Journal of Data Analytics and Intelligent Decision-making

Online ISSN: 3115-8161

Journal Homepage: jdaid.qom.ac.ir

University of Qom

Proportional Representation in Artificial Intelligence: Clustering, Alignment, and Beyond

Mehdi Fazli[1]✉, and **Saeed Alipour**[2]

1. Corresponding Author, Assistant Prof, Department of Mathematics, Ard.C., Islamic Azad University, Ardabil, Iran. Email: mehdi.fazli.s@gmail.com
2. Assistant Prof, Department of Accounting, Ard.C., Islamic Azad University, Ardabil, Iran. Email: saeed.alipour@iau.ac.ir

Article Info

Article type:
Research Article

Article history:
Received 14 February 2026
Received in revised form 7 April 2026
Accepted 5 June 2026
Available online 30 June 2026

Keywords:
Proportional representation, Artificial intelligence, Computational social choice, Clustering, AI alignment.

ABSTRACT

Proportional representation is a foundational concept in social choice theory, seeking to ensure that the preferences of distinct groups are reflected fairly in collective decisions. As algorithmic systems increasingly shape high-stakes decisions in society, there is a growing need for principled methods that enable artificial intelligence (AI) to account for heterogeneous human values and preferences. This article explores how proportional representation can be extended beyond its classical role in voting and elections to address key challenges in modern AI. We focus on two central domains. First, we examine clustering, when data points naturally represent individuals or agents with diverse characteristics or preferences. We review recent advances that reinterpret clustering as a representation problem, introduce formal notions as a representation problem for both centroid-based and non-centroid-based clustering, and highlight algorithmic guarantees ensuring that large, cohesive groups receive influence proportional to their size. Second, we consider AI alignment, particularly reinforcement learning from human feedback (RLHF) in the presence of heterogeneous preferences. We argue that learning a single global reward function is fundamentally insufficient to capture population-level diversity and may violate basic social choice principles. To address this, we present a framework based on committees of reward functions, designed so that pairwise preferences induced by the committee proportionally reflect those of human annotators. We discuss theoretical guarantees showing that small committees suffice to achieve low proportionality error, as well as empirical evidence demonstrating substantial improvements over any single deterministic reward model.

Cite this article: Fazli, M., Alipour, S (2026). Proportional Representation in Artificial Intelligence: Clustering, Alignment, and Beyond. *Journal of Data Analytics and Intelligent Decision-Making (JDAID)*, *2*(2), 31-47. https://doi.org/10.22091/jdaid.2026.2605.1051

© Author(s) retain the copyright.

Publisher: University of Qom.
DOI: https://doi.org/10.22091/jdaid.2026.2605.1051

Introduction

Artificial intelligence systems are increasingly shaping decisions that affect millions of people, from content recommendations and hiring algorithms to medical diagnosis and autonomous systems. In these contexts, the diversity of human preferences and values poses a fundamental challenge: how can AI systems make decisions that fairly represent a heterogeneous population? Traditional approaches often optimize a single global objective, which may overlook minority groups, reinforce existing biases, or fail to capture nuanced social preferences (Anyanwu et al., 2025; Mahpouya et al., 2026).

Proportional representation, a principle long studied in social choice and political science, offers a promising solution. By ensuring that the influence of any group is roughly proportional to its size, it provides a framework for fairness that respects population diversity. While this principle has been primarily applied in voting and elections, recent research suggests its relevance to AI tasks, particularly in clustering and alignment, where algorithms must aggregate individual preferences or characteristics into collective outcomes (Amiri et al., 2023; Mansouri et al., 2024).

In clustering, proportional representation allows algorithms to form groups that reflect the underlying population structure rather than being dominated by dense regions or outliers. In AI alignment, especially in reinforcement learning from human feedback (RLHF), it enables the design of reward models that more faithfully capture heterogeneous human judgments, avoiding the pitfalls of a single global objective.

This article explores how proportional representation can be systematically integrated into AI design. We review formal definitions and algorithmic guarantees in clustering, introduce committee-based approaches for RLHF, and discuss potential extensions to other AI domains such as federated learning, recommender systems, and large-scale model evaluation. By bridging social choice theory and AI, we aim to provide principled methods for building systems that respect population diversity while remaining computationally tractable.

The more general problem of aggregating diverse preferences and making collective decisions has been studied for centuries in the field of social choice theory (Arrow et al., 1964), providing a rich foundation for addressing conflicts and trade-offs in group decision-making. Rooted in economics, political science, and mathematics, social choice theory offers formal tools and axiomatic frameworks for a variety of settings such as voting rules for electing candidates or choosing public policies even when preferences are only partially known, when participants may act strategically, and when other practical constraints are present. Building on this foundation, the field of computational social choice (Peters, 2024) has emerged over the past few decades, bringing computational thinking into social choice theory by combining ideas from theoretical computer science and artificial intelligence. This research extends classical questions through an algorithmic and complexity-theoretic lens, and more broadly applies computational techniques to formalize, analyze, and redesign social choice frameworks. Work in computational social choice is now a regular presence at major AI conferences, often featured in dedicated tracks—such as Social Choice and Voting, the very kind of session our hypothetical Italian-food-loving grad student might have wandered into (Pezeshgi et al., 2025; Safizadeh et al., 2026).

Over the past ten years, the computational social choice community has increasingly focused on the principle of proportional representation. This principle dictates that each group of agents should influence the outcome in proportion to its size. Put differently, a group constituting x% of the total population should have the ability to determine x% of the outcome. Much of this research has concentrated on the committee selection scenario, where the objective is to choose k candidates as winners of an election while accounting for voters' diverse preferences (Gambhir, 2026; Sahoo et al., 2025). A significant practical application of this research has

emerged in participatory budgeting, a civic engagement process in which community members decide how to allocate public funds across proposed projects (Peters & Skowron, 2020; Peters et al., 2021). In this context, proportional representation ensures that cohesive voter groups, with strong preferences for particular projects, receive funding roughly proportional to their size, rather than being consistently overruled by a majority with different priorities. This limitation is commonly observed before the implementation of algorithms designed to enforce proportional representation (Fairstein et al., 2023; Pezeshgi et al., 2026).

However, the idea of proportional representation can be valuable far beyond classical voting and election scenarios. As algorithms play an increasingly central role in decisions that shape our daily lives, aligning them with societal values and ethical principles—meaning designing systems whose behavior reliably reflects human goals and ethical norms—has become both urgent and essential. This need has been further highlighted by recent breakthroughs in AI, particularly by the emergence of large language models (LLMs), which now perform tasks ranging from content generation and code synthesis to applications with higher societal impact, such as offering student feedback (Malik et al., 2021) and providing mental health support (Hua et al., 2025; Mukherjee et al., 2026). However, societal values are far from uniform and often in conflict. As a result, even identifying which values an AI system should reflect poses a fundamental challenge that requires some form of collective decision-making. With AI systems are increasingly deployed in high-stakes domains, it is crucial to develop principled methods that articulate precisely and mathematically how a broad spectrum of perspectives should be reflected across applications.

As the importance of eliciting and integrating human preferences into AI systems becomes increasingly evident, researchers have highlighted the critical role such efforts play in ensuring that large language models (LLMs) behave in ways consistent with human expectations. A key challenge in this context is the aggregation of potentially conflicting viewpoints, which has emerged as a central concern in the alignment of AI systems with diverse human values. These considerations have, in turn, emphasized the relevance of social choice theory and preference aggregation methods in AI research.

In a broadcast by Google DeepMind, titled *"AI Safety… ok doomed: with Anca Dragan"* (Dragan, 2024), Anca Dragan, director of AI safety and alignment at DeepMind, addressed the complexity of value alignment in a world where individuals hold heterogeneous preferences:

> *How do you value alignment when you acknowledge that there are different people in the world with different values? I don't claim to have the answers here, but there are areas of science, research, and economics that have thought about this. If you think about voting, there's this notion of social choice of preference aggregation, and the whole idea is that we have to make a decision that's acceptable for multiple people who might have competing objectives. So, there are a few things you could start doing.*

Dragan's remarks highlight that AI alignment cannot rely solely on a single, uniform definition of "optimal" behavior. Instead, it requires methods capable of systematically considering the preferences of multiple stakeholders, many of which may be in tension. Social choice theory offers a well-established framework for such scenarios, providing tools for preference aggregation that have been studied extensively in economics, political science, and decision theory. By applying these principles to AI, researchers can begin to design systems that respect the distribution of human values rather than privileging the majority or a single metric of utility.

Despite the remarkable success of modern AI systems, a fundamental challenge remains unresolved: how can an AI system make decisions that fairly reflect the preferences of a heterogeneous population? Most existing AI methods rely on optimizing a single objective

function, selecting a single model, or producing a single collective outcome. While such approaches often achieve high aggregate performance, they may systematically underrepresent minority viewpoints, overlook important subpopulations, and fail to capture the diversity of human values. This limitation becomes particularly pronounced in applications such as clustering, recommender systems, federated learning, and reinforcement learning from human feedback, where decisions are inherently based on aggregating information from many individuals. Consequently, there is a growing need for principled frameworks that ensure different groups exert influence on algorithmic outcomes in proportion to their presence in the population. Proportional representation provides a natural candidate for addressing this challenge by offering mathematically grounded guarantees that diverse groups are fairly reflected in collective decisions.

The primary goal of this article is to illustrate how the principle of proportional representation has been applied to two key AI challenges: (a) Clustering, where each data point can naturally be interpreted as representing an individual or agent with distinct preferences or attributes, and (b) AI alignment, specifically in designing reward functions for reinforcement learning from human feedback (RLHF) in the presence of heterogeneous preferences. In concluding section, we also discuss additional potential applications where proportional representation could offer significant benefits.

This article is a survey and perspective-based research. While proportional representation has been studied extensively in computational social choice, its role in artificial intelligence has only recently begun to attract attention across multiple research communities. The novelty of this article lies in providing a unified perspective on these developments. Specifically, we bring together recent advances from clustering, AI alignment, reinforcement learning from human feedback, and related AI applications under the common framework of proportional representation. By identifying shared principles across these domains and highlighting emerging research opportunities, the article offers a broader conceptual understanding of how proportional representation can serve as a general design principle for modern AI systems. Recent years have witnessed a growing interest in incorporating social choice principles into artificial intelligence. Beyond classical applications in voting and collective decision-making, concepts such as proportional representation, fairness, and preference aggregation have recently been studied in reinforcement learning from human feedback, recommender systems, federated learning, evaluation of large language models, and participatory AI systems. Recent works have highlighted the need for AI systems that account for heterogeneous human preferences, rather than optimizing a single global objective, further motivating the integration of computational social choice techniques into modern AI pipelines. By identifying shared principles across these domains and highlighting emerging research opportunities, the article offers a broader conceptual understanding of how proportional representation can serve as a general design principle for modern AI systems

Table 1 summarizes the main strands of research on proportional representation in AI and highlights how the present article differs by providing a unified perspective across clustering, AI alignment, and emerging AI applications.

Table 1. Representative Studies on Proportional Representation in AI

Study	Domain	Main Idea	Representation Guarantee	Limitation
Chen et al. (2019)	Centroid-based clustering	Introduced proportional representation in clustering	Large groups receive at least one representative center	Focused only on centroid-based clustering
Micha & Shah (2020)	Euclidean clustering	Improved approximation guarantees	Constant-factor approximation	Limited to geometric settings
Aziz et al. (2024)	Clustering	Multi-representation guarantees for cohesive groups	Groups receive representation proportional to size	Restricted to clustering applications
Caragiannis, Micha, & Peters (2024)	Non-centroid clustering	Proportional representation without cluster centers	Cohesive groups can form independent clusters	Limited theoretical understanding for some objectives
Ge et al. (2024)	RLHF and AI alignment	Demonstrated failures of single reward models under heterogeneous preferences	Highlights need for preference diversity	Does not provide proportional representation mechanism
Halpern et al. (2026)	RLHF and AI alignment	Committee of reward functions	Preferences represented proportionally through reward ensembles	Focused on pairwise preference aggregation
This Article	Survey and perspective	Unified review of proportional representation across clustering and AI alignment	Identifies common principles and future directions	Highlights open research challenges

Proportional Representation in Committee Selection

We start by providing a short overview of the definitions of proportional representation as developed in the literature on computational social choice theory (Faliszewski et al., 2017) In this context, the objective is to determine a group of winning candidates based on a given set of candidates and the preference profiles of voters. More precisely, a committee selection problem involves n voters (or agents), each expressing preferences over the candidates, with the goal of selecting a committee of predetermined size k. Members of the selected committee can be interpreted as representatives of the electorate. A fundamental principle in the study of committee selection is proportional representation, which requires that the influence of each group of voters on the outcome correspond to its size. Concretely, any group containing at least n/k voters should be entitled to elect one committee member, and more generally, a group of size, at least $\ell \cdot n/k$, for any integer $\ell \geq 1$, should be able to secure ℓ representatives. This principle aims to guarantee that sufficiently large groups of voters with aligned preferences are fairly and proportionally reflected in the final committee. To capture what it means for voters to have aligned or similar preferences, the literature has introduced a variety of formal definitions. Dummett (1984) was among the first to formalize preference similarity through the notion of solid coalitions. Such a coalition consists of voters who unanimously rank a particular set of candidates above all others. In the context of multiline elections, Dummett's proportionality principle stipulates that whenever a coalition is sufficiently large to warrant ℓ representatives, the selected committee should include at least ℓ candidates from that coalition's most preferred set. Subsequent research has shifted attention toward approval-based models, in which agents simply approve subsets of candidates (Lackner & Skowron, 2022). Although the concept of solid coalitions extends to this setting, it often fails to capture stronger forms of proportionality. To address this limitation, Halpern et al. (2026) introduced more robust formal

frameworks based on cohesive groups. Under this approach, a group is deemed ℓ-cohesive if it contains at least $\ell \cdot \mathrm{n}/k$ voters who all approve at least ℓ common candidates.

Building on these foundations, we next examine how analogous ideas have been adapted to clustering and AI alignment, clarifying how proportional representation is defined in these domains and how preference similarity is characterized.

Proportional Representation in Clustering

We now turn from classical voting and committee selection to unsupervised learning, illustrating how proportional representation can be naturally incorporated into clustering when data points correspond to individuals or agents with diverse attributes. Clustering is a core unsupervised learning task that organizes data points into groups so that those within the same cluster exhibit greater similarity than those assigned to different clusters. In most formulations, data points are embedded in a metric space, and similarity is quantified using a distance function, such as Euclidean distance.

Beyond its traditional role in data analysis, clustering is widely employed in settings where data points represent agents. In recommendation systems, for example, users may be grouped based on their interaction histories or stated preferences, enabling more targeted recommendations (Sarwar et al., 2002; Zhang et al., 2016). In federated learning, clients often corresponding to distinct devices or organizations can be clustered according to statistical characteristics of their local datasets, helping to mitigate data heterogeneity and improve training efficiency (Sattler et al., 2020). Clustering is also used in civic engagement platforms, such as Polis (Small et al., 2021), where participants' responses to policy proposals are analyzed to identify groups with shared viewpoints and to inform policymakers about the structure of public opinion.

A widely studied and fundamental paradigm in clustering is centroid-based clustering. In this approach, the algorithm selects a set of k representative points, or centers, from the metric space and assigns each data point to its nearest center. This framework encompasses several classical objectives: (a) k-means clustering, which minimizes the sum of squared distances between points and their assigned centers; (b) k-median clustering, which minimizes the sum of (non-squared) distances; and (c) k-center clustering, which minimizes the maximum distance between any point and its assigned center. These objectives differ in their sensitivity to outliers and the trade-offs they achieve between overall versus worst-case performance. Beyond traditional AI applications, centroid-based clustering also serves as a natural model for problems in social planning and operations research, such as determining the optimal locations of (k) facilities—including hospitals, schools, and service centers—to best serve a population (Farahani & Hekmatfar, 2009; Farahani et al., 2010).

Intuitively, the threshold ·n/k captures the minimum size a group must have to justify influence over one of the k selected representatives or clusters. In a pioneering work, Chen et al. (2019) reinterpreted centroid-based clustering as a committee selection problem. Given the locations of n agents, the task is to choose k centers from the set of feasible cluster centers so that the selected centers adequately represent the agents. They introduced the notion of proportional representation, requiring that every group of at least n / k agents be represented by at least one center. More formally, a clustering is proportionally representative if there does not exist a group of at least (n/k) agents and a feasible center such that every member of the group would strictly prefer that center; that is, each agent is closer to the proposed center than to any of the existing (k) centers. This guarantee applies to all sufficiently large groups (i.e., of size at least n / k)· without assuming any predefined structure on the data.

To illustrate how proportional representation differs from classical clustering objectives, the following example, similar to those presented by Chen et al. (2019), is proposed: Suppose that nearly half of the agents are located at position 0, nearly half at position 1, and the remaining two agents are positioned far from both the main groups and from each other. With $k = 3$, classical objectives, such as k-means or k-center, would typically place one center on each outlier and a single center to cover both large groups, since the groups are relatively close to each other compared to the outliers. In contrast, a proportional representation approach ensures that each of the two major groups, each comprising nearly half of the agents, receives its own center (See Figure 1). This example demonstrates the robustness of proportional representation to outliers and its ability to more accurately capture the underlying structure of the population.

Chen et al. (2024) demonstrated that a proportionally representative solution, as they formally defined it, does not necessarily exist. To address this, they introduced the concept of an α-approximation, a multiplicative relaxation in which not every member of a large group is required to reduce their distance by more than a factor of $\alpha \geq 1$ when moving to a new center. They proved that a solution achieving an approximation factor of $1 + \sqrt{2}$, with respect to proportional representation, consistently exists and can be computed in polynomial time for any metric space. In the case of Euclidean spaces, this bound can be further improved, yielding an approximation factor of 2 (Micha & Shah, 2020).

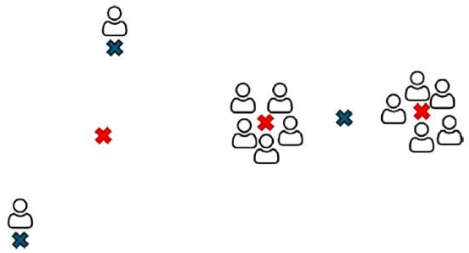


Figure 1. Classical Clustering Objectives (e.g., k-Means, k-Median, k-Center) Allocate Centers to Outliers (Blue), Whereas Proportional Representation Assigns Centers to Large Groups Entitled to Representation (Red).

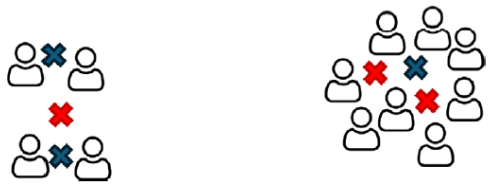


Figure 2. Comparison of Two Notions of Proportional Representation for $n = 12$ and $k = 3$. Chen et al. (2019) Guarantee at Least One Center for Any Group of Size $\ell \cdot n/k$(Blue), Whereas Aziz et al. (2024) Guarantee ℓcenters for Such a Group (Red).

The concept of proportional representation was further revisited and extended by Aziz et al. (2024). In their framework, the guarantee is parameterized by an integer $\ell \geq 1$, requiring that any group of agents of size at least $\ell \cdot n/k$ be represented by at least ℓdistinct centers. Formally, for any subset S of agents with $| S | \geq \ell \cdot n/k$ and maximum pairwise distance r, the definition mandates the existence of at least ℓdifferent centers, each located within distance r of some agent in S. Intuitively, this ensures that large, cohesive groups whose members are all "close" to one another in the metric space receive a number of centers proportional to their size, rather than being constrained to share a single center. In other words, while Chen et al. (2019) required

each group of size $\ell \cdot n/k$ to be assigned at least one center, Aziz et al. (2024) extend this to require at least ℓcenters for such a group, as illustrated in Figure 2. They further proved that clustering solutions, satisfying this form of proportional representation, always exist and can be computed in polynomial time, providing an explicit polynomial-time algorithm. This work has inspired a growing line of research on proportionality in centroid-based clustering (Kalayci et al., 2024; Kellerhals & Peters, 2024; Li et al., 2021).

Although centroid-based clustering is a central paradigm in clustering research, many widely used clustering methods fundamentally depart from this framework. A prominent class of such approaches is non-centroid-based clustering, which partitions agents into $k$groups based on internal similarity rather than distance to a central representative. In these settings, an agent's relevance is determined not by proximity to a center, but by the composition of its group. This distinction is particularly important in applications such as clustered federated learning (Sattler et al., 2020), where agents are grouped to collaboratively train local models. In this context, the quality of a cluster depends on the compatibility of the members' data distributions: each agent benefits from being grouped with others whose data is similar, allowing the shared model to generalize effectively to their local data. Consequently, an agent's performance is dictated by its similarity to other members of the cluster, rather than its distance to a center, underscoring the need for new representational guarantees that extend beyond the centroid-based paradigm.

In the non-centroid-based clustering setting, the goal is to ،assign n agents to k clusters without relying on explicit centers. Unlike centroid-based clustering where utility is measured by proximity to a central representative, what matters here is *who else* an agent is grouped with. An agent's loss or utility depends not on its distance to a center, but on its similarity or dissimilarity to other members of its assigned cluster. In a recent work (Caragiannis, Micha & Peters 2024), we consider the non-centroid setting and aim to ensure that sufficiently large and internally coherent groups are entitled to form their own cluster. Specifically, we adapt the notion of proportionality by requiring that any group of at least n/k data points should be allowed to form a separate cluster, provided that all its members strictly improve their losses by doing so. A natural question is whether a centroid-based clustering solution that is proportionally representative also induces a proportionally representative clustering in the non-centroid-based setting. Every centroid-based clustering naturally defines a non-centroid-based clustering by grouping together the data points assigned to the same center. However, the answer is strongly negative: a solution that is proportionally representative in the centroid-based setting may arbitrarily violate proportional representation in the non-centroid-based setting.

To illustrate, consider a group of (*n/k*) agents located at the same point. In the centroid-based setting, one can place a center at that location and assign additional, arbitrarily distant agents to the same center without affecting the clustering objective. In contrast, such an assignment is undesirable in the non-centroid-based setting. The cohesive group of (*n/k*) co-located agents should form a cluster of its own, without being grouped together with unrelated distant agents. Consequently, a clustering that is proportionally representative under the centroid-based notion need not satisfy proportional representation when clusters are evaluated directly as groups of agents.

In this work, Caragiannis, Micha, and Peters (2024), we study clustering settings in which agents evaluate the quality of their cluster based on their distances to other cluster members. We focus on two loss functions: (i) the maximum distance to any other agent in the same cluster and (ii) the average distance to all other agents in the same cluster. A non-centroid-based clustering is said to be (*alpha*)-approximately proportionally representative if there is no subset of at least (*n/k*) agents that could each improve their loss by a factor of (*alpha*) or more by forming a cluster among themselves.

For the maximum-distance objective, we show that a constant-factor approximately proportionally representative clustering can always be computed efficiently. In contrast, for the average-distance objective, we obtain only an (O(*n/k*))-approximation guarantee. Whether substantially stronger guarantees are possible for this objective remains an open question.

Proportional Representation in AI Alignment and RLHF

In the previous section, we discussed how the principle of proportional representation has been applied to different clustering paradigms. Here, we explore its use in addressing the challenge of aligning AI decisions with the potentially heterogeneous preferences of human users. A standard approach for aligning AI models with societal ethical values is RLHF, which has been applied in domains such as robotics (Bıyık et al., 2024; Kupcsik et al., 2018), recommendation systems (Ailon & Mohri, 2010; Viappiani & Boutilier, 2010), and more recently, finetuning large language models (LLMs) (Ouyang et al., 2022; Stiennon et al., 2020; Ziegler et al., 2019). A typical RLHF pipeline begins by training a reward model on human feedback, often using a pretrained LLM. This reward model is then used to fine-tune the LLM via reinforcement learning. Feedback is collected in various forms, with pairwise comparisons, where annotators indicate which of two responses to the same prompt they prefer, being the most common. Most current RLHF algorithms make the modeling assumption that human preferences follow a random utility model, typically the Bradley Terry (BT) model (Christiano et al., 2017; Ouyang et al., 2022). Under this model, each prompt response pair is assigned a score by an unknown reward function, and the probability of one response being preferred over another by a user that increases with the difference in their scores. Crucially, this reward function is assumed to be shared across all annotators. The reward model is trained to approximate this underlying unknown function by maximizing the likelihood of the observed preferences.

Despite the widespread adoption of this approach, in a recent work (Ge et al., 2024), we show a surprising failure of it in the presence of heterogeneous preferences; that is, when different annotators do not share a common reward function, but rather each holds their own belief, modeled by a separate individual reward function. Specifically, under the widely adopted theoretical assumption that reward functions are linear, an assumption motivated by computational tractability, interpretability, and favorable generalization guarantees (Ge et al., 2024; Zhu et al., 2023; Zhong et al., 2024), the resulting learned reward can violate one of the most fundamental principles in social choice theory, known as Pareto optimality. That is, it may assign a higher score to a candidate response B than to candidate response A, even when all human annotators unanimously prefer A over B. This result highlights the importance of explicitly accounting for heterogeneous preferences when designing reward functions. Social choice theory offers a rich and principled foundation for this goal, as evidenced by a rapidly growing body of work that advocates incorporating social choice principles into RLHF to better reflect variation in human preferences (Chakraborty et al., 2024; Conitzer et al., 2024; Dai & Fleisig, 2024; Ge et al., 2024; Mishra, 2023; Park et al., 2024; Siththaranjan et al., 2024; Swamy et al., 2024; Zhong et al., 2024).

However, designing a single reward function makes it impossible to adequately capture the complexity of population preferences (Dumoulin et al., 2024; Park et al., 2024). For example, if 60% of users prefer $a \succ b \succ c$ and 40% prefer $c \succ b \succ a$, the model will likely adopt the popular ranking, effectively ignoring the preferences of the remaining users. Such effects have been observed in current RLHF systems (Khalifa et al., 2021; Kirk et al., 2024b; Lake et al., 2025; Perez et al., 2022; Poddar et al., 2024; Shypula et al., 2025). One increasingly popular approach to address this limitation is personalization, in which distinct reward functions are tailored to individual users. However, this strategy also introduces significant societal risks, as

some users may interact with LLMs in ways that conflict with broadly accepted social norms or ethical views (Anwar et al., 2024; Kirk et al., 2024a; Perez et al., 2023; Weidinger et al., 2021). In response, recent work has explored bounded personalization methods (Kirk et al., 2024a) which aim to balance individual alignment with collective well-being.

In another recent work (Halpern et al., 2026), we advocate for a middle ground between full personalization and a single global reward function: a committee of reward functions that collectively capture different preferences of human annotators or users. Fine-tuning a separate language model for each reward function yields a distribution over policies that can be leveraged at inference time to represent a range of viewpoints, whether by synthesizing responses that combine multiple perspectives, selecting the most suitable policy for each user, or sampling from the distribution to reflect a variety of opinions (Feng et al., 2024; Sorensen et al., 2024). In practice, we envision the committee to be much smaller than the total number of users, both for practical reasons, such as computational efficiency, and to avoid overfitting to individual annotators or amplifying extreme or outlier views. This approach thus offers a thoughtful compromise between the complexity and overfitting risks of full personalization, and the simplicity of a single global model, which may fail to adequately represent the range of user preferences. This idea of creating a committee of reward functions has already received some support (Feng et al., 2024; Sorensen et al., 2024). However, existing work primarily explores this direction through experimental setups where the desired reward functions are already given. Our focus is on laying the mathematical and algorithmic foundations for constructing a committee of k reward functions that proportionally reflects the heterogeneous preferences of n human annotators.

In our recent work (Halpern et al., 2026), we introduce a method for achieving proportional representation by constructing a committee of reward functions whose collective preferences reflect those of human annotators. Specifically, for any pair of candidate responses A and B, the fraction of reward functions in the committee that assign a higher score to A than to B is designed to match the proportion of annotators who prefer A over B. This approach relies solely on aggregated pairwise comparison data—a common feedback format in RLHF rather than on each annotator's complete preference ranking. In our model, each annotator has a personal reward function that induces a ranking over all candidates; when asked to compare A and B, the annotator selects the higher-ranked option. By aggregating such comparisons across many annotators, we can estimate the empirical fraction that prefers A to B. The committee is then constructed so that, in its induced rankings, the proportion of reward functions ranking A above B aligns with this empirical fraction, thereby ensuring proportional representation (see Figure 3).

We first observe that while it is trivial to satisfy pairwise calibration with an ensemble large enough to match the number of annotators, constructing such ensembles is NP-hard. However, the primary goal is to identify practical ensembles with small support, meaning that the value of k is significantly smaller than the value of n. To this end, we show that an ensemble of size $(1/\varepsilon)$ suffices to achieve an average proportional representation error of ε. Furthermore, we show that proportional representation can be indeed learned from a limited number of pairwise comparisons queries without having access to the exact preferences of all users.

We also conducted experiments with four public datasets, constructing ensembles of reward models of size 8, using a simple greedy algorithm. Our goal was to investigate whether a small number of reward functions can overcome the limitations of relying on a single reward function. To this end, we compared the committee to a theoretically optimal single deterministic reward model that, for each comparison, always selects the majority-preferred answer. Since the induced preference relations are not necessarily transitive, this serves as the strongest possible

benchmark achievable by a single reward function. We measured the average proportionality error, defined as the average squared difference between the proportion of annotators who prefer A to B and the proportion of reward functions that assign a higher reward to A than to B. Our results show that, in many cases, an ensemble of only 2–4 reward functions already reduce the error on held-out prompts substantially, from 0.20 to 0.05. Therefore, combining a small number of reward models can yield significantly better proportional representation than any single deterministic model can achieve. Our contribution in this setting can be summarized as follows. First, we identify a fundamental limitation of learning a single reward function under heterogeneous human preferences. Second, we propose proportional representation as a guiding principle for aggregating preferences via a committee of reward functions. Third, we provide both theoretical guarantees and empirical evidence, indicating that small committees can achieve low proportionality error.

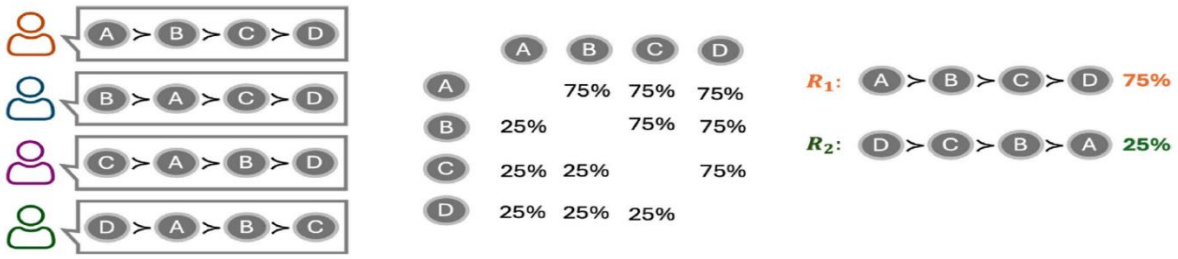


Figure 3. Proportional Representation via a Committee of Reward Functions. Agents Hold Preferences over Alternatives (Left), from Which Pairwise Comparisons Yield Population-Level Preference Proportions (Middle). A Small Committee of Reward Functions Induces Rankings That Reflect These Proportions (Right).

The Road Ahead for Proportional Representation in AI

We have highlighted two prominent settings, clustering and AI alignment, where proportional representation arises naturally and can ensure that agents receive their proportional share. However, the list is by no means exhaustive. For example, proportionality can also be directly applied to rankings in recommender systems, as recently explored by Revel et al. (2025). In another recent work, Procaccia et al. (2025) applied proportional representation to the challenge of selecting a subset of metrics from a larger suite for LLM evaluation, with the goal of choosing a representative subset. Peters (2024) highlighted further potential applications in AI, including mixing outputs from multiple language models so that the combined response reflects their relative strengths rather than defaulting to a single dominant model; training preference models from multiple human labelers; and "virtual democracy," where AI-based preference models act as proxies for individual agents in collective decision-making, ensuring that outcomes proportionally reflect the preferences encoded by these virtual voters. Here, we discuss two additional potential applications of proportional representation, which, to the best of our knowledge, have not yet been explored and could prove highly impactful.

One application arises in federated learning, where each training round involves the client selection problem choosing which subset of available clients (e.g., mobile devices, sensors, or organizations) will participate. Selecting all clients is typically infeasible due to constraints such as limited bandwidth, computation, energy, and communication costs. The goal is to select a representative and informative set of clients while addressing key challenges, such as the heterogeneity (non-IID nature) of the data (Wang et al., 2020) and the need to ensure fairness through equitable selection probabilities among clients (Lai et al., 2021; Shi et al., 2023).We view proportional representation as a natural framework for this setting, as it can enforce the selection of client subsets that preserve data diversity by ensuring that each group of clients with similar datasets is adequately represented. In particular, we can require that in each round,

the selected clients reflect the underlying client population, while over time, every client has an equal probability of participation. This perspective connects naturally with our recent works (Caragiannis, Micha, & Shah, 2024; Ebadian & Micha, 2025), indicating that when agents are embedded in a metric space, it is possible to select subsets that proportionally represent the entire population while also providing participation guarantees. Applied to federated learning, such an approach could ensure that the benefits and burdens of participation are shared fairly, while maintaining representative updates to the global model.

A different application arises in meta-learning, or learning to learn, where the goal is to train models that can quickly adapt to new tasks with limited data by leveraging knowledge from past tasks. Conventional supervised methods require large datasets, whereas real-world scenarios such as robotic learning (Yu et al., 2020) often involve multiple tasks, each with only a few labeled examples. This few-shot learning challenge has motivated optimization-based meta-learning methods (Finn et al., 2017), which train a meta-model whose parameters can be fine-tuned for a new task using only a few samples. However, these methods often assume that the same knowledge acquired from past tasks can be applied uniformly to all new tasks, a limitation when task heterogeneity makes a one-size-fits-all transfer inadequate (Si et al., 2024). One way to address this challenge is to replace the single meta-model with a committee of trained models that collectively cover the diversity of tasks by proportional representation. Specifically, for each sufficiently large subset of tasks, large enough to be encountered frequently, we can maintain a dedicated model that can be fine-tuned with just a few samples, ensuring that even heterogeneous tasks receive tailored starting points. For instance, different users may prefer a robot to operate in different ways (Biyik & Sadigh, 2018). In such cases, the objective is for the robot to adapt to an individual's preferences using only limited feedback (Hejna & Sadigh, 2023), with the appropriate committee member providing the most relevant initialization for fine-tuning.

The rapid emergence of foundation models and increasingly personalized AI systems is expected to further amplify the importance of proportional representation in future research. As AI systems become more deeply integrated into societal decision-making processes, ensuring that diverse stakeholder preferences are represented proportionally is likely to become a central design objective rather than a secondary consideration.

Real-world applications further illustrate the practical value of proportional representation in AI. In recommendation systems, proportional representation can help ensure that minority user communities are not systematically overshadowed by majority preferences, leading to more diverse and inclusive recommendations. In federated learning, it can guide client selection procedures so that updates reflect the diversity of participating users and data distributions. In large language models trained through RLHF, proportional representation can prevent reward models from exclusively reflecting dominant viewpoints, and instead, preserve a broader range of human values. Similarly, in public-sector applications, such as participatory budgeting, policy analysis, and civic engagement platforms, proportional representation can help AI-supported decisions better reflect the composition of the populations they serve. These examples demonstrate that proportional representation is not merely a theoretical concept but a practical design principle with direct implications for fairness, inclusiveness, and robustness in real-world AI systems.

More broadly, we believe proportional representation has the potential to be transformative for AI. Many AI systems combine information from multiple sources, optimize across competing objectives, or rely on feedback from diverse evaluators. Applying proportional representation in these contexts ensures that the selected outputs or models reflect the full distribution of inputs or preferences, rather than being dominated by a single prevailing

tendency. This perspective can guide the design of decision-making algorithms, leading to AI systems whose choices more faithfully represent the range of relevant viewpoints. Beyond proportional representation, other foundational ideas from social choice theory, such as envy-freeness, have also been applied to a variety of machine learning domains, including classification (Balcan et al., 2019) and recommendation systems (Freeman et al., 2021), and beyond (Shah, 2023). These examples highlight the broader potential of social choice, as its principles offer a rigorous framework for reasoning about fairness, representation, and collective decision-making, and can be systematically adapted to guide AI design. Overall, this article argues that proportional representation should be viewed as a general design principle for AI systems that aggregate data preferences, or evaluations from diverse sources, offering mathematical guarantees that diversity is not merely acknowledged but formally respected.

Conclusion

This article has explored proportional representation as a unifying framework for addressing diversity, fairness, and heterogeneous preferences in artificial intelligence. Drawing on ideas from computational social choice, we showed how proportional representation can be extended beyond its traditional role in voting and committee selection to serve as a general design principle for AI systems that aggregate information, preferences, or evaluations from multiple sources. Our discussion highlighted two domains in particular. In clustering, proportional representation provides formal guarantees that sufficiently large and cohesive groups receive influence proportional to their size, both in centroid-based and non-centroid-based settings. In AI alignment and reinforcement learning from human feedback (RLHF), we argued that a single reward function is often insufficient to capture diverse human preferences and presented recent advances based on committees of reward functions that proportionally reflect population-level preferences. We also discussed broader opportunities in recommender systems, federated learning, meta-learning, AI evaluation, and other settings where heterogeneous stakeholders, datasets, or objectives must be represented fairly.

At the same time, important challenges remain. Translating proportional representation principles into practical AI systems requires balancing representation with other objectives such as accuracy, efficiency, robustness, privacy, and scalability. Moreover, different applications may require different notions of similarity, cohesion, and representation, raising both theoretical and algorithmic questions.

Overall, we believe that proportional representation offers a powerful and broadly applicable perspective for the design of future AI systems. As AI increasingly mediates decisions that affect diverse populations, incorporating proportional representation principles can help ensure that interests, preferences, and characteristics of different groups are not only recognized but formally reflected in algorithmic outcomes. Developing this perspective further represents a promising direction for future research at the intersection of artificial intelligence, fairness, and computational social choice.

Theoretical Implications

This article contributes to the growing literature at the intersection of artificial intelligence and computational social choice by highlighting proportional representation as a unifying principle across multiple AI domains. The review demonstrates that concepts originally developed for collective decision-making and voting can be systematically adapted to clustering, AI alignment, recommender systems, and federated learning. By bringing together these diverse research streams, the article provides a common conceptual framework that may facilitate future theoretical developments and encourage the transfer of representation guarantees across different AI settings.

Managerial Implications

The findings discussed in this article have important implications for practitioners and decision-makers responsible for deploying AI systems. In recommendation platforms, proportional representation can help ensure that minority user groups are not systematically underrepresented, leading to more inclusive and trustworthy services. In organizations adopting AI-assisted decision support systems, proportional representation offers a principled approach for incorporating diverse stakeholder preferences into automated decisions. Similarly, managers overseeing federated learning infrastructures, public-sector digital services, or AI-driven customer engagement platforms may use proportional representation principles to improve fairness, transparency, and user acceptance. These practical benefits suggest that proportional representation should be considered not only as a theoretical fairness concept but also as a managerial tool for designing socially responsible AI systems.

References

Amiri, M. K., Zaferani, S. P. G., Emami, M. R. S., Zahmatkesh, S., Pourhanasa, R., Namaghi, S. S., Klemeš, J., Bokhari, A., & Hajiaghaei-Keshteli, M. (2023). Multi-objective optimization of thermophysical properties of powders-DW/EG NF by RSM, NSGA-II, ANN, MLP and ML. *Energy*, *280*, 128176. https://doi.org/10.1016/j.energy.2023.128176

Anyanwu, K., Mansouri, S., & Adei, D. (2025). Towards declarative blockchains: A SHACL-based model for robust and efficient transactions. In *The 2025 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*. https://doi.org/10.1109/ICBC64466.2025.11114583

Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., Singh Lubana, E., Jenner, E., Casper, S., Sourbut, O., Edelman, B., Zhang, Z., Günther, M., Korinek, A., Hernandez-Orallo, J., Hammond, L., Bigelow, E., Pan, A., Langosco, L., Korbak, T., & … Krueger, D. (2024). Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*. https://doi.org/10.48550/arXiv.2404.09932

Arrow, K. J. (1964). *Social choice and individual values* (Vol. 2). Wiley.

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

Aziz, H., Lee, B. E., Morota Chu, S., & Vollen, J. (2024). Proportionally representative clustering. In *Proceedings of the 20th Conference on Web and Internet Economics (WINE)* (pp. 1075–1083).

Balcan, M.-F., Dick, T., Northgate, R., & Procaccia, A. D. (2019). Envy-free classification. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS)* (pp. 1238–1248).

Bıyık, E., Huynh, N., Kochenderfer, M. J., & Sadigh, D. (2024). Active preference-based Gaussian process regression for reward learning and optimization. *International Journal of Robotics Research*, *43*(5), 665–684.

Biyik, E., & Sadigh, D. (2018). Batch active preference-based learning of reward functions. In *Conference on Robot Learning* (pp. 519–528). PMLR.

Caragiannis, I., Micha, E., & Peters, J. (2024). Can a few decide for many? The metric distortion of sortition. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Caragiannis, I., Micha, E., & Shah, N. (2024). Proportional fairness in non-centroid clustering. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Chakraborty, S., Qiu, J., et al. (2024). MaxMin-RLHF: Alignment with diverse human preferences. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 6116–6135).

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

Chen, A., Malladi, S., Zhang, L. H., Chen, X., Zhang, Q., Ranganath, R., & Cho, K. (2024). Preference learning algorithms do not learn preference rankings. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS)* (pp. 101928–101968).

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., et al. (2024). Social choice should guide AI alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Dai, J., & Fleisig, E. (2024). *Mapping social choice theory to RLHF* [arXiv preprint]. arXiv:2404.13038.

Dragan, A. (2024). *Google DeepMind: AI safety …ok doomer: With Anca Dragan* [Video]. YouTube. https://www.youtube.com/watch?v=ZXA2dmFxXmg

Dummett, M. (1984). *Voting procedures*. Oxford University Press.

Dumoulin, V., Johnson, D. D., Castro, P. S., Larochelle, H., & Dauphin, Y. (2024). A density estimation perspective on learning from pairwise human preferences. *arXiv preprint arXiv:2311.14115*. https://doi.org/10.48550/arXiv.2311.14115

Ebadian, S., & Micha, E. (2025). Boosting sortition via proportional representation. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. Forthcoming.

Fairstein, R., Benadè, G., & Gal, K. (2023). Participatory budgeting designs for the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (pp. 5633–5640).

Faliszewski, P., Skowron, P., Slinko, A., & Talmon, N. (2017). Multiwinner voting: A new challenge for social choice theory. *Trends in Computational Social Choice*, *74*(2017), 27–47.

Farahani, R. Z., & Hekmatfar, M. (2009). *Facility location: Concepts, models, algorithms and case studies*. Springer Science & Business Media.

Farahani, R. Z., SteadieSeifi, M., & Asgari, N. (2010). Multiple criteria facility location problems: A survey. *Applied Mathematical Modelling*, *34*(7), 1689–1709.

Feng, S., Sorensen, T., Liu, Y., Fisher, J., Park, C. Y., Choi, Y., & Tsvetkov, Y. (2024). Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4151–4171).

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning* (pp. 1126–1135). PMLR.

Freeman, R., Micha, E., & Shah, N. (2021). Two-sided matching meets fair division. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)* (pp. 203–209).

Gambhir, M. (2026). *A brief literature review and summary of reinforcement learning with human feedback*.

Ge, L., et al. (2024). Axioms for AI alignment from human feedback. *Advances in Neural Information Processing Systems*, *37*, 80439–80465.

Halpern, D., Micha, E., Procaccia, A., & Shapira, I. (2026). Pairwise calibrated rewards for pluralistic alignment. *Advances in Neural Information Processing Systems*, *38*, 57882–57916.

Hejna, D. J., & Sadigh, D. (2023). Few-shot preference learning for human-in-the-loop RL. In *Conference on Robot Learning* (pp. 2014–2025). PMLR. https://doi.org/10.48550/arXiv.2212.03363

Hua, Y., Na, H., Li, Z., Liu, F., Fang, X., Clifton, D., & Torous, J. (2025). A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine*, *8*(1), 230.

Kalayci, Y. H., Kempe, D., & Kher, V. (2024). Proportional representation in metric spaces and low-distortion committee selection. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)* (pp. 9815–9823).

Kellerhals, L., & Peters, J. (2024). Proportional fairness in clustering: A social choice perspective. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*.

Khalifa, M., Elsahar, H., & Dymetman, M. (2021). A distributional approach to controlled text generation. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.

Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2024a). The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, *6*(4), 383–392.

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., & Raileanu, R. (2024b). Understanding the effects of RLHF on LLM generalisation and diversity. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Kupcsik, A., Hsu, D., & Lee, W. S. (2018). Learning dynamic robot-to-human object handover from human feedback. In *Robotics Research: Volume 1* (pp. 161–176). Springer.

Lackner, M., & Skowron, P. (2022). Approval-based committee voting. In *Multi-winner voting with approval preferences* (pp. 1–7). Springer.

Lai, F., Zhu, X., Madhyastha, H. V., & Chowdhury, M. (2021). Oort: Efficient federated learning via guided participant selection. In *15th USENIX Symposium on Operating Systems Design and Implementation (OSDI 21)* (pp. 19–35).

Lake, T., Choi, E., & Durrett, G. (2025). From distributional to overton pluralism: Investigating large language model alignment. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (pp. 6794–6814).

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

Mahpouya, F., Burris, C. J., Paul, H., & Nikolaev, A. (2026). Maximizing the expected value of experimentation for finding top-κ rank via aggregation of pairwise comparisons. *IISE Transactions*, 1–20.

Mansouri, S., Mohammed, H., Korchiev, N., & Anyanwu, K. (2024). Taming smart contracts with blockchain transaction primitives: A possibility? In *the 2024 IEEE International Conference on Blockchain (Blockchain)*.

Micha, E., & Shah, N. (2020). Proportionally fair clustering revisited. In *Proceedings of the 47th International Colloquium on Automata, Languages, and Programming (ICALP)* (Article 85, pp. 1–16).

Mukherjee, A., et al. (2026). SharedRep-RLHF: A shared representation approach to RLHF with diverse preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)* (pp. 27730–27744).

Park, C., Liu, M., Kong, D., Zhang, K., & Ozdaglar, A. (2024). RLHF from heterogeneous feedback via personalization and preference aggregation. *arXiv:2405.00254*.

Peters, D. (2024). Proportional representation for artificial intelligence. In *27th European Conference on Artificial Intelligence*. IOS Press.

Peters, D., Pierczyński, G., & Skowron, P. (2021). Proportional participatory budgeting with additive utilities. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS)* (pp. 12726–12737).

Pezeshgi, A., Abarghoei, M. V., Naeimi, M., & Family, Q. (2026). Trusting the machine: How consumer trust in artificial intelligence shapes future adoption intentions. *Management Science Advances*, *3*(1), 237–245.

Pezeshgi, A., Naeimi, M., & Pezeshgi, Q. (2025). Buying on impulse in the age of AI: Mechanisms, evidence, and moral dilemmas. *Evidence, and Moral Dilemmas (August 07, 2025)*.

Peters, D., & Skowron, P. (2020). Proportionality and the limits of welfarism. In *Proceedings of the 21st ACM Conference on Economics and Computation* (pp. 793–794).

Poddar, S., Wan, Y., Ivison, H., Gupta, A., & Jaques, N. (2024). Personalizing reinforcement learning from human feedback with variational preference learning. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*.

Procaccia, A., Schiffer, B., Wang, S., & Zhang, S. (2025). *Metritocracy: Representative metrics for lite benchmarks* [arXiv preprint]. arXiv:2506.09813.

Revel, M., Milli, S., Lu, T., Watson-Daniels, J., & Nickel, M. (2025). Representative ranking for deliberation in the public sphere. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Safizadeh, M., Yazdanparast, A., & Felix, R. (2026). Taking pride in vegan consumption: A construal level theory account of ad message appeal and future self connectedness. *Psychology & Marketing*.

Sahoo, S., et al. (2025). Position: The complexity of perfect AI alignment—Formalizing the RLHF trilemma [arXiv preprint]. arXiv:2511.19504.

Sattler, F., Müller, K.-R., & Samek, W. (2020). Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(8), 3710–3722.

Shi, Y., Liu, Z., Shi, Z., & Yu, H. (2023). Fairness-aware client selection for federated learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 324–329). IEEE.

Shypula, A., Li, S., Zhang, B., Padmakumar, V., Yin, K., & Bastani, O. (2025). *Evaluating the diversity and quality of LLM generated content* [arXiv preprint]. arXiv:2504.12522.

Small, C., Bjorkegren, M., Erkkilä, T., Shaw, L., & Megill, C. (2021). Polis: Scaling deliberation by mapping high dimensional opinion spaces. *Recerca: Revista de Pensament i Anàlisi*, *26*(2), 1–26.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., & Choi, Y. (2024). A roadmap to pluralistic alignment. arXiv preprint arXiv:2402.05070. https://doi.org/10.48550/arXiv.2402.05070

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. (2020). Learning to summarize from human feedback. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS)* (pp. 3008–3021).

Swamy, G., Dann, C., Kidambi, R., Wu, Z. S., & Agarwal, A. (2024). A minimaximalist approach to reinforcement learning from human feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Viappiani, P., & Boutilier, C. (2010). Optimal Bayesian recommendation sets and myopically optimal choice query sets. In *Proceedings of the 6th Annual Conference on Neural Information Processing Systems (NeurIPS)* (pp. 2352–60).

Wang, H., Z. Kaplan, D. Niu, & Li, B. (2020). Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications* (pp. 1698–707). IEEE.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L.,... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., & Levine, S. (2020, May). Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning* (pp. 1094-1100). PMLR.

Zhang, J., Lin, Y., Lin, M., & Liu, J. (2016). An effective collaborative filtering algorithm based on user preference clustering. *Applied Intelligence*, *45*(2), 230-240.

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

Zhu, B., Jordan, M., & Jiao, J. (2023, July). Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning* (pp. 43037-43067). PMLR.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*. https://doi.org/10.48550/arXiv.1909.08593