Position: Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback

Vincent Conitzer [1 2] **Rachel Freedman** [3] **Jobst Heitzig** [4] **Wesley H. Holliday** [5] **Bob M. Jacobs** [6] **Nathan Lambert** [7] **Milan Mossé** [5] **Eric Pacuit** [8] **Stuart Russell** [3] **Hailey Schoelkopf** [9] **Emanuel Tewolde** [1] **William S. Zwicker** [10 11]

Abstract

Foundation models such as GPT-4 are fine-tuned to avoid unsafe or otherwise problematic behavior, such as helping to commit crimes or producing racist text. One approach to fine-tuning, called *reinforcement learning from human feedback*, learns from humans' expressed preferences over multiple outputs. Another approach is *constitutional AI*, in which the input from humans is a list of high-level principles. But how do we deal with potentially diverging input from humans? How can we aggregate the input into consistent data about "collective" preferences or otherwise use it to make collective choices about model behavior? In this paper, we argue that the field of *social choice* is well positioned to address these questions, and we discuss ways forward for this agenda, drawing on discussions in a recent workshop on Social Choice for AI Ethics and Safety held in Berkeley, CA, USA in December 2023.

[1]Foundations of Cooperative AI Lab, Computer Science Department, Carnegie Mellon University, Pittsburgh, USA [2]Institute for Ethics in AI, University of Oxford, Oxford, United Kingdom [3]Center for Human-Compatible AI, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA [4]Potsdam Institute for Climate Impact Research, Potsdam, Brandenburg, Germany [5]Department of Philosophy, University of California, Berkeley, USA [6]Department of Philosophy and Moral Sciences, Ghent University, Ghent, Begium [7]Allen Institute for AI, Berkeley, California, USA [8]Department of Philosophy, University of Maryland, College Park, USA [9]EleutherAI [10]Department of Mathematics, Union College, Schenectady, USA [11]Murat Sertel Center for Advanced Economic Studies, Istanbul Bilgi University, Istanbul, Turkey. Correspondence to: Vincent Conitzer <conitzer@cs.cmu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Recently, *reinforcement learning from human feedback* (RLHF) has become the primary strategy that leading AI companies such as OpenAI (OpenAI, 2023), Anthropic (Anthropic, 2023), Meta (Meta, 2023), and Google (Google, 2023) use to make pretrained large language models (LLMs) more capable and controllable (Christiano et al., 2017; Ziegler et al., 2019) and to align them with human values. However, RLHF faces many limitations and challenges (Casper et al., 2023; Lambert & Calandra, 2023), including unrepresentative data (Prabhakaran et al., 2021; Feffer et al., 2023), unrealistic models of human decision-making (Hong et al., 2022; Freedman et al., 2021; Siththaranjan et al., 2023; Lambert et al., 2023), and insufficient modeling of human diversity (Kirk et al., 2023; Freedman et al., 2023) which may lead to political bias (Motoki et al., 2023; Rozado, 2024). We propose that ideas from social choice theory (Arrow, 2012; Fishburn, 1973; Kelly, 1988; Brandt et al., 2015)—e.g., concerning whose preferences should be integrated into decisions and how this should be done—are needed to solve many of these open problems.

While models that are solely pretrained on internet data may produce repetitive or harmful text, RLHF trains models to follow instructions (Ouyang et al., 2022) and produce helpful and "harmless" outputs (Bai et al., 2022a) based on human judgments. RLHF gathers example outputs from a pretrained LLM or examples written by humans. Next, humans are asked to select the outputs that best meet specified criteria (such as being "helpful" or "unbiased"). These judgments, often called *preferences*, are then used to fine-tune the LLM. From a social choice perspective, this method raises several critical questions: Which humans are asked to judge outputs? What criteria do they use? How are their judgments combined? And how do their expressed judgments relate to their actual preferences?

Constitutional AI (CAI), which involves reinforcement learning from AI feedback (RLAIF), is an alternative addressing some of these questions (Bai et al., 2022b). Humans produce a "constitution" that specifies principles that the LLM is trained to align with. However, one must still

decide who has input on the constitution and how it is constructed. Bai et al. (2022b) construct it "in a fairly ad-hoc way [. . .] for research purposes", but developing safe and ethical AI requires a more principled approach as in Ganguli et al. (2023) or OpenAI (2024). How then should one aggregate diverse preferences into a representative constitution?

Social choice theory has long studied similar questions. **This position paper argues that methods from social choice should be applied to address questions such as which humans should provide input, what type of feedback should be collected, and how it should be aggregated and used.** By taking into account the lessons from social choice theory, one can avoid naïve mistakes and reinventing the wheel, while leveraging feedback to address challenging design problems (Dobbe et al., 2021). We also highlight areas in which new work is required to extend social choice to new problems unique to training safe and ethical AI.

There are several advantages to addressing the above problems in a principled way. First, it will likely result in a fairer system, taking into account the input of a broader group of people. Second, it promises to give generally more accurate feedback about questions of truthfulness, as suggested by a significant body of literature on "epistemic democracy"—voting to settle questions about facts (Pivato, 2017). Intuitively, having input from diverse people makes it less likely that something important is missed. Third, it will likely result in broader buy-in into the system.

One may have concerns about this approach; for example, will feedback from diverse people be inconsistent and result in inconsistent system behavior? Social choice theory provides examples where naïve aggregation of preferences or judgments leads to seemingly irrational collective choices, such as cyclical preferences (Schwartz, 2018) or logically inconsistent conclusions (List & Pettit, 2002). Then again, social choice theory also provides the tools for thinking about such issues and preventing them.

Social choice is not new to computer scientists; *computational social choice* (Brandt et al., 2015) is a well-studied topic, with a biennial workshop since 2006. However, while many of those researchers affiliate with the AI community, there has not yet been much work connecting computational social choice to the alignment of modern AI systems.

In the following, we first give background on value alignment, RLHF, and social choice. Then we discuss several questions at their intersection. We think that significant further research is needed to answer these questions well and that good answers are needed in order to build AI systems in a responsible way based on potentially diverging feedback from multiple stakeholders. In contrast, ad-hoc approaches may result in systems that fail to represent their stakeholders well, that marginalize significant groups of stakeholders, and that create a basis for conflict between groups of people or the multiple AI systems that represent them.

2. Background

Our proposed research agenda requires background from so far mostly disjoint communities. Readers familiar with some of the following can skip the corresponding subsections.

2.1. Value Alignment

As AI systems become more capable, it becomes critical that they act in alignment with human and societal values (Gabriel, 2020). Many approaches to *value alignment* exist, such as, e.g., formal games in which AI agents must align with humans to solve them (Shah et al., 2020), empirics on the relation between neural network activations and morally relevant output features (Zou et al., 2023), and evaluations of the ethical behavior of LLMs (Pan et al., 2023). RLHF is a particularly popular but so far limited approach.

2.2. Reinforcement Learning from Human Feedback

RLHF begins with generating and evaluating a dataset of model outputs $\mathcal{Y}$. In vanilla RLHF, humans are then shown paired completions $\{y_0, y_1\} \in \mathcal{Y} \times \mathcal{Y}$ to prompts $x \in \mathcal{X}$ and asked which output $y \in \{y_0, y_1\}$ they prefer (Christiano et al., 2017; Lee et al., 2021). Other RLHF variants ask humans to rank or provide scores for groups of outputs (Ziegler et al., 2019; Ouyang et al., 2022), and many additional variations exist (Wu et al., 2023).

The next step is to fit a parameterized reward model $\varrho_\theta \colon \mathcal{Y} \to \mathbb{R}$. For LLMs, ϱ_θ is typically a neural network with weights θ. RLHF methods assume that there is a ground-truth reward function ϱ_{θ^*} that the human preferences reflect (up to noise). The reward model is then optimized to match the likelihoods of the human preferences observed in the data. If the training data comes from diverse sources, this implicitly amounts to a rather intransparent or flawed form of preference aggregation (Siththaranjan et al., 2023; Xu et al., 2023; Chakraborty et al., 2024; Ge et al., 2024).

The final step is to use reinforcement learning to train a policy that maximizes rewards from the reward model. This involves many design decisions—which RL algorithm to use, how to regularize the updates, and whether to gather further online feedback during training. See Uc-Cetina et al. (2023) for a survey on the use of RL to train LLMs.

2.3. Alternate Preference-Based Fine-Tuning Objectives

Since common RL methods like PPO (Schulman et al., 2017) can be unstable, novel techniques for optimizing a

model based on collected preference data have been proposed. Rafailov et al. (2023) introduce Direct Preference Optimization (DPO), which recasts RLHF to converge to the best solution by directly optimizing a loss on the preference label dataset, rather than sampling online from the LLM policy or training an explicit reward model. Another variant emerged to remove the dependency on pairwise data. Ethayarajh et al. (2023) propose a loss function termed Kahneman–Tversky Optimization (KTO) that enables learning a policy from *unpaired* preferences. The authors further claim that the effectiveness of various losses for RLHF depends on the properties they share with proposed human utility functions (Kahneman & Tversky, 1979).

2.4. Constitutional AI

Bai et al. (2022b) further explore the design space with *Constitutional AI (CAI),* which relies on *RL from AI Feedback (RLAIF).* RLAIF is a larger set of techniques for using AI to augment or generate feedback data, including pairwise preferences (Lee et al., 2023; Sharma et al., 2024; Castricato et al., 2024). By employing a human-written set of principles, which they term a *constitution*, Bai et al. (2022b) use a separate LLM to generate artificial preference and instruction data used for fine-tuning. A constitution $\mathcal{C}$ is a set of written principles indicating specific aspects to focus on during a critique phase. The instruction data is curated by repeatedly sampling a principle $c_i \in \mathcal{C}$ and asking the model to revise its latest output y^i to the prompt x to align with c_i. This yields a series of instruction variants $\{y^0, y^1, \cdots, y^n\}$ from the principles $\{c_0, c_1, \cdots, c_{n-1}\}$ used for critique. The final data point is the prompt x together with the final completion y^n, for some n.

The preference data is constructed in a similar, yet simpler way by using a subset of principles from $\mathcal{C}$ as context for a feedback model. The feedback model is presented with a prompt x, a set of principles $\{c_0, \cdots, c_n\}$, and two completions y_0 and y_1 labeled as answers (A) and (B) from a previous RLHF dataset. The feedback models' probability of outputting either (A) or (B) is recorded as a training sample for the reward model, as discussed in Section 2.2.

2.5. Social Choice

Modern social choice theory began in the 1950s with Arrow's Impossibility Theorem (Arrow, 1951; McLean & Urken, 1995). Arrow considered the problem of aggregating multiple individuals' preferences—in his case rankings—into a social preference, subject to some desiderata. In particular, he required the aggregation function to be defined for any family of individual preferences; that the social preference relation be complete and transitive (it is then called a *social welfare function*); that the social preference between alternatives A and B should depend only on individual preferences between A and B (Independence of Irrelevant Alternatives, IIA); and that unanimous individual preference for A over B should imply social preference for A over B. Arrow proved that the only aggregation functions satisfying these desiderata are *dictatorships*: there is one individual d such that no matter what others prefer, if d prefers A to B, then the social preference ranks A above B as well.[1] A similar theorem (see Taylor 2005, § 1.3) holds for *social choice functions* where, instead of asking for a social ranking, we ask for just a single winner, or even just a set of choice-worthy alternatives.

4	4	9	4	2
A	A	B	C	C
B	C	C	A	B
C	B	A	B	A

Borda Count: CBA
Instant Runoff: ABC
Ranked Pairs: BCA

Figure 1. Individual rankings on the left (4 voters say ABC, 4 say ACB, etc.) lead to different aggregations on the right, depending on the aggregation rule. Borda Count gives an alternative 0, 1, or 2 points for each voter who ranks it last, second, or first, respectively; alternatives are then ordered by score. Instant Runoff ranks C last since C has the fewest first-place rankings; after removing C, B has the fewest first-place rankings, so B is in second and A first. For Ranked Pairs, notice there is a *majority cycle*: a majority of voters prefer A to B, a majority prefer B to C, and a majority prefer C to A; the smallest margin of victory is for A over B, so we drop this majority preference, yielding BCA.

Arrow's Theorem stimulated a huge literature exploring the consequences of weakening his desiderata (see, e.g., Campbell & Kelly 2002, Holliday & Pacuit 2020). The general takeaway is that for ordinal preference aggregation, in order to avoid dictatorships, oligarchies and vetoers, one must weaken IIA and allow the social preference between A and B to depend in part on preferences involving other alternatives. This allows for many alternative methods of aggregating individual preferences (see, e.g., Brams & Fishburn 2002; Zwicker 2016; Pacuit 2019 and the voting methods in the Preferential Voting Tools library). Figure 1 gives an example in which three well-known methods disagree. The costs and benefits of these and other methods are systematically studied from various angles (axiomatic, computational, empirical, etc.) in social choice theory.

Since Arrow, social choice theory has grown to study aggregation not only of individuals' preferences, both ordinal and cardinal (d'Aspremont & Gevers, 2002), but also of their *approvals* of alternatives (Laslier & Sanver, 2010), *grades* given to alternatives (Balinski & Laraki, 2010), *judgments* about propositions (Grossi & Pigozzi, 2022), *subjective probabilities* for propositions (Dietrich & List, 2016), and other types of objects (Rubinstein & Fishburn, 1986). Re-

[1] Mishra (2023) applies Arrow's Theorem to RLHF.

cent work has also started to explore social choice methods for RLHF (Song et al., 2024; Swamy et al., 2024; Dai & Fleisig, 2024; Ge et al., 2024). In the following, we discuss some of the aggregation problems that might arise in the context of AI alignment.

3. What Are the Collective Decision Problems and their Alternatives in this Context?

If we want to use social choice methods for aligning AI systems, we first need to specify what the concrete options/objects are, then collect preferences over them, and finally make actual or simulated collective choices between them. These options are called *alternatives*. In some contexts, the set of alternatives is easy to comprehend and enumerate, e.g. candidates for a position. In other settings, there are exponentially many alternatives, but the set is still easy to comprehend, e.g., if each of n propositions must be either accepted or rejected (Lang, 2007).

For AI alignment, it is harder to see how best to determine the set of alternatives for evaluation. It could be the set of all possible *parameterizations* of a given network architecture, but this would surely be conceptually intractable.

For an LLM, most RLHF approaches ask the evaluator to choose between a small, explicit set of alternative *responses* to some prompt, handwritten or sampled from the pretrained LLM. Alternately, we could consider *all* possible responses as alternatives. Evaluators could then indicate their preference by providing the preferred response themselves. Such exemplars are often used for fine-tuning and can be used to learn evaluators' preferences and generate responses that well-represent them (Fish et al., 2023). While this does not address questions about how to generalize beyond a single prompt, it is a useful way of conceptualizing the alternatives.

One might conceive of the alternatives as *probability distributions* over responses; LLMs are anyway configured to respond stochastically. This might be desirable not only for creativity but also to promote fairness and representativeness of responses. In response to a controversial question, fairness might militate against always giving the same answer, as any one answer will inevitably omit some relevant considerations on one side of a debate. There is a large literature on social choice rules that output probability distributions. Their input could be the evaluators' stated preferences between distributions (Fishburn 1973, Ch. 18), or stated preferences between plain alternatives (Brandt, 2017), since evaluators might have difficulties comparing probability distributions. Indeed, the type of objects chosen by a social choice rule (e.g., distributions over responses) need not match the type of objects about which individuals state preferences (e.g., responses).

Multi-winner rules (Faliszewski et al., 2017; Elkind et al., 2017) form a middle ground between deterministic single-winner rules and probabilistic rules. They pick a small, predetermined number of answers that best reflect what the voters want, which could be combined into a single response that lists these answers as bullet points to provide the user with a representative overview of possible answers.

4. Who Provides the Human Feedback?

Let us assume a *stakeholder population* of people who will be affected by an AI system and whose preferences would therefore ideally be considered in aligning the AI.[2] When it is infeasible to elicit feedback from all stakeholders, we must select a smaller group to query. One can try to select a suitably representative subset such that the alignment obtained using their feedback sufficiently approximates the alignment that would be obtained from all stakeholders' feedback. Here one could draw on ongoing work in social choice theory on how to select citizens' assemblies that are representative of a full population (e.g., Flanigan et al. 2021; Landemore & Fourniau 2022), as well as work in statistics on efficient stratified sampling (e.g., Meng 2013).

Alternatively, one could let stakeholders vote on their representatives, e.g., with a voting procedure designed to elect proportionally representative assemblies (see, e.g., Ch. 4 of Lackner & Skowron 2023). Stakeholders might also delegate their feedback rights to others (who may in turn delegate, etc.), as in *liquid democracy* (see Paulin 2020).

Work up to now has used evaluator recruitment methods such as Mechanical Turk (Freedman et al., 2020; Bai et al., 2022a); Upwork, Scale AI, or Lionbridge (Stiennon et al., 2020; Ziegler et al., 2019); and purpose-built platforms (Noothigattu et al., 2018). We believe this component of the RLHF pipeline deserves a more in-depth discussion, including one informed by social choice theory.

5. What Is the Format of Human Feedback?

As discussed, human feedback for AI systems can have various forms; which of these are most natural and useful? Here, we can draw on a significant literature on *preference elicitation* (see, e.g., Sandholm & Boutilier 2006), studying how best to query agents for their preferences in several domains.[3] This literature is closely tied to that of *communication complexity* (e.g., Kushilevitz & Nisan 1997), which is concerned with minimizing the number of bits needed to communicate something. Both of these topics have also been studied in voting settings (Conitzer & Sandholm, 2002; 2005; Service & Adams, 2012).

[2] Some stakeholders, such as small children and non-human animals, whose preferences we cannot easily elicit might need to be represented by dedicated others.

[3] Incidentally, Li et al. (2023) propose to use LLMs for this.

5.1. Multiple Format Options

In general, we want the type of input or feedback that we ask of humans to be (1) natural to give, (2) informative about their preferences and values, and (3) of a type that can be used to align AI systems. For example, having humans comment on an AI output in an open-ended text box may satisfy 1 and 2, but not 3 (at least, with current methods). Having them sort responses alphabetically may satisfy 1 and 3, but not 2. Having them directly rank neural networks based on inspecting their weights may satisfy 3 but not 1 or 2. Needless to say, different choices for the type of input may lead to differently aligned systems and have different behavioral effects on humans (cf. Section 8).

Perhaps one should let individuals *choose* the format in which they give input or feedback. In traditional social choice this is uncommon—though there may be some flexibility in how preferences are expressed (e.g., allowing voters to only rank a few alternatives rather than all (Halpern et al., 2023), or to give numerical ratings instead of rankings) and some variety in the interaction mechanism (e.g., one can vote for candidates individually, or pull a lever to vote for all candidates of a single party at once). Going forward, however, it is easy to imagine giving evaluators choices from a range of different ways to give their input. For example, the input could be individual responses, whole dialogue sessions, longterm interactions with the same user, or published guiding principles. The feedback could be expressed as approval/disapproval votes, pairwise comparisons of alternatives, full or partial rankings of the alternatives, giving precise or imprecise ratings of the form "I rate A between 7 and 9", or even in free-form verbal feedback that an LLM can interpret as some formal preference data such as a partial ordering. Moreover, the evaluation could be on various aspects of the system's behavioral patterns, as done in fine-grained RLHF (Wu et al., 2023), or in RLHF that optimizes for multiple attributes (Dong et al., 2023) such as helpfulness, humor, toxicity, etc. Section 5.2 discusses how we may process all this heterogeneous data.

5.2. Processing Diverse and Informal Feedback

Recall that in RLHF, human feedback is typically used to train a reward (or "preference") model, which maps any possible AI system response to a numerical rating. The concept of reward models could also be used to convert an evaluator's diverse input into a common form, in order to then aggregate it with other evaluators' input.

First, an *individual evaluation interpretation model* ϕ could be trained to map a tuple of inputs of the form $(x, \mathcal{Y}, f_i, e, y)$ to a numerical evaluation r. As before, x represents a prompt to the AI system, $\mathcal{Y}$ the set of possible AI responses, and $y \in \mathcal{Y}$ a particular response. Vector f_i represents the relevant features of a certain evaluator i, and e is a language representation of i's feedback on possible responses $\mathcal{Y}$ to x, containing preference- and evaluation-related statements of various types (see Section 5.1). In practice, ϕ would likely be based on an LLM pretrained to understand the texts x, $\mathcal{Y}$, e, and y, that is then fine-tuned to the interpretation task described above. Then the output $r = \phi(x, \mathcal{Y}, f_i, e, y)$ of ϕ is a numerical rating of y given by evaluator i that is trained to be (approximately) consistent with the verbal evaluation e of that evaluator. We note that this task can be seen as a form of meta-learning.

One could then use the trained evaluation interpretation model ϕ to train another model—an *individual preference model* ψ—that skips verbal evaluations and directly maps inputs $(x, \mathcal{Y}, f_i, y)$ to ratings $r = \psi(x, \mathcal{Y}, f_i, y)$. Namely, any tuple $(x, \mathcal{Y}, f_i, e)$ can be converted into supervised training data $\big((x, \mathcal{Y}, f_i, y), \phi(x, \mathcal{Y}, f_i, e, y)\big)_{y \in \mathcal{Y}}$ for ψ, containing simulated ratings $r = \phi(x, \mathcal{Y}, f_i, e, y)$. The hope is that the individual preference model ψ would be able to simulate the rating of any evaluator (represented by their features f_i), as long as the evaluator, prompt, and response set come from the same distribution as the one ψ was trained on. Similar to the preference models used in current RLHF, ψ could finally be used to fine-tune the actual AI system or steer its behavior in real time. In fact, if the evaluators' features f_i are omitted in the training process sketched above, ψ reduces to the standard preference model as is already used in RLHF. This is vulnerable to, for example, evaluators that strategically misreport (Siththaranjan et al., 2023), or to issues that arise from a disproportionate representation in the set of evaluators. We add evaluators' features f_i to ψ so that preferences can later be aggregated in a transparent and deliberate manner via an additional *social choice step*, as we discuss next.

6. How Do We Incorporate Diverse Individual Feedback?

Here we sketch several variants of two approaches for including diverse input or feedback into AI systems in a consistent way using methods from social choice theory. The first proposes adding an additional *preference aggregation step* somewhere during training, thereby turning RLHF into RLCHF: Reinforcement Learning from Collective Human Feedback. The second approach instead proposes adding an additional *simulated collective decision step* somewhere in the training or the system's real-time decision procedure, similar to Bakker et al. (2022) and Jarrett et al. (2023).

6.1. Proposal: Reinforcement Learning from Collective Human Feedback (RLCHF)

Preference aggregation could be incorporated into RLHF in several ways, early on or rather late. For simplicity, assume

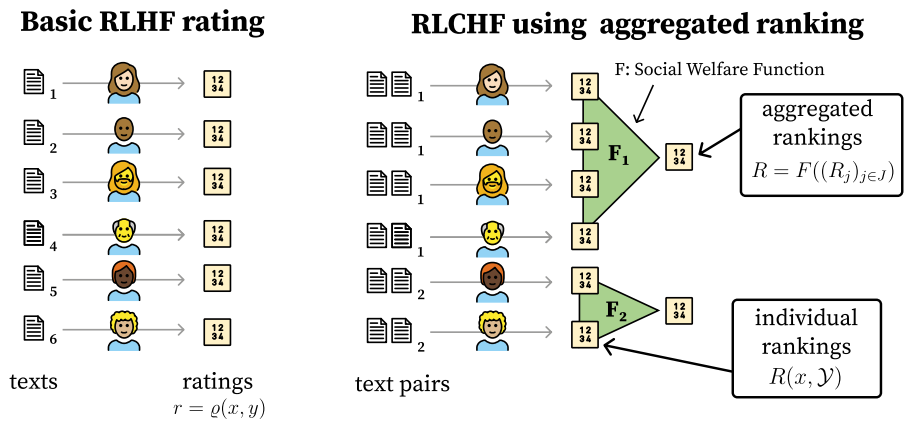

Figure 2. **RLCHF using aggregated rankings**. The core addition to the standard RLHF process is the call-out of an explicit social welfare function, F, which determines how preferences are aggregated.

a base version of *rankings-based* RLHF that (1) takes a database of prompts x together with corresponding sets of possible responses $\mathcal{Y}$, (2) asks one associated evaluator $i(x, \mathcal{Y})$ to provide a ranking $R(x, \mathcal{Y})$ of the elements of $\mathcal{Y}$, (3) turns this ranking into $|\mathcal{Y}|$ many data points for training a common preference model ϱ that produces numerical ratings $r = \varrho(x, y)$, and (4) uses these ratings as rewards in fine-tuning the actual LLM via reinforcement learning.

The earliest point one may introduce preference aggregation is between steps (2) and (3). Instead of a single evaluator, we may ask the members of a jury $J(x, \mathcal{Y})$ to provide individual rankings R_j. Using some ordinal social welfare function F, those rankings can then be aggregated into a collective ranking $R = F((R_j)_{j \in J})$ to be used in step (3). We call this "RLCHF using aggregated rankings" (Fig. 2).

Alternatively, one could use *cardinal* preference aggregation between steps (3) and (4). For this, change step (3) so that a model of *individual* preferences is trained, mapping pair $(x, \mathcal{Y})$ and evaluator i with features f_i to predicted ratings $r_i = \psi(x, f_i, y)$. Also, generate a large sample of feature vectors $f_1, \ldots, f_N$ that is representative of the stakeholder population. Then a *cardinal* social welfare function W can be used to aggregate simulated individual ratings into a social rating $\varrho(x, y) = W(\psi(x, f_1, y), \ldots, \psi(x, f_N, y))$ which can be used in step (4). We call this "RLCHF using evaluator features and aggregated ratings" (Fig. 3).

6.2. Proposal: Simulated Collective Decisions

RLCHF, as described above, keeps the reinforcement learning step that requires numerical rewards, and it uses ordinal or cardinal preference aggregation to produce those rewards for all possible responses $y \in \mathcal{Y}$. A different approach would replace reinforcement learning by something else and introduce social choice methods in the form of simulated collective decisions rather than preference aggregation.

For one thing, one could modify "RLCHF using evaluator features and aggregated ratings" into "Supervised Learning from Simulated Collective Decisions", as shown in Fig. 4. For this, in step (3) from above, use the individual preference model $r_i = \psi(x, f_i, y)$ and feature vectors $f_1, \ldots, f_N$ not to produce an aggregated rating but to simulate a collective choice that picks a single *winning* response $y^* = C\big((\psi(x, f_j, y))_{y \in \mathcal{Y}, j=1,\ldots,N}\big)$. Here, C is now a single-winner social *choice* function. Then in step (4), use data point (x, y^*) to train the actual AI system via supervised (rather than reinforcement) learning. Instead of picking a single winner y^*, we could also use a multi-winner social choice function C that outputs, say, a set of three responses (y', y'', y'''). These can then be (creatively) combined into a single response, e.g., by merging them into a bullet-point list and adding "The following are (three) typical answers to your question: ..." at the beginning.

A more radical modification would drop the fine-tuning-via-learning step altogether (leaving the LLM only pretrained) and instead simulate the collective choice at inference time. Whenever the live system is prompted with some x, generate $k \gg 1$ many candidate responses y_i and $N \gg 1$ many evaluator feature vectors f_j representative of the stakeholder population for the problem $(x, \mathcal{Y})$, and then directly return the winner $y^* = C\big((\psi(x, f_j, y_i))_{j,i=1}^{N,k}\big)$ of the simulated collective choice. Here, too, C could be a multi-winner or probabilistic social choice rule.

7. Which Traditional Social-Choice-Theoretic Concepts Are Most Relevant?

Social choice studies a wide variety of concepts, the relevance of which depends on the specific application. For example, consider the concept of *false-name-proofness* (Yokoo et al., 2001; 2004; Conitzer & Yokoo, 2010), meaning that no-one can benefit from participating multiple times under

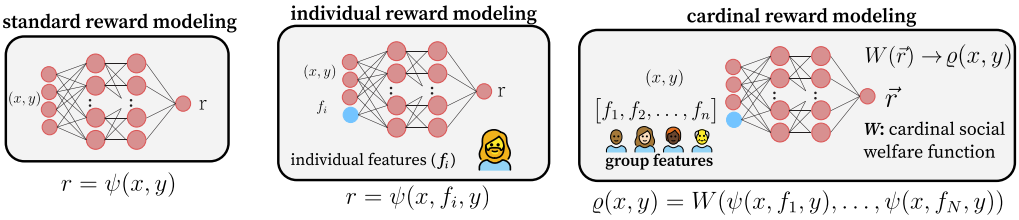

Figure 3. **RLCHF using evaluator features and aggregated ranks**. We show how an individuals' features can be used as an additional input to reward models within the RLHF process.

multiple accounts. This is relevant when voting over the internet, but irrelevant for in-person faculty meetings where faculty vote publicly by raising their hands.

So, rather than studying every single social-choice-theoretic concept in the context of aligning AI systems, we should be careful to evaluate which traditional concepts are most relevant. In the following, we give just a few examples.

7.1. Independence of Clones

In social choice problems, sometimes multiple alternatives, say A and B, compare very similarly against every other alternative X, according to the preferences of individuals. Such alternatives are referred to as *clones*, a notion that can be formalized in several ways. According to a strict notion of clones (Tideman, 1987), A and B are clones if, for every individual, if that individual prefers A to some other alternative X, then they also prefer B to X, and if they instead prefer X to A, then they also prefer X to B. According to a more liberal notion (Laffond et al., 1996), A and B are clones if, whenever a majority of individuals prefer A to some other alternative X, then a majority prefers B to X as well, and whenever a majority prefers some X to A, then a majority prefers X to B as well.

Sometimes the introduction of a clone can affect the outcome of an election. Suppose a group of people are voting over where to go for dinner, and the only two alternatives are a Chinese restaurant and an Indian restaurant. 52% of the voters prefer the Chinese restaurant. But then, someone points out that the Chinese restaurant has two floors and argues that the two floors should be considered separate options. So now the alternatives are C_1, C_2, and I. Nobody really cares all that much about the floor, but suppose that 26% of the voters prefer $C_1 \succ C_2 \succ I$, and 26% of the voters prefer $C_2 \succ C_1 \succ I$ (adding up to the original 52%). Further suppose that the voting rule used is Plurality, in which the alternative that appears at the very top of voters' rankings the most often wins. Then the Indian restaurant ends up winning now with 48% of the vote. This seems like an undesirable property for a voting rule to have; it would be better for the introduction of a clone never to make a difference.[4] This latter desirable property is called *independence of clones*. Perhaps when choosing restaurants, this is not that relevant, as restaurants will rarely be clones (unless the floors of restaurants are treated separately). On the other hand, when choosing responses for a chatbot, it may be quite common for two responses to be very close to each other, indicating its importance to this context. We note that Borda Count (see Figure 1), which is implicitly used in some standard approaches to RLHF (Siththaranjan et al. 2023), badly violates independence of clones.

7.2. Strategic Voting

Another concern is *strategic voting* (or strategic feedback). Strategic voting consists of casting a vote that does not reflect one's true preferences, in order to obtain a better result for oneself. For example, consider an election with plurality voting, as described above. A voter might perceive that her top-ranked alternative has no chance of winning and therefore strategically vote for another alternative. Strategic voting poses a problem because we can no longer take votes (or feedback) at face value. Unfortunately, in general, every reasonable voting rule will sometimes introduce incentives to manipulate (Gibbard, 1973; Satterthwaite, 1975). These incentives to manipulate might be reduced if voters lack full information about the preferences of other voters (Conitzer et al., 2011) or about the voting rule that will be used (Holliday & Pacuit, 2019). But we often cannot guarantee such ignorance, just as we often cannot guarantee computer security through obscurity.

What form might strategic voting in a context such as RLHF take? If rating responses range on a scale from (say) 0 to 10, a natural strategy is to overreport. E.g., if one evaluator does not really like a response (at the level of a 3), but suspects that others would like it (say, two other evaluators give it a 6), then this evaluator may strategically give a

[4]More precisely, introducing a clone should not affect whether a non-clone (e.g., the Indian restaurant in our example) is selected or which non-clone is selected. But it may affect which clone, if any, is selected. For instance, a clone-independent rule could select C_1 over C_2 in our example, if among the 48% of people who prefer I, a strict majority of them prefer C_1 to C_2.

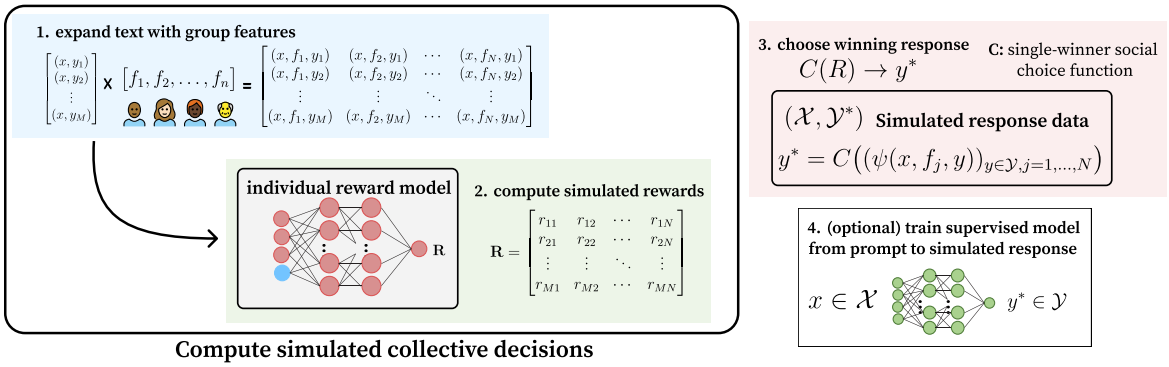

Figure 4. **Supervised Learning from Simulated Collective Decisions**. We show that with an individual or cardinal reward model, as presented in Figure 3, responses y to a prompt x can be simulated. This process expands the scope of studying preferences within RLHF and opens future work on personalization and other topics.

rating of 0 to "compensate" for the other reviewers. This manipulation would be successful if we eventually aggregate ratings by taking their average: the average will be pulled down to 4, instead of the 5 that would result from reporting truthfully, so that the average is closer to the 3 that the evaluator believes is ideal. If instead we use the median as the aggregate, then this manipulation is ineffective—the median would remain 6. Indeed, the median is *strategy-proof* in this context: misreporting one's preferences never helps, as long as one's only goal is to move the median rating closer to one's "true" rating (cf. Moulin 1980).

7.3. Anonymity

In democratic contexts, a standard desideratum on voting rules is *anonymity*: if two voters swap their ballots before submitting them, the output of the voting rule will not change (the rules in Figure 1 all satisfy anonymity in this sense). This captures the idea that the voting rule should not favor some voters over others. Anonymity prohibits not only the extremes of dictatorship (recall Section 2.5) but even any kind of weighted voting wherein some voters' votes count for more than others. However, in the context of AI development, one might consider aggregating human feedback in a way that violates anonymity (cf. the *weighted majority rule* discussed in Nitzan & Paroush 1982). Perhaps some evaluators are more experienced or more highly rated than others; perhaps some are influenced by others, so their input should not be considered completely independent inputs for aggregation; and so on. In general, whether the same democratic norms applied to voting also apply in an AI context is an important question for discussion.

7.4. Principles as Voters

While it is standard in social choice for the voters to be humans, this is not required by the social choice theory framework. In some applications of social choice to AI ethics and safety, possibly including Constitutional AI (recall Section 2.4), we might regard different ethical principles as the "voters" who can rank or otherwise evaluate the outputs of an AI system (cf. Greene et al. 2016). This is analogous to applications of social choice theory in the philosophy of science, where the "voters" are theoretical virtues that may rank scientific theories differently (Okasha, 2011), or to multi-criteria decision-making, where the "voters" are relevant factors that may rank the options differently (Arrow & Raynaud, 1986). Of course, such ethical principles could themselves be outputs of some prior social choice procedure in which the voters are humans (cf. Collective Constitutional AI in Ganguli et al. 2023).

This suggests a possible alternative architecture for applying social choice to AI—one sitting somewhere between the extremes of a spectrum that ranges from Constitutional AI at one end (in which principles are the whole show, while social choice does not appear) to the RLHF version of reinforcement learning as described in previous sections (where principles play no role at all). In this alternative model, each respondent would be required to justify her rankings of alternative AI responses in terms of their level of satisfaction of each of a number of principles taken from a fixed menu. The AI system would use the results to train for several independent tasks: for each principle, separately learn how to rate responses to queries based on that principle alone; and learn how to aggregate those separate ratings into an overall rating of the responses. These would be composed to form the final stage of a simulated collective decision—the stage in which the voters are the principles.

8. How Should We Account for Behavioral Aspects and Human Cognitive Structures?

Preference elicitation often makes idealized assumptions, e.g., that each queried individual has well-defined and con-

sistent preferences and will answer in a way that is consistent with them (up to some random noise). But in reality, myriad behavioral effects kick in. For example, McElfresh et al. (2021) study how (in the context of kidney allocation) humans can become indecisive when asked to give input that will have moral implications.

This leads to a variety of questions about how to align AI systems in the context of behavioral effects. Should we correct for them? That would seem to require having a model where such a behavioral effect obscures humans' "true" values. But do these "true" values correspond to anything real in the world? Do we run the risk of the "correction" actually removing valuable information? Could the ability to make such "corrections" in fact be abused to intentionally remove inconvenient feedback?

9. How Do We Navigate a Multiplicity of AIs?

Consider, again, a group of people voting on a restaurant for their dinner. If there is a significant disagreement, then rather than forcing a minority to go someplace they really do not like, it can make sense to split into multiple groups, each going to their favorite restaurant. Similarly, perhaps it makes sense to create multiple AI systems—for example, to recognize strong inter- and intra-cultural variations that have been identified in some non-homogenous populations (Awad et al., 2018; Peters & Carman, 2024). Depending on the situation, the people providing feedback might be split into groups *ex ante* (e.g., each country makes a system based on their own citizens' feedback), but also *ex post*, where we first collect feedback and then form groups of people. The latter approach is closely related to the topic of *representation* in voting theory (Faliszewski et al., 2017).

There is also the slightly different scenario where one AI system is in place, and some group of people believe that it is not serving them well. Hence, they might decide to pool their resources and create their own system. The literature on *cooperative game theory* (cf. Chalkiadakis et al. 2011), sometimes referred to as *coalitional game theory*, touches on these considerations (and indeed also plays a role in questions of representation, as in Aziz et al. 2017).

Finally, let us highlight possible shortcomings to creating multiple AI systems. As in the restaurant example, it may have the result of unnecessarily dividing people into separate groups. Moreover, splitting into groups may not be feasible if it does not dovetail with existing social structures. For example, the US Federal Government may want to adopt a single system that will impact all its citizens, and adopting two systems would be tantamount to splitting the country in two. Finally, unlike in the case of the restaurants, the multiple AI systems may have to interact with each other, creating the risk of conflict between AIs with different goals. The nascent literature on *cooperative AI* (Dafoe et al., 2021; Conitzer & Oesterheld, 2023) may help ensure these interactions do not result in adverse outcomes. Nonetheless, it might be best to see if we can completely avoid having multiple AIs with competing goals, or at least design them in a way that makes conflict between them less likely.

10. Conclusion

It is important that a variety of stakeholders are involved in giving input or feedback on how AI systems, such as those based on LLMs and other foundation models, should function. But those stakeholders are likely to give conflicting input. If so, how do we aggregate this input or otherwise use it for real or simulated collective decisions to end up with a sensible system? As we have argued in this paper, the field of social choice is well placed to help address this question—conceptually, due to its focus on methods for making consistent collective decisions, e.g., via aggregating preferences, judgments, and other inputs in a consistent way, as well as pragmatically, with many researchers in the computational social choice community being well prepared to engage with AI alignment researchers on these problems.

That said, it is important to acknowledge that aggregating conflicting input or feedback can be a complex task. It requires careful consideration of various factors, such as who the stakeholders are, which humans should provide the feedback, how their input is collected and weighed, the level of expertise and credibility of their input, and potential biases. Additionally, incorporating transparency and accountability measures into the aggregation process can help ensure that the final system reflects a fair and balanced representation of the stakeholders and their input. Significant research is needed to deepen our understanding of the possibilities of using social choice for these purposes and the different effects that this will have.

Needless to say, the questions considered are multifaceted and cannot be adequately addressed without complementary (and not necessarily AI-specific) research. The involved practical decisions and associated legal and political considerations are important topics for future research as well.

Last but not least, we have put a particular focus on RLHF in this paper as it is an especially important and fruitful point of contact between social choice and AI. But the insights afforded by social choice theory bear on countless problems. Social choice can be used to more generally determine the objectives that AI systems pursue, the data on which they are trained, and which systems we build in the first place. Given the rapid development of AI systems underway, we urge researchers to begin forging these connections between social choice and AI alignment.

Impact Statement

This paper highlights the need for further research and collaboration between experts in social choice and AI ethics and safety to ensure that AI systems are designed and deployed in a way that aligns with societal values and promotes accountability and transparency. As we discussed briefly in the introduction, we believe the approach proposed in this paper would result in systems that are fairer, that are less likely to have blind spots due to nobody having raised an issue, and that people buy into more broadly.

As we briefly discussed in the introduction, there is perhaps a concern that feedback from a broader set of participants is more likely to be inconsistent and that consequently the resulting system will behave erratically. The idea that naïve aggregation of votes or judgments leads to inconsistency is a familiar one from social choice theory. For example, if three voters rank three alternatives A, B, and C respectively as follows: $A \succ_1 B \succ_1 C$, $C \succ_2 A \succ_2 B$, and $B \succ_3 C \succ_3 A$, then a majority of voters prefers A to B, a majority prefers B to C, and a majority prefers C to A. This illustrates that majority preferences are cyclical and thus arguably irrational. (See Figure 1 for another example.) We encounter similar issues in *judgment aggregation* (for an overview, see Endriss 2015). To illustrate this in our own context, say that it is broadly agreed that an output should be given if and only if it is both safe and helpful. Suppose evaluator 1 believes the output is safe but not helpful, and therefore should not be given. Evaluator 2 believes the output is helpful but not safe, and therefore should not be given. Evaluator 3 believes the output is both safe and helpful, and therefore should be given. Then a majority believes that the output is safe, a majority believes that it is helpful, but a majority believes that it should not be given—so that majority judgments are logically inconsistent. However, social choice theory is precisely concerned with how we should actually obtain *consistent* aggregations and is therefore well placed to address this issue. For example, one common strategy is to restrict to rational or consistent outputs only and among these find one that is in some sense "closest" to the reports (see, e.g., Elkind & Slinko 2015). Therefore, social choice theory is actually well positioned to *help* with the issue of inconsistencies from aggregation.

Acknowledgements

This paper grew out of the workshop on Social Choice for AI Ethics and Safety held at UC Berkeley in December 2023. We thank all the participants of the workshop for fruitful discussions. We are also grateful to Open Philanthropy for a grant that made the workshop possible, as well as further support from the Center for Human-Compatible AI, the C3.ai Digital Transformations Institute, and the Kavli Center for Ethics, Science, and the Public. For helpful comments on this paper, we thank Aditi Raghunathan and anonymous reviewers.

References

Anthropic. Introducing claude, 2023. https://www.anthropic.com/index/introducing-claude, retrieved 2024-01-31.

Arrow, K. J. *Social Choice and Individual Values*. John Wiley & Sons, Inc., New York, 1st edition, 1951.

Arrow, K. J. *Social Choice and Individual Values*. Yale University Press, 2012.

Arrow, K. J. and Raynaud, H. *Social Choice and Multicriterion Decision-Making*. The MIT Press, 1986.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., and Rahwan, I. The Moral Machine experiment. *Nature*, 563(7729):59–64, November 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0637-6. URL https://doi.org/10.1038/s41586-018-0637-6.

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

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional AI: Harmlessness from AI feedback, 2022b. arXiv:2212.08073.

Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

Balinski, M. and Laraki, R. *Majority Judgement: Measuring, Ranking and Electing*. MIT Press, Boston, 2010. doi: 10.7551/mitpress/9780262015134.001.0001.

Brams, S. J. and Fishburn, P. C. Voting procedures. In Arrow, K. J., Sen, A. K., and Suzumura, K. (eds.), *Handbook of Social Choice and Welfare*, volume 1, pp. 173–236. North-Holland, Amsterdam, 2002. doi: 10.1016/s1574-0110(02)80008-x.

Brandt, F. Rolling the dice: Recent results in probabilistic social choice. In Endriss, U. (ed.), *Trends in Computational Social Choice*, pp. 3–26. AI Access, 2017.

Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D. *Handbook of Computational Social Choice*. Cambridge University Press, 2015.

Campbell, D. E. and Kelly, J. S. Impossibility theorems in the Arrovian framework. In Arrow, K. J., Sen, A. K., and Suzumura, K. (eds.), *Handbook of Social Choice and Welfare*, volume 1, pp. 35–94. North-Holland, Amsterdam, 2002.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Castricato, L., Lile, N., Anand, S., Schoelkopf, H., Verma, S., and Biderman, S. Suppressing pink elephants with direct principle feedback, 2024.

Chakraborty, S., Qiu, J., Yuan, H., Koppel, A., Huang, F., Manocha, D., Bedi, A. S., and Wang, M. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. *CoRR*, abs/2402.08925, 2024. doi: 10.48550/ARXIV.2402.08925. URL https://doi.org/10.48550/arXiv.2402.08925.

Chalkiadakis, G., Elkind, E., and Wooldridge, M. *Computational Aspects of Cooperative Game Theory*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2011.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Conitzer, V. and Oesterheld, C. Foundations of cooperative AI. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, pp. 15359–15367, Washington, DC, USA, 2023.

Conitzer, V. and Sandholm, T. Vote elicitation: Complexity and strategy-proofness. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pp. 392–397, Edmonton, AB, Canada, 2002.

Conitzer, V. and Sandholm, T. Communication complexity of common voting rules. In *Proceedings of the ACM Conference on Electronic Commerce (EC)*, pp. 78–87, Vancouver, BC, Canada, 2005.

Conitzer, V. and Yokoo, M. Using mechanism design to prevent false-name manipulations. *AI Magazine*, 31(4): 65–77, 2010.

Conitzer, V., Walsh, T., and Xia, L. Dominating manipulations in voting with partial information. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI-11)*, pp. 638–643. AAAI Press, 2011.

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson K., and Graepel, T. Cooperative AI: machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

Dai, J. and Fleisig, E. Mapping social choice theory to RLHF. *CoRR*, abs/2404.13038, 2024. doi: 10.48550/ARXIV.2404.13038. URL https://doi.org/10.48550/arXiv.2404.13038.

d'Aspremont, C. and Gevers, L. Social welfare functionals and interpersonal comparability. In Arrow, K. J., Sen, A. K., and Suzumura, K. (eds.), *Handbook of Social Choice and Welfare*, volume 1, pp. 459–541. Elsevier Science B.V., 2002.

Dietrich, F. and List, C. Probabilistic opinion pooling. In Hajek, A. and Hitchcock, C. (eds.), *Oxford Handbook of Philosophy and Probability*. Oxford University Press, Oxford, 2016.

Dobbe, R., Gilbert, T. K., and Mintz, Y. Hard choices in artificial intelligence. *Artificial Intelligence*, 300:103555, 2021.

Dong, Y., Wang, Z., Sreedhar, M. N., Wu, X., and Kuchaiev, O. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. *arXiv preprint arXiv:2310.05344*, 2023.

Elkind, E. and Slinko, A. Rationalizations of voting rules. In Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D. (eds.), *Handbook of Computational Social Choice*, chapter 8. Cambridge University Press, 2015.

Elkind, E., Faliszewski, P., Skowron, P., and Slinko, A. Properties of multiwinner voting rules. *Social Choice and Welfare*, 48:599–632, 2017.

Endriss, U. Judgment aggregation. In Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D. (eds.), *Handbook of Computational Social Choice*, chapter 17. Cambridge University Press, 2015.

Ethayarajh, K., Xu, W., Jurafsky, D., and Kiela, D. Human-centered loss functions (halos). Technical report, Contextual AI, 2023. https://github.com/ContextualAI/HALOs/blob/main/assets/report.pdf, retrieved 2024-01-31.

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

Feffer, M., Heidari, H., and Lipton, Z. C. Moral machine or tyranny of the majority? *arXiv preprint arXiv:2305.17319*, 2023.

Fish, S., Gölz, P., Parkes, D. C., Procaccia, A. D., Rusak, G., Shapira, I., and Wüthrich, M. Generative social choice. *arXiv preprint arXiv:2309.01291*, 2023.

Fishburn, P. C. *The Theory of Social Choice*. Princeton Legacy Library. Princeton University Press, 1973.

Flanigan, B., Gölz, P., Gupta, A., Hennig, B., and Procaccia, A. D. Fair algorithms for selecting citizens' assemblies. *Nature*, 596:548–552, 2021.

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

Freedman, R., Shah, R., and Dragan, A. Choice set misspecification in reward inference. *arXiv preprint arXiv:2101.07691*, 2021.

Freedman, R., Svegliato, J., Wray, K., and Russell, S. Active teacher selection for reinforcement learning from human feedback. *arXiv preprint arXiv:2310.15288*, 2023.

Gabriel, I. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

Ganguli, D. et al. Collective constitutional AI: Aligning a language model with public input. Anthropic, 2023. https://www.anthropic.com/index/collective-constitutional-ai-aligning-a-language-model-with-public-input, retrieved 2024-01-31.

Ge, L., Halpern, D., Micha, E., Procaccia, A. D., Shapira, I., Vorobeychik, Y., and Wu, J. Axioms for ai alignment from human feedback. *CoRR*, abs/2405.14758, 2024. doi: 10.48550/ARXIV.2405.14758. URL https://doi.org/10.48550/arXiv.2405.14758.

Gibbard, A. Manipulation of voting schemes: a general result. *Econometrica*, 41:587–601, 1973.

Google. Bard, 2023. https://bard.google.com/, retrieved 2024-01-31.

Greene, J., Rossi, F., Tasioulas, J., Venable, K. B., and Williams, B. C. Embedding ethical principles in collective decision support systems. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 4147–4151, Phoenix, AZ, USA, 2016.

Grossi, D. and Pigozzi, G. *Judgment Aggregation: A Primer*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer Cham, 1 edition, 2022. doi: 10.1007/978-3-031-01568-7.

Halpern, D., Kehne, G., Procaccia, A. D., Tucker-Foltz, J., and Wüthrich, M. Representation with incomplete votes. In Williams, B., Chen, Y., and Neville, J. (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 5657–5664. AAAI Press, 2023. doi: 10.1609/AAAI.V37I5.25702. URL https://doi.org/10.1609/aaai.v37i5.25702.

Holliday, W. H. and Pacuit, E. Strategic voting under uncertainty about the voting method. In Moss, L. S. (ed.), *Theoretical Aspects of Rationality and Knowledge: Proceedings of the 2019 Conference (TARK 2019)*, volume 297 of *Electronic Proceedings in Theoretical Computer Science*, pp. 252–272. EPTCS, 2019. doi: 10.4204/EPTCS.297.17.

Holliday, W. H. and Pacuit, E. Arrow's decisive coalitions. *Social Choice and Welfare*, 54:463–505, 2020. doi: 10.1007/s00355-018-1163-z.

Hong, J., Bhatia, K., and Dragan, A. On the sensitivity of reward inference to misspecified human models. *arXiv preprint arXiv:2212.04717*, 2022.

Jarrett, D., Pislar, M., Bakker, M. A., Tessler, M. H., Koster, R., Balaguer, J., Elie, R., Summerfield, C., and Tacchetti, A. Language agents as digital representatives in collective decision-making. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

Kahneman, D. and Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. ISSN 00129682, 14680262.

Kelly, J. S. *Social Choice Theory: An Introduction*. Springer, Berlin, 1988. doi: 10.1007/978-3-662-09925-4.

Kirk, H. R., Vidgen, B., Röttger, P., and Hale, S. A. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453*, 2023.

Kushilevitz, E. and Nisan, N. *Communication Complexity*. Cambridge University Press, 1997.

Lackner, M. and Skowron, P. *Multi-Winner Voting with Approval Preferences*. SpringerBriefs in Intelligent Systems. Springer Cham, 2023. doi: 10.1007/978-3-031-09016-5.

Laffond, G., Lainé, J., and Laslier, J. Composition-consistent tournament solutions and social choice functions. *Social Choice and Welfare*, 13:75–93, 1996. doi: 10.1007/BF00179100.

Lambert, N. and Calandra, R. The alignment ceiling: Objective mismatch in reinforcement learning from human feedback. *arXiv preprint arXiv:2311.00168*, 2023.

Lambert, N., Gilbert, T. K., and Zick, T. The history and risks of reinforcement learning and human feedback. *arXiv preprint arXiv:2310.13595*, 2023.

Landemore, H. and Fourniau, J.-M. Citizens' assemblies, a new form of democratic representation? *Participations: Revue de sciences sociales sur la démocratie et la citoyenneté*, 34:5–36, 2022.

Lang, J. Vote and aggregation in combinatorial domains with structured preferences. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1366–1371, Hyderabad, India, 2007.

Laslier, J.-F. and Sanver, M. R. (eds.). *Handbook on Approval Voting*. Studies in Choice and Welfare. Springer Berlin Heidelberg, 1 edition, 2010. doi: 10.1007/978-3-642-02839-7.

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., and Rastogi, A. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Lee, K., Smith, L., and Abbeel, P. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021.

Li, B. Z., Tamkin, A., Goodman, N., and Andreas, J. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*, 2023.

List, C. and Pettit, P. Aggregating sets of judgments: An impossibility result. *Economics & Philosophy*, 18(1): 89–110, 2002.

McElfresh, D. C., Chan, L., Doyle, K., Sinnott-Armstrong, W., Conitzer, V., Borg, J. S., and Dickerson, J. P. Indecision modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 5975–5983, 2021.

McLean, I. and Urken, A. (eds.). *Classics of Social Choice*. The University of Michigan Press, Ann Arbor, 1995.

Meng, X. Scalable simple random sampling and stratified sampling. In *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013.

Meta. Meta and microsoft introduce the next generation of llama, 2023. https://about.fb.com/news/2023/07/llama-2/, retrieved 2024-01-31.

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

Motoki, F., Neto, V. P., and Rodrigues, V. More human than human: measuring chatgpt political bias. *Public Choice*, 198:3–23, 2023.

Moulin, H. On strategy-proofness and single peakedness. *Public Choice*, 35:437–455, 1980. doi: 10.1007/BF00128122.

Nitzan, S. and Paroush, J. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, 23(2):289–297, 1982.

Noothigattu, R., Gaikwad, S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., and Procaccia, A. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Okasha, S. Theory choice and social choice: Kuhn versus Arrow. *Mind*, 120(477):83–115, 2011.

OpenAI. GPT-4 technical report, 2023.

OpenAI. Democratic inputs to ai grant program: lessons learned and implementation plans, 2024. https://openai.com/blog/democratic-inputs-to-ai-grant-program-update, retrieved 2024-01-31.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Pacuit, E. Voting methods. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2019 edition, 2019.

Pan, A., Chan, J. S., Zou, A., Li, N., Basart, S., Woodside, T., Zhang, H., Emmons, S., and Hendrycks, D. Do the rewards justify the means? measuring trade-offs between

rewards and ethical behavior in the machiavelli benchmark. In *International Conference on Machine Learning*, pp. 26837–26867. PMLR, 2023.

Paulin, A. An overview of ten years of liquid democracy research. In *Proceedings of the 21st Annual International Conference on Digital Government Research*, pp. 116–121, 2020. doi: 10.1145/3396956.3396963.

Peters, U. and Carman, M. Cultural bias in explainable ai research: A systematic analysis. *J. Artif. Int. Res.*, 79, mar 2024. ISSN 1076-9757. doi: 10.1613/jair.1.14888. URL https://doi.org/10.1613/jair.1.14888.

Pivato, M. Epistemic democracy with correlated voters. *Journal of Mathematical Economics*, 72:51–69, 2017.

Prabhakaran, V., Davani, A. M., and Diaz, M. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*, 2021.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Rozado, D. The political preferences of llms. *arXiv preprint arXiv:2402.01789*, 2024.

Rubinstein, A. and Fishburn, P. C. Algebraic aggregation theory. *Journal of Economic Theory*, 38(1):63–77, 1986.

Sandholm, T. and Boutilier, C. Preference elicitation in combinatorial auctions. In Cramton, P., Shoham, Y., and Steinberg, R. (eds.), *Combinatorial Auctions*, chapter 10, pp. 233–263. MIT Press, 2006.

Satterthwaite, M. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10:187–217, 1975.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Schwartz, T. *Cycles and Social Choice: The True and Unabridged Story of a Most Protean Paradox*. Cambridge University Press, 3 2018. doi: 10.1017/9781316848371.

Service, T. C. and Adams, J. A. Communication complexity of approximating voting rules. In *Proceedings of the Eleventh International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pp. 593–602, Valencia, Spain, 2012.

Shah, R., Freire, P., Alex, N., Freedman, R., Krasheninnikov, D., Chan, L., Dennis, M. D., Abbeel, P., Dragan, A., and Russell, S. Benefits of assistance over reward learning. In *NeurIPS Workshop on Cooperative AI*, 2020.

Sharma, A., Keh, S., Mitchell, E., Finn, C., Arora, K., and Kollar, T. A critical evaluation of ai feedback for aligning large language models, 2024.

Siththaranjan, A., Laidlaw, C., and Hadfield-Menell, D. Distributional preference learning: Understanding and accounting for hidden context in rlhf. *arXiv preprint arXiv:2312.08358*, 2023.

Song, F., Yu, B., Li, M., Yu, H., Huang, F., Li, Y., and Wang, H. Preference ranking optimization for human alignment. In Wooldridge, M. J., Dy, J. G., and Natarajan, S. (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 18990–18998. AAAI Press, 2024. doi: 10.1609/AAAI.V38I17.29865. URL https://doi.org/10.1609/aaai.v38i17.29865.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Swamy, G., Dann, C., Kidambi, R., Wu, Z. S., and Agarwal, A. A minimaximalist approach to reinforcement learning from human feedback. *CoRR*, abs/2401.04056, 2024. doi: 10.48550/ARXIV.2401.04056. URL https://doi.org/10.48550/arXiv.2401.04056.

Taylor, A. D. *Social Choice and the Mathematics of Manipulation*. Cambridge University Press, Cambridge, 2005. doi: 10.1017/cbo9780511614316.

Tideman, T. N. Independence of clones as a criterion for voting rules. *Social Choice and Welfare*, 4(3):185–206, 1987.

Uc-Cetina, V., Navarro-Guerrero, N., Martin-Gonzalez, A., Weber, C., and Wermter, S. Survey on reinforcement learning for language processing. *Artificial Intelligence Review*, 56(2):1543–1575, 2023.

Wu, Z., Hu, Y., Shi, W., Dziri, N., Suhr, A., Ammanabrolu, P., Smith, N. A., Ostendorf, M., and Hajishirzi, H. Fine-grained human feedback gives better rewards for language model training. *arXiv preprint arXiv:2306.01693*, 2023.

Xu, W., Dong, S., Lu, X., Lam, G., Wen, Z., and Roy, B. V. RLHF and IIA: perverse incentives. *CoRR*, abs/2312.01057, 2023. doi: 10.48550/ARXIV.2312.01057. URL https://doi.org/10.48550/arXiv.2312.01057.

Yokoo, M., Sakurai, Y., and Matsubara, S. Robust combinatorial auction protocol against false-name bids. *Artificial Intelligence*, 130(2):167–181, 2001.

Yokoo, M., Sakurai, Y., and Matsubara, S. The effect of false-name bids in combinatorial auctions: New fraud in Internet auctions. *Games and Economic Behavior*, 46(1): 174–188, 2004.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

Zwicker, W. S. Introduction to the theory of voting. In Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D. (eds.), *Handbook of Computational Social Choice*, pp. 23–56. Cambridge University Press, New York, 2016. doi: 10.1017/cbo9781107446984.003.