the concept *best responses* to a given pure strategy. We say a pure strategy $\boldsymbol{\theta}_i \in S_i$ is a *best response* to $\boldsymbol{\theta}_{-i} \in S_{-i}$ in G if for all $\hat{\boldsymbol{\theta}}_i \in S_i$,

$$u_i(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{-i}) \geq u_i(\hat{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_{-i}).$$

Given a strategy $\boldsymbol{\theta}_{-i}$ for player $-i$, we will use the notation $BR_i(\boldsymbol{\theta}_{-i})$ to refer to the set of all pure best responses of i to $\boldsymbol{\theta}_{-i}$.

Definition 2 (Nash Equilibrium)**.** *A strategy profile $(\boldsymbol{\theta}'_A, \boldsymbol{\theta}'_D)$ is a pure Nash Equilibrium for G if $\boldsymbol{\theta}'_A$ is a best-response to $\boldsymbol{\theta}'_D$ and vice versa.*

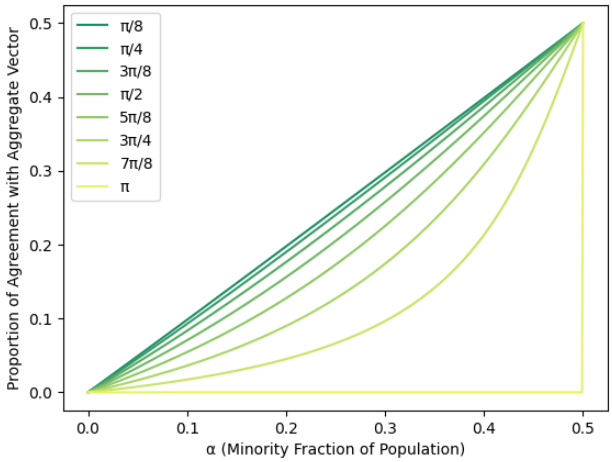

Figure 2: As α increases, the probability that the aggregate agrees with the minority group increases, but this relationship is sub-proportional in α and depends on $\measuredangle(\boldsymbol{\theta}_D^*, \boldsymbol{\theta}_A^*)$.

From here on, we restrict our analysis to the case where $d = 2$. As we argue in the extended version of our paper, it may suffice to consider the 2-dimensional plane containing the origin, $\boldsymbol{\theta}_A^*$, and $\boldsymbol{\theta}_D^*$.

Findings

We are now ready to introduce our primary findings.

Disproportionate majority voice via the averaging mechanisms Assume that all participants truthfully report their preferences. Even here, the averaging mechanism has some strange properties. Notably, focusing on the fraction of all disputed cases where group D prevails, specifically

$$\mathbb{P}[\text{sign}(\boldsymbol{\theta}_C \cdot (\mathbf{y}-\mathbf{x})) = \text{sign}(\boldsymbol{\theta}_D^* \cdot (\mathbf{y}-\mathbf{x})) | \text{sign}(\boldsymbol{\theta}_A^* \cdot (\mathbf{y}-\mathbf{x})) \neq \text{sign}(\boldsymbol{\theta}_D^* \cdot (\mathbf{y}-\mathbf{x}))] \quad (1)$$

we find that this is sub-proportional in α (the pattern follows a sigmoid-like shape). Moreover, the degree of sub-proportionality depends on the angle between the true preference vectors because even though $\boldsymbol{\theta}_C$ is a weighted sum of $\boldsymbol{\theta}_A^*$ and $\boldsymbol{\theta}_D^*$, the direction of $\boldsymbol{\theta}_C$ (and in turn the levels of agreement with the minority and majority groups) depends on the directions of $\boldsymbol{\theta}_A^*$ and $\boldsymbol{\theta}_D^*$ in addition to α. This can be seen clearly in Figure 2 where the probability of group D prevailing is computed via Equation 1 under varying settings of their population share α and for various levels of agreement between the two groups (see the extended version of our paper for more details). Note that sub-proportionality becomes more extreme as the angle between the true preference vectors increases.

Majority Group Can Create Aggregate Vector In Any Direction Next, we address the setting where groups can report their preferences strategically. First, we find that for any preference vector $\boldsymbol{\theta}_D$ reported by the minority group, the majority group can always choose some preference vector to report such that the aggregate vector is identical to their true preferences $\boldsymbol{\theta}_C = \boldsymbol{\theta}_A^*$. This implies that if a pure strategy Nash equilibrium exists, group A always gets their way.

(In our running example, this implies that the autonomous vehicle always operates in accordance with group A's preferences).

Lemma 1. *For any fixed vector $\boldsymbol{\theta}_D$ played by the minority group, the majority group can report a vector $\boldsymbol{\theta}_A$, such that $\boldsymbol{\theta}_C = \boldsymbol{\theta}^*_A$.*

Proof sketch. In order to ensure $\boldsymbol{\theta}^*_A = \boldsymbol{\theta}_C$, player A must report $\boldsymbol{\theta}_A$ such that $\frac{\alpha\boldsymbol{\theta}_D+(1-\alpha)\boldsymbol{\theta}_A}{\|\alpha\boldsymbol{\theta}_D+(1-\alpha)\boldsymbol{\theta}_A\|} = \boldsymbol{\theta}^*_A$. For any vector $\boldsymbol{\theta}_D$, it is easy to see that the above equation is equivalent to:

$$\boldsymbol{\theta}_A = \frac{\left[\alpha(\boldsymbol{\theta}_D \cdot \boldsymbol{\theta}^*_A) + \sqrt{\alpha^2(\boldsymbol{\theta}_D \cdot \boldsymbol{\theta}^*_A)^2 - 2\alpha + 1}\right]\boldsymbol{\theta}^*_A - \alpha\boldsymbol{\theta}_D}{(1-\alpha)}. \tag{2}$$

(See the extended version of our paper for the full derivation.) ■

Conditions for Pure Strategy Nash Equilibrium

In this section, we present a necessary condition for the existence of a pure Nash Equilibrium in the above game. First, we derive the maximum amount by which the minority group can pull the aggregate vector based on their relative population size by reporting their preference strategically.

Lemma 2. *Consider a fixed reported majority group vector $\boldsymbol{\theta}_A$. Suppose the minority group reports $\boldsymbol{\theta}_D$ to yield an aggregate vector $\boldsymbol{\theta}_C$. Then for any $\boldsymbol{\theta}_D \in S_D$,*

$$\measuredangle(\boldsymbol{\theta}_C, \boldsymbol{\theta}_A) \leq \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right). \tag{3}$$

The equality occurs if and only if $\boldsymbol{\theta}_D$ is orthogonal to $\boldsymbol{\theta}_C$.

Proof Sketch. It is easy to see that $\boldsymbol{\theta}_D$ yields an aggregate vector $\boldsymbol{\theta}_C$ such that $\measuredangle(\boldsymbol{\theta}_C, \boldsymbol{\theta}_A) = \sin^{-1}\left(\frac{\alpha \sin(\measuredangle(\boldsymbol{\theta}_C, \boldsymbol{\theta}_D))}{1-\alpha}\right)$. Note that because $\sin(\measuredangle(\boldsymbol{\theta}_C, \boldsymbol{\theta}_D)) \leq 1$, $\left(\frac{\alpha \sin(\measuredangle(\boldsymbol{\theta}_C, \boldsymbol{\theta}_D))}{1-\alpha}\right) \leq \left(\frac{\alpha}{1-\alpha}\right)$. Moreover, $\sin^{-1}(.)$ is a monotonic function in $[-1, 1]$. Therefore, $\measuredangle(\boldsymbol{\theta}_C, \boldsymbol{\theta}_A) = \sin^{-1}\left(\frac{\alpha \sin(\measuredangle(\boldsymbol{\theta}_C, \boldsymbol{\theta}_D))}{1-\alpha}\right) \leq \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$. Additionally, $\measuredangle(\boldsymbol{\theta}_C, \boldsymbol{\theta}_A) = \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$ if and only if $\sin(\measuredangle(\boldsymbol{\theta}_C, \boldsymbol{\theta}_D)) = 1$. This in turn happens if and only if $\boldsymbol{\theta}_C \perp \boldsymbol{\theta}_D$.

(See the extended version of our paper for the full derivation.) ■

These two lemmas yield the subsequent three lemmas, which are proved in our paper's extended version:

First, if a pure strategy Nash equilibrium exists, the majority group can always report its preference vector such that the aggregate matches their true preferences.

Lemma 3. *Consider the game, G, described in the previous section. If $(\boldsymbol{\theta}'_D, \boldsymbol{\theta}'_A)$ is a pure strategy Nash equilibrium for G, then $\boldsymbol{\theta}_C = \boldsymbol{\theta}^*_A$.*

Second, we derive an upper bound on the angle between the aggregate and minority group's true preferences.

Lemma 4. *For any $\boldsymbol{\theta}_A \in S_A$, there exists $\boldsymbol{\theta}_D \in S_D$ such that $\measuredangle(\boldsymbol{\theta}_C, \boldsymbol{\theta}^*_D) \leq \pi - \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$.*

Lastly, in a pure strategy Nash equilibrium, the minority group's best response is orthogonal to the aggregate.

Lemma 5. *In any pure strategy Nash equilibrium defined by best responses $\boldsymbol{\theta}'_A$ and $\boldsymbol{\theta}'_D$, $\boldsymbol{\theta}'_D \perp \boldsymbol{\theta}_C$, and $\measuredangle(\boldsymbol{\theta}_C, \boldsymbol{\theta}'_A) = \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$.*

Now, we show that a pure strategy Nash equilibrium does not always exist.

Theorem 1. *For G to have a pure strategy Nash equilibrium, it must be the case that*

$$\measuredangle(\boldsymbol{\theta}^*_A, \boldsymbol{\theta}^*_D) < \pi - \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right).$$

Proof. Suppose, via contradiction, there exists a pure strategy Nash equilibrium and $\measuredangle(\boldsymbol{\theta}^*_A, \boldsymbol{\theta}^*_D) \geq \pi - \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$. Two cases are possible:

1. $\measuredangle(\boldsymbol{\theta}^*_A, \boldsymbol{\theta}^*_D) > \pi - \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$;
2. $\measuredangle(\boldsymbol{\theta}^*_A, \boldsymbol{\theta}^*_D) = \pi - \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$.

First, consider the case where $\measuredangle(\boldsymbol{\theta}^*_A, \boldsymbol{\theta}^*_D) > \pi - \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$. By Lemma 3, $\boldsymbol{\theta}^*_A = \boldsymbol{\theta}_C$ in this equilibrium. However, by Lemma 4, $\measuredangle(\boldsymbol{\theta}_C, \boldsymbol{\theta}^*_D) \leq \pi - \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$ given any response from the majority group. This means the minority group will report a best response $\boldsymbol{\theta}'_D$ that does not yield $\boldsymbol{\theta}^*_A$. This in turn raises a contradiction and indicates that such an equilibrium cannot exist if $\measuredangle(\boldsymbol{\theta}^*_A, \boldsymbol{\theta}^*_D) > \pi - \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$.

Second, consider the case where $\measuredangle(\boldsymbol{\theta}^*_A, \boldsymbol{\theta}^*_D) = \pi - \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$. By Lemma 3, $\boldsymbol{\theta}^*_A = \boldsymbol{\theta}_C$ in this equilibrium. Because $\boldsymbol{\theta}^*_A = \boldsymbol{\theta}_C$, by Lemma 5, $\measuredangle(\boldsymbol{\theta}'_A, \boldsymbol{\theta}^*_A) = (\boldsymbol{\theta}'_A, \boldsymbol{\theta}_C) = \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$. Therefore $\measuredangle(\boldsymbol{\theta}'_A, \boldsymbol{\theta}^*_D) = \measuredangle(\boldsymbol{\theta}'_A, \boldsymbol{\theta}^*_A) + \measuredangle(\boldsymbol{\theta}^*_A, \boldsymbol{\theta}^*_D) = \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right) + \pi - \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right) = \pi$, and $\boldsymbol{\theta}'_A$ is diametrically opposed to $\boldsymbol{\theta}^*_D$. Note that there are two best responses for the minority group, both of which pull the aggregate (one by $\sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$, the other by $-\sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$) towards $\boldsymbol{\theta}^*_D$ to yield $\measuredangle(\boldsymbol{\theta}_C, \boldsymbol{\theta}^*_D) = \pi - \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$. However, only *one* yields $\boldsymbol{\theta}_C = \boldsymbol{\theta}^*_A$. This is a contradiction because $\boldsymbol{\theta}_C$ does not necessarily match $\boldsymbol{\theta}^*_A$. Therefore, this is not a pure strategy Nash equilibrium.

Thus, for G to have a pure strategy Nash equilibrium, $\measuredangle(\boldsymbol{\theta}^*_A, \boldsymbol{\theta}^*_D) < \pi - \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$. ■

Theorem 1 reveals that equilibrium does not exist if (i) the groups are close to diametric opposition (in terms of their preference vectors), and (ii) the groups are close in size.

Form of Pure Strategy Nash Equilibrium

Whenever necessary conditions outlined in Theorem 1 are met, the equilibrium takes a certain form. Theorem 2 specifies this form exactly and proves the conditions are also sufficient. We denote $R = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ as the $\frac{\pi}{2}$ radians rotation matrix.

Theorem 2. *If $\measuredangle(\boldsymbol{\theta}^*_A, \boldsymbol{\theta}^*_D) < \pi - \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$, then there exists a pure strategy Nash equilibrium. Specifically, $(\boldsymbol{\theta}'_A, \boldsymbol{\theta}'_D)$ form a pure strategy Nash equilibrium iff:*

1. *$\boldsymbol{\theta}'_A$ and $\boldsymbol{\theta}'_D$ take the following form:*

$$\boldsymbol{\theta}'_A = \frac{\left[(\sqrt{1-2\alpha})I - \alpha \text{sign}(\boldsymbol{\theta}^*_D \cdot R\boldsymbol{\theta}^*_A)R\right]\boldsymbol{\theta}^*_A}{1-\alpha},$$
$$\boldsymbol{\theta}'_D = \text{sign}(\boldsymbol{\theta}^*_D \cdot R\boldsymbol{\theta}^*_A)R\boldsymbol{\theta}^*_A.$$

2. *$\measuredangle(\boldsymbol{\theta}'_A, \boldsymbol{\theta}'_D) = \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right) + \frac{\pi}{2}$. That is, players' equilibrium strategies will always point in opposing directions.*

Proof. Because $\measuredangle(\boldsymbol{\theta}^*_A, \boldsymbol{\theta}^*_D) < \pi - \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$, the minority group will always report the unique vector $\boldsymbol{\theta}'_D$ orthogonal to $\boldsymbol{\theta}_C$ (as per Lemma 5) that maximizes agreement between $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}^*_D$. By Lemma 3, $\boldsymbol{\theta}_C = \boldsymbol{\theta}^*_A$ at equilibrium, meaning that $\boldsymbol{\theta}'_D$ is orthogonal to $\boldsymbol{\theta}^*_A$. $\boldsymbol{\theta}'_D$ is therefore either $R\boldsymbol{\theta}^*_A$ or $-R\boldsymbol{\theta}^*_A$.

If $\boldsymbol{\theta}^*_D \cdot R\boldsymbol{\theta}^*_A > 0$, then the minority group should report $\boldsymbol{\theta}'_D = R\boldsymbol{\theta}^*_A$ to maximize their utility. Otherwise, the minority group should report $\boldsymbol{\theta}'_D = -R\boldsymbol{\theta}^*_A$. In either case, $\boldsymbol{\theta}'_D = \text{sign}(\boldsymbol{\theta}^*_D \cdot R\boldsymbol{\theta}^*_A)R\boldsymbol{\theta}^*_A$. (Note that $\text{sign}(\boldsymbol{\theta}^*_D \cdot R\theta^*_A) = 0$ if and only if the two groups are in total agreement or are diametrically opposed. The proof preconditions that the groups disagree and that $\measuredangle(\boldsymbol{\theta}^*_A, \boldsymbol{\theta}^*_D) < \pi - \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right)$ prevent this result from occurring.)

Using Equation 2 with $\boldsymbol{\theta}'_D$ and $\boldsymbol{\theta}^*_A$ yields $\boldsymbol{\theta}'_A$ and $\boldsymbol{\theta}'_D$ in item 1. Our paper's extended version shows the steps involved in this process. Moreover, item 2 follows from Lemma 5 because $\measuredangle(\boldsymbol{\theta}'_A, \boldsymbol{\theta}'_D) = \measuredangle(\boldsymbol{\theta}'_A, \boldsymbol{\theta}'_C) + \measuredangle(\boldsymbol{\theta}'_C, \boldsymbol{\theta}'_D) = \sin^{-1}\left(\frac{\alpha}{1-\alpha}\right) + \frac{\pi}{2}$. See Figure 3 for an illustration of all items. ■

Discussion

Our analysis of the averaging mechanism shows the following: (a) even when groups are truthful, concessions to the minority group are less than proportional to their share of the population; (b) this sub-proportionality depends on the cosine similarity between the true preference vectors; and (c) if participants respond strategically, tyranny of the majority results whenever an equilibrium exists. We now build on this foundation by exploring: (i) how our results obtain even without intragroup collusion; (ii) the benefits and pitfalls of other aggregation mechanisms, (iii) more general implications for computational social choice and participatory ML algorithms.

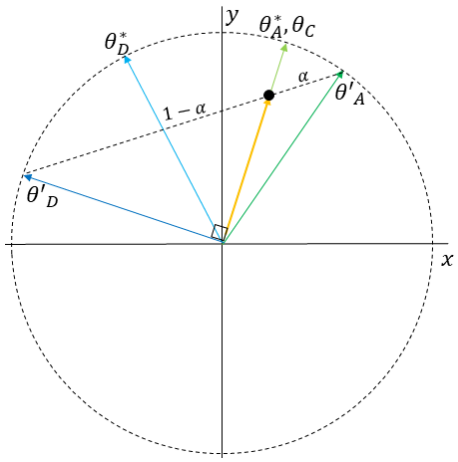

Figure 3: Example pure strategy Nash equilibrium with best responses $\boldsymbol{\theta}'_A$ and $\boldsymbol{\theta}'_D$ given minority population fraction α, and true preference vectors $\boldsymbol{\theta}^*_A$, and $\boldsymbol{\theta}^*_D$. Note that $\boldsymbol{\theta}'_D \perp \boldsymbol{\theta}_C$ and the aggregate vector $\boldsymbol{\theta}_C$ points in the same direction as $\boldsymbol{\theta}^*_A$.

No intra-group collusion required for results. All of our analysis up to this point utilized a setup in which each group acts as a monolithic player. However, we now informally argue that collusion is *not* required. Namely, as long as individuals are aware of the aggregate vector's direction and the relative sizes of each group, no collusion is necessary.

Proof Sketch. Suppose colluding and not colluding result in different aggregate vectors. Decomposing each aggregate vector into contributions from the minority and majority group members (i.e. for a given group, summing the individual preference vectors of members from that group and dividing by the number of members) will result in different contributing vectors from each group. The contributing vectors that stem from collusion are guaranteed to maximize the utility of each group. However, this means at least one of the contributing vectors in the other set (resulting from independent contributions) does *not* maximize the utility of its corresponding group because it does not match the vector for that same group via collusion. However, this raises a contradiction because maximizing group utility involves maximizing each group member's individual utility, regardless of whether collusion is involved. If maximal group utility is not achieved, at least one member is not maximizing their own utility. This cannot possibly happen because we assume that all individuals are rational actors seeking to maximize their utility, so this is an impossible case. Therefore, no collusion is required to obtain the results described in previous subsections. ■

Median-based approaches: geometric medians and

coordinate-wise medians. El-Mhamdi et al. (2021) mentions that average-based aggregation mechanisms are susceptible to manipulation even by an individual. In response, they describe two types of median-based aggregation approaches that generalize to d dimensions: geometric medians and coordinate-wise medians. Given a set of preference vectors, the geometric median is a vector that minimizes the Euclidean distance to each vector in the set, and the coordinate-wise median is a vector whose ith component is the one-dimensional median computed from the ith components of all provided vectors. The main result of El-Mhamdi et al. (2021) is that the geometric median is not generally strategy-proof, and they point to existing work (Sui and Boutilier 2015) that proves the coordinate-wise median is strategy-proof but not group strategy-proof. In our setup, the coordinate-wise median is both incentive-compatible and stable, but it does not alleviate the dominance of the majority's voice. In fact, this mechanism might be considered *worse*, because even when both groups report preferences truthfully, the majority group always prevails.[2]

Randomized Dictatorship. In contrast to median-based approaches, the method of randomized dictatorship solves all problems related to averaging aggregation. In this method, the preferences of an individual selected at random from the population are used directly as the aggregated result. As reported in Zeckhauser (1973), it is strategy-proof and *probabilistically linear*. In the setting we consider, participants would not be incentivized to lie (because any participant's reported preferences could be applied to everyone), and concessions to the minority group are proportional instead of sub-proportional (because the minority group preferences will be selected with probability α and the majority group preferences will be selected with probability $1 - \alpha$). Thus, the randomized dictatorship mechanism is both incentive-compatible and proportional.

In the economics literature (Gibbard 1973, 1977), researchers have contemplated such mechanisms but have argued that they are "unattractive" despite being strategy-proof because they "[leave] too much to chance" and ignore input from all individuals except the one selected at random. Zeckhauser (1973) also posits that using the preferences of one over many may not be appropriate when dealing with "momentous social decisions". Deciding the ethics of autonomous vehicles may be one such decision. Conitzer, Brill, and Freeman (2015) additionally notes that one may be more confident in their colleagues' preferences than those of a random member of the population, so it may not meet requirements of procedural fairness despite giving proportional voice to the minority in expectation.

Additional Considerations. Zeckhauser (1973) proves that "No voting system that relies on individuals' self-interested balloting can guarantee a nondictatorial outcome that is Pareto-optimal." In their conclusion, they note that while this is a pessimistic result, "perhaps we should not ask despairingly, Where do we go from here?, but rather inquire, Have we been employing the appropriate mind-set all along?" In other words, the correct approach may not be to build a "one-size-fits-all" strategy-proof voting scheme but rather something that holistically considers all preferences regarding a given social issue and surrounding context. Conitzer, Brill, and Freeman (2015) notes that participation requires context and locality, both of which are lost when we crowdsource moral dilemmas (as in Awad et al. (2018)). Pugnetti and Schläpfer (2018) echoes this message; they find that Swiss residents had different preferences about autonomous vehicle ethics relative to those of other countries. Additionally, Conitzer, Brill, and Freeman (2015) calls attention to the importance of the featurization process of the alternatives. If these alternatives are not represented properly, the elicitation and aggregation processes cannot hope to arrive at a result that accurately reflects participants' true judgments. Moreover, Landemore and Page (2015) and Pierson (2017) suggest deliberation and discussion used alongside voting can lead to better overall outcomes than voting alone.

Conclusion. The central impulse of participatory machine learning is to integrate input from various stakeholders directly into the process of developing machine learning systems. For all its promise, participatory ML also raises challenging questions about the precise form that such an integration should take. Our work highlights some of the challenges of designing such mechanisms, especially when individuals may hold radically different values and act strategically. While we draw heavily on the previous literature in economics and computational social science, our work also reveals that some of the ways of combining preferences that seem natural from a machine learning perspective can be unstable and lead to strange strategic behavior in ways that do not map so neatly onto known analyses. Surprisingly, while preference elicitation has gained considerable attention in the participatory ML literature, few papers address cases in which stakeholders hold genuinely conflicting values. While some part of this work going forward will surely be to study such mechanisms formally, we also stress that better mechanism design is no panacea for reconciling conflicting values in the real world. Beyond theoretical analysis, participatory ML systems may also require channels by which communication, debate, and reconciliation of competing values could potentially take place.

Acknowledgements

Authors acknowledge support from NSF (IIS2040929) and PwC (through the Digital Transformation and Innovation Center at CMU). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the National Science Foundation and other funding agencies. The first author was additionally supported by a GEM Fellowship and an ARCS Scholarship.

[2]It may also possible to transform the two-dimensional problem on a circle into a one-dimensional problem on a line, where the single-peaked preferences result of Moulin (1980) can be applied.

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.

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

Brill, M.; and Conitzer, V. 2015. Strategic voting and strategic candidacy. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM Conference on Fairness, Accountability and Transparency (FAccT)*.

Chamberlin, J. R. 1985. An investigation into the relative manipulability of four voting systems. *Behavioral Science*, 30(4): 195–203.

Conitzer, V.; Brill, M.; and Freeman, R. 2015. Crowdsourcing Societal Tradeoffs. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Conitzer, V.; Freeman, R.; Brill, M.; and Li, Y. 2016. Rules for choosing societal tradeoffs. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Costinot, A.; and Kartik, N. 2007. On optimal voting rules under homogeneous preferences. *V Manuscript, Department of Economics, University of California San Diego*.

El-Mhamdi, E.-M.; Farhadkhani, S.; Guerraoui, R.; and Hoang, L.-N. 2021. On the strategyproofness of the geometric median. *arXiv preprint arXiv:2106.02394*.

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

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

Gibbard, A. 1977. Manipulation of schemes that mix voting with chance. *Econometrica: Journal of the Econometric Society*, 665–681.

Hiranandani, G.; Mathur, J.; Narasimhan, H.; and Koyejo, O. 2020. Quadratic Metric Elicitation for Fairness and Beyond. *arXiv preprint arXiv:2011.01516*.

Hiranandani, G.; Narasimhan, H.; and Koyejo, S. 2020. Fair performance metric elicitation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hurley, W.; and Lior, D. 2002. Combining expert judgment: On the performance of trimmed mean vote aggregation procedures in the presence of strategic voting. *European Journal of Operational Research*, 140(1): 142–147.

Ilvento, C. 2020. Metric learning for individual fairness. In *Foundations of Responsible Computing (FORC)*.

Johnston, C. M.; Blessenohl, S.; and Vayanos, P. 2020. Preference Elicitation and Aggregation to Aid with Patient Triage during the COVID-19 Pandemic. In *ICML Workshop on Participatory Approaches to Machine Learning*.

Jung, C.; Kearns, M.; Neel, S.; Roth, A.; Stapleton, L.; and Wu, Z. S. 2021. An algorithmic framework for fairness elicitation. In *Foundations of Responsible Computing (FORC)*.

Kemmer, R.; Yoo, Y.; Escobedo, A.; and Maciejewski, R. 2020. Enhancing collective estimates by aggregating cardinal and ordinal inputs. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Krishna, V.; and Morgan, J. 2012. Voluntary voting: Costs and benefits. *Journal of Economic Theory*, 147(6): 2083–2123.

Landemore, H.; and Page, S. E. 2015. Deliberation and disagreement: Problem solving, prediction, and positive dissensus. *Politics, philosophy & economics*, 14(3): 229–254.

Lee, M. K.; Kusbit, D.; Kahng, A.; Kim, J. T.; Yuan, X.; Chan, A.; See, D.; Noothigattu, R.; Lee, S.; Psomas, A.; et al. 2019. WeBuildAI: Participatory framework for algorithmic governance. In *ACM Conference on Computer- Supported Cooperative Work And Social Computing (CSCW)*.

Marchese, C.; and Montefiori, M. 2011. Strategy versus sincerity in mean voting. *Journal of Economic Psychology*, 32(1): 93–102.

Moulin, H. 1980. On strategy-proofness and single peakedness. *Public Choice*, 35(4): 437–455.

Noothigattu, R.; Gaikwad, S.; Awad, E.; Dsouza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. 2018. A voting-based system for ethical decision making. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Pierson, E. 2017. Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124*.

Pugnetti, C.; and Schläpfer, R. 2018. Customer preferences and implicit tradeoffs in accident scenarios for self-driving vehicle algorithms. *Journal of Risk and Financial Management*, 11(2): 28.

Renault, R.; and Trannoy, A. 2005. Protecting minorities through the average voting rule. *Journal of Public Economic Theory*, 7(2): 169–199.

Renault, R.; and Trannoy, A. 2011. Assessing the extent of strategic manipulation: the average vote example. *SERIEs*, 2(4): 497–513.

Sui, X.; and Boutilier, C. 2015. Approximately Strategyproof Mechanisms for (Constrained) Facility Location. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Thomson, J. J. 1985. The trolley problem. *The Yale Law Journal*, 94(6): 1395–1415.

Wang, T.; Zhao, J.; Yu, H.; Liu, J.; Yang, X.; Ren, X.; and Shi, S. 2019. Privacy-preserving crowd-guided AI decision-making in ethical dilemmas. In *ACM International Conference on Information and Knowledge Management (CIKM)*.

Zeckhauser, R. 1973. Voting systems, honest preferences and Pareto optimality. *American Political Science Review*, 67(3): 934–946.

Zhang, H.; Cheng, Y.; and Conitzer, V. 2019. A better algorithm for societal tradeoffs. In *Association for the Advancement of Artificial Intelligence (AAAI)*.