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**[3], **Bob M. Jacobs**[5], **Nathan Lambert**[6], **Milan Mossé**[3], **Eric Pacuit**[7], **Stuart Russell**[3], **Hailey Schoelkopf**[8], **Emanuel Tewolde**[1] and **William S. Zwicker**[9,10]

[1]Carnegie Mellon University
[2]University of Oxford
[3]University of California, Berkeley
[4]Potsdam Institute for Climate Impact Research
[5]Ghent University
[6]Allen Institute for AI
[7]University of Maryland, College Park
[8]EleutherAI
[9]Union College
[10]Istanbul Bilgi University
Correspondence to: conitzer@cs.cmu.edu

Abstract

Foundation models such as GPT-4 are fine-tuned to avoid unsafe or otherwise problematic behavior, so that, for example, they refuse to comply with requests for help with committing crimes or with 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 Introduction

Over the past year, *reinforcement learning from human feedback* (RLHF) has played a key role in making large language models (LLMs) more capable and controllable [Christiano *et al.*, 2017; Ziegler *et al.*, 2019]. RLHF is now the primary strategy that leading AI companies such as OpenAI [OpenAI, 2023], Anthropic [Anthropic, 2023], Meta [Meta, 2023], and Google [Google, 2023] use to align pretrained LLM models with human values. However, RLHF faces many limitations and concrete challenges [Casper *et al.*, 2023; Lambert and 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]. We hold the position that core ideas from social choice theory [Arrow, 2012; Fishburn, 1973; Kelly, 1988; Brandt *et al.*, 2015]—primarily concerning whose preferences should be integrated into decisions and how this should be done—are needed to solve many of the open problems facing RLHF.

While models that are solely pretrained on internet data may produce repetitive or harmful text, RLHF enables training 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 an LLM that has been pretrained to predict a text corpus. Next, humans are asked to select the outputs that best meet specified criteria (such as being "helpful" or "unbiased"). Humans may also manually write the outputs to be compared, but due to cost, human input is often limited to these comparative judgements. These judgments, often called *preferences*, are then used to fine-tune the LLM to produce more desirable outputs. From a social choice perspective, this method raises several critical questions: Which humans are asked to judge models? 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 alternate approach that directly addresses some of these questions [Bai *et al.*, 2022b]. Humans produce a "constitution" that explicitly specifies principles to guide the LLM training process. The

LLM is then trained to align with this constitution. However, we must still decide who has input on the constitution and how it is constructed. Bai *et al.* [2022b] construct their constitution "in a fairly ad-hoc way [. . .] for research purposes", but developing safe and ethical AI requires a more principled approach, as exemplified in Ganguli and others [2023] or announced in OpenAI [2024]. How then should one aggregate diverse preferences into a representative constitution?

Social choice theory has long studied similar questions, and by taking into account its lessons, one can avoid making naïve mistakes and reinventing the wheel. In this paper, we argue that tools and theories from social choice should be applied to these open problems, in particular in RLHF, to help bridge challenging design problems to sociotechnical questions [Dobbe *et al.*, 2021]. Specifically, we demonstrate how such tools can be used to begin addressing which humans should provide input or feedback, what type of feedback they should provide, and how that feedback should be aggregated and used. 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 a number of advantages to addressing these problems in a principled way. First, it is likely to result in a fairer system that takes into account the input or feedback of a broader group of people. Second, there are reasons to believe that this will result in generally more accurate feedback about questions of truthfulness; cf. the literature on "epistemic democracy" – voting to settle questions about facts [Pivato, 2017]. Intuitively, having input from a more diverse group of people makes it less likely that something important is missed. Third, it will likely result in broader buy-in into the system. For example, important issues such as political biases of LLMs [Motoki *et al.*, 2023] have been hypothesized to emerge from the finetuning phase that follows pretraining [Rozado, 2024].

One may also have concerns about this approach; for example, is feedback from a diverse group of people going to be inconsistent and consequently result in inconsistent behavior from the system? Social choice theory provides a number of examples where naïve aggregation of preferences or judgments leads to choices in the aggregate that seem irrational, such as cyclical preferences [Schwartz, 2018] or logically inconsistent conclusions [List and Pettit, 2002]. Then again, social choice theory also provides the tools for thinking about such issues and preventing them.

In the remainder of this paper, we first give background on value alignment, RLHF, and social choice. Then we discuss a number of questions at the intersection of these topics. We believe that significant further research is required to answer each of these questions well and that good answers to them are needed to build AI systems in a responsible way based on potentially diverging feedback from multiple stakeholders.

2 Background

Our proposed research agenda requires background on topics that have so far been studied by mostly disjoint communities. A reader familiar with some of these topics can skip the corresponding subsections.

2.1 Value Alignment

As advanced AI systems become increasingly capable, it becomes critical that they act in a way that aligns with human and societal values [Gabriel, 2020]. There are many approaches to *value alignment*, including theoretical work to define formal games that AI agents must align with humans to solve [Shah *et al.*, 2020], empirical investigation of the relationship between neural network activations and morally relevant output features [Zou *et al.*, 2023], and evaluations of the ethical behavior of state-of-the-art models [Pan *et al.*, 2023]. RLHF is a particularly popular approach to value alignment, but it faces many limitations in its current form.

2.2 Reinforcement Learning from Human Feedback

Preference data collection. The first step in RLHF is to generate and evaluate 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 $x \in \mathcal{X}$ of these outputs and asked to select which output $p \in \{y_0, y_1\}$ they prefer from each pair [Christiano *et al.*, 2017; Lee *et al.*, 2021]. Other RLHF variants require 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].

Reward model training. The next step is to fit a parameterized reward model $\varrho_\theta : \mathcal{Y} \to \mathbb{R}$. For LLMs, the reward model 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 probabilistic 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 form of preference aggregation [Siththaranjan *et al.*, 2023].

Optimizing the policy with RL. 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 of methods and limitations for using RL to train LLMs.

2.3 Constitutional AI

Bai *et al.* [2022b] further explore the design space by introducing Constitutional AI (CAI), which relies on RL from AI Feedback (RLAIF). RLAIF is a set of techniques for using an AI model to augment or generate feedback data in the form of pairwise preferences or other signals [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*, they use a separate LLM to generate artificial preference and instruction data that can be used for model fine-tuning. A constitution $\mathcal{C}$ is made up of a set of written principles c_i that indicate specific aspects to focus on during a critique phase. The instruction data, which is largely out of the scope of this paper, is curated by repeatedly sampling a principle c_i and asking the model to revise the current completion y_k^0 to the prompt x_k. This yields a series of instruction variants $\{y_k^0, y_k^1, \cdots, y_k^n\}$ from the principles $\{c_{i_0}^0, c_{i_1}^1, \cdots, c_{i_{n-1}}^{n-1}\}$ used for critique at each step. The final data point is the prompt x_k with the final completion y_k^n, for some suitable n.

The preference data is constructed in a similar, yet simpler way by using a subset of principles from the constitution $\mathcal{C}$ 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.

2.4 Social Choice

Modern social choice theory began in the 1950s with Arrow's Impossibility Theorem [Arrow, 1951] (for its long prehistory, see McLean and Urken [1995]). Arrow considered the problem of aggregating multiple individuals' preferences—in the form of complete and transitive rankings of some set of alternatives—into a social preference, subject to a list of normative desiderata. In particular, Arrow assumed that the aggregation function should be defined for any family of individual preferences to be aggregated (Universal Domain); that the outputted social preference relation should be complete and transitive, like individual preferences, in which case the aggregation function is called a *social welfare function*; that the social preference between two alternatives A and B should depend only on individual preferences between A and B (Independence of Irrelevant Alternatives); and that unanimous individual preference for A over B should imply social preference for A over B (Pareto). Arrow proved that if there are at least three alternatives, then the only aggregation functions satisfying these desiderata are *dictatorships*: there is one individual d such that no matter what others prefer, if d strictly prefers A to B, then the social preference ranks A over B as well. A similar theorem (see Taylor 2005, § 1.3) holds for *social choice functions* where, instead of asking for a social ranking of alternatives, we more modestly ask for just a set of choice-worthy alternatives. This also includes the special case of social choice functions that always pick a single winner.

Arrow's Theorem stimulated a huge literature exploring the consequences of weakening Arrow's desiderata (see, e.g., Campbell and Kelly 2002, Holliday and Pacuit 2020, and references therein). The general takeaway is that for ordinal preference aggregation, in order to avoid dictatorships and related pathologies such as oligarchies and vetoers, one must weaken the Independence of Irrelevant Alternatives (IIA) and allow the social preference between two alternatives to depend in part on individual preferences involving other alternatives. With this freedom to relax IIA comes a vast proliferation of alternative methods of aggregating individual preferences (see, e.g., Brams and Fishburn 2002; Zwicker 2016; Pacuit 2019 and the voting methods implemented in the Preferential Voting Tools library). Figure 2 gives an example in which three well-known methods disagree. The costs and benefits of these and other methods are systematically studied from different 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 and Gevers, 2002], but also of their *approvals* of alternatives [Laslier and Sanver, 2010], *grades* given to alternatives [Balinski and Laraki, 2010], *judgments* about propositions [Grossi and Pigozzi, 2022], *subjective probabilities* for propositions [Dietrich and List, 2016], and other types of objects [Rubinstein and Fishburn, 1986]. 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 methods from social choice for the purpose of aligning AI systems, we first need to specify what the concrete options are, before we can start collecting preferences over them and make actual or simulated collective choices between them. These options are called *alternatives* in social choice theory. In some contexts, the set of alternatives is easy to comprehend and enumerate, as when the alternatives are candidates for a position or an award. In other settings, there are exponentially many alternatives, but the set is still easy to comprehend, e.g., when there are n propositions and each of them must be either accepted or rejected [Lang, 2007].

When considering the alignment of AI systems, it is harder to see exactly how best to think about the relevant set of alternatives for evaluation. In principle, it could be the set of all AI systems or all possible parameterizations of a given network architecture, but this would be conceptually intractable.

In the context of an LLM, the RLHF approach traditionally asks the evaluator to choose between a small, explicit set of alternative responses to a single prompt, with each response sampled from the LLM's output distribution. Alternately, we could consider all possible responses as alternatives. While this response set is too large to explicitly enumerate, the evaluators can still 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. This is natural, as LLMs are typically configured to respond stochastically to a prompt. This might be desirable not only for creativity but also to promote fairness and representativeness of responses. For example, in response to a controversial question, fairness might militate against an LLM 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 whose outputs are probability distributions. The inputs to such a rule could be the evaluators' stated explicit preferences between distributions (Fishburn 1973, Ch. 18), but they could also be stated preferences between plain alternatives [Brandt, 2017]. 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 their preferences or evaluations (e.g., responses). This is important, since probability distributions over large sets of responses may be particularly difficult for evaluators to reliably compare.

4 Who Provides the Human Feedback?

Let us assume that there is a population of people, the *stakeholder population*, who will be affected by an AI system and whose preferences would therefore ideally be taken into account in aligning the AI system.[1] Unfortunately, it may be infeasible to elicit feedback from all members of the stakeholder population, so we must select some smaller group from which to elicit feedback. For example, one could try to select a suitably representative subset of the population such that the alignment obtained using feedback from the subset sufficiently approximates the alignment that would be obtained using feedback from the full stakeholder population. 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 and Fourniau 2022), as well as work in statistics on efficient stratified sampling (e.g., Meng 2013).

Another approach would be to allow the full stakeholder population to vote on their representatives in some way. This could be done, for example, with a voting procedure that is designed to elect assemblies that are proportionally representative (see, e.g., Ch. 4 of Lackner and Skowron 2023). Additionally, stakeholders might be allowed to delegate their feedback rights to others (who may in turn delegate, etc.), as in *liquid democracy* (see Paulin 2020).

As of now, earlier work 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 we have discussed, human feedback for AI systems can come in 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 and Boutilier 2006), studying how best to query agents for their preferences in a variety of domains.

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, with current methods, having humans comment on an AI output in an open-ended text box may satisfy 1 and 2, but not 3. 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.

It should be noted that different choices for the type of input or feedback can lead to differently aligned systems, especially if we do not understand the behavioral effects of the different types of input. For example, McElfresh *et al.* [2021] introduce (in the context of feedback on kidney allocation) an *indecision option* among the available choices and reject several natural hypotheses about how the resulting data relate to those obtained without that option.

One question is whether we should actually let individual humans *choose* the format in which they give input or feedback. In traditional social choice, this is uncommon, although there may be some flexibility in how preferences are expressed (e.g., allowing voters to not give a complete ranking but rather only rank a few alternatives [Halpern *et al.*, 2023], or to give numerical ratings instead of ordinal rankings), as well as some variety in the interaction mechanism to get to that expression of preferences (e.g., one can vote for candidates individually but also pull a lever that corresponds to voting for exactly the candidates of a single party).

It is easy to imagine giving evaluators the choice between a range of different ways to give their input or feedback on various aspects of the system's behavior or behavioral patterns or rules (e.g., individual responses, whole dialogue sessions, longterm interaction with the same user, or published guiding principles) and various dimensions of desirability, which is emerging as fine-grained RLHF [Wu *et al.*, 2023] or optimizing attributes in the data [Dong *et al.*, 2023], relating to various values such as "truthfulness", "harmlessness", "fairness", etc., and to allow them to give that feedback in various ways: approving/disapproving, making pairwise comparison statements of the form "I like A better than B", giving full or partial rankings of the form "A is best, B 2nd-best, ...", giving precise or imprecise ratings of the form "I rate A between 7 and 9", or even by giving free-form verbal feedback that the LLM then interprets and converts into some formal data such as a partial ordering. This heterogeneous data could then be transformed in some formal way into a common, sufficiently expressive data structure, such as a utility function.

5.2 Dealing with Diverse and Informal Feedback

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

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. Moreover, vector f_i represents the relevant features of a certain evaluator i, and e shall be a language representation of i's feedback on possible responses $\mathcal{Y}$ to x, containing preference- and evaluation-related statements of whatever type (see Section 5.1). In

[1]There may also be stakeholders, such as small children and nonhuman animals, whose feedback we cannot easily elicit. In that case, we may consider feedback from humans who are charged with representing their interests.

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 were omitted in the training process sketched above, ψ would be a preference model of the same type as is already used in RLHF and could readily be used for it. This would, however, conflate the evaluations of the (possibly not proportionally representative) set of evaluators used in training in a rather uncontrolled and potentially confusing way. An arguably better way of making use of ψ is, therefore, to indeed make use of evaluators' features f_i in training and add an additional *social choice step* to the RLHF pipeline or the AI system's real-time decision-making procedure. Below we sketch several ways in which this might be done.

6 How can Diverse Individual Input or Feedback be Incorporated?

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 suggests adding an additional *preference aggregation step* somewhere during training, thereby turning RLHF into RLCHF: Reinforcement Learning from Collective Human Feedback. The second approach instead suggests 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 as an additional step into RLHF in several ways, from early to rather late in the RLHF pipeline. For clarity of exposition, assume a simple 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 to introduce preference aggregation in this pipeline would be between steps (2) and (3). Instead of a single evaluator $i(x, \mathcal{Y})$, we may ask the members of a jury $J(x, \mathcal{Y})$ of evaluators 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 use it in step (3). This approach could be termed "RLCHF using aggregated rankings", see Fig. 3.

Alternatively, one could use cardinal rather than ordinal preference aggregation at a later point in the pipeline: between steps (3) and (4). For this, change step (3) so that not a model of common but 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 collection 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 into one rating $\varrho(x, y) = W(\psi(x, f_1, y), \ldots, \psi(x, f_N, y))$ which can be used in step (4). This approach could be termed "RLCHF using evaluator features and aggregated ratings", see Fig. 4.

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 these said 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. 1. 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, for example, by merging them into a bullet-point list and adding a sentence "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 rather 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 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?

A wide variety of concepts is studied in social choice. We should be careful to evaluate which traditional concepts are most relevant to aligning AI systems. 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. For example, 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. It turns out 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. This results in the Indian restaurant now actually winning 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. This latter desirable property is called *independence of clones*. Perhaps when choosing restaurants, this is not that important, 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.

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 and 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 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 that give a 6), then this evaluator may strategically give a 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 7. 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.

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 2 all satisfy anonymity). This captures the idea that the voting rule should not favor some voters over others. Anonymity not only prohibits the extremes of dictatorship (recall Section 2.4), 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 and Paroush 1982). Perhaps some evaluators have more experience or a better rating; perhaps some are influenced by others, so their input should not be considered completely independent inputs for aggregation; etc. 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 human agents, this is not the only interpretation of the mathematical framework of social choice theory. In some applications of social choice to AI ethics and safety, possibly including Constitutional AI (recall Section 2.3), 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 and Raynaud, 1986]. Of course, such ethical principles could themselves be outputs of some prior social choice procedure in which the voters are humans.

This principles as voters idea suggests a possible alternative architecture for applying social choice to AI—one sitting

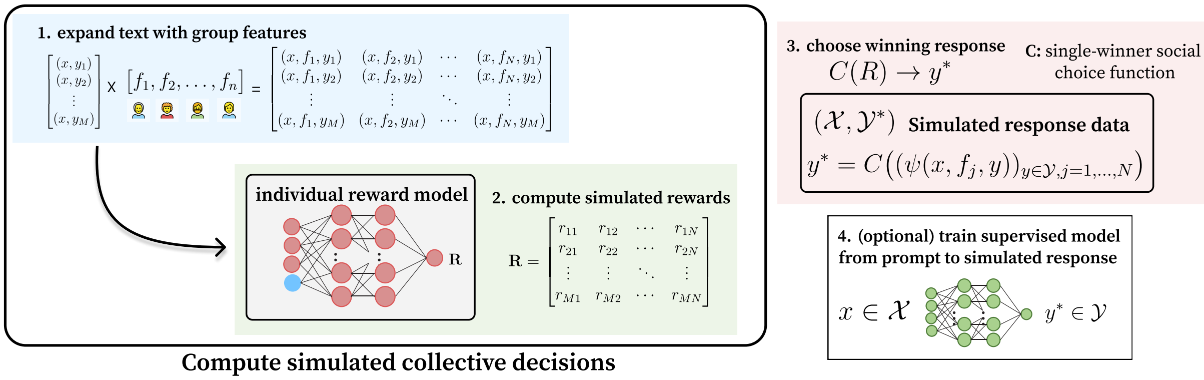

Figure 1: **Supervised Learning from Simulated Collective Decisions**. We show that with an individual or cardinal reward model, as presented in Figure 4, responses y to a prompt x can be simulated. This process expands the scope of studying preferences within RLHF and opens future work on personalization and other topics.

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 above (in which 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 to Navigate a Multiplicity of AIs?

Consider the example of a group of people voting over the restaurant where they will go for dinner. If there is significant disagreement in the votes, rather than forcing a minority to go to a restaurant that they really do not like, it can make sense to split the people into multiple groups, each going to their own 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 and Carman, 2024]. Depending on the situation, the people providing feedback might be split into groups *ex ante* (for example, country A makes one system based on the feedback of A's citizens and country B another based on the feedback of B's citizens), but also *ex post*, where we first collect feedback and then consider which people it makes sense to group together. 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 believes 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, 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 and Oesterheld, 2023] may help keep these kinds of interactions from going horribly wrong. 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.

9 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 and effects of using social choice for these purposes. Needless to say, the questions considered above are multifaceted and, as such, cannot be adequately addressed without complementary (not necessarily AI-specific) research. How best to make practical decisions, as well as associated legal and political considerations, provide further important avenues for future research.

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.

References

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

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

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

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

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The Moral Machine experiment. *Nature*, 563(7729):59–64, November 2018.

Haris Aziz, Markus Brill, Vincent Conitzer, Edith Elkind, Rupert Freeman, and Toby Walsh. Justified representation in approval-based committee voting. *Social Choice and Welfare*, 48(2):461–485, 2017.

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback, 2022. arXiv:2212.08073.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

Michel Balinski and Rida Laraki. *Majority Judgement: Measuring, Ranking and Electing*. MIT Press, Boston, 2010.

Steven J. Brams and Peter C. Fishburn. Voting procedures. In Kenneth J. Arrow, Amartya K. Sen, and Kotaro Suzumura, editors, *Handbook of Social Choice and Welfare*, volume 1, pages 173–236. North-Holland, Amsterdam, 2002.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia. *Handbook of Computational Social Choice*. Cambridge University Press, 2015.

Felix Brandt. Rolling the dice: Recent results in probabilistic social choice. In Ulle Endriss, editor, *Trends in Computational Social Choice*, pages 3–26. AI Access, 2017.

Donald E. Campbell and Jerry S. Kelly. Impossibility theorems in the Arrovian framework. In Kenneth J. Arrow, Amartya K. Sen, and Kotaro Suzumura, editors, *Handbook of Social Choice and Welfare*, volume 1, pages 35–94. North-Holland, Amsterdam, 2002.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Louis Castricato, Nathan Lile, Suraj Anand, Hailey Schoelkopf, Siddharth Verma, and Stella Biderman. Suppressing pink elephants with direct principle feedback, 2024.

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

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

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

Vincent Conitzer, Toby Walsh, and Lirong Xia. Dominating manipulations in voting with partial information. In

Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI-11), pages 638–643. AAAI Press, 2011.

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson and Thore Graepel. Cooperative AI: machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

Claude d'Aspremont and Louis Gevers. Social welfare functionals and interpersonal comparability. In Kenneth J. Arrow, Amartya K. Sen, and Kotaro Suzumura, editors, *Handbook of Social Choice and Welfare*, volume 1, pages 459–541. Elsevier Science B.V., 2002.

Franz Dietrich and Christian List. Probabilistic opinion pooling. In Alan Hajek and Christopher Hitchcock, editors, *Oxford Handbook of Philosophy and Probability*. Oxford University Press, Oxford, 2016.

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. Hard choices in artificial intelligence. *Artificial Intelligence*, 300:103555, 2021.

Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. *arXiv preprint arXiv:2310.05344*, 2023.

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

Michael Feffer, Hoda Heidari, and Zachary C Lipton. Moral machine or tyranny of the majority? *arXiv preprint arXiv:2305.17319*, 2023.

Sara Fish, Paul Gölz, David C Parkes, Ariel D Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. Generative social choice. *arXiv preprint arXiv:2309.01291*, 2023.

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

Bailey Flanigan, Paul Gölz, Anupam Gupta, Brett Hennig, and Ariel D. Procaccia. Fair algorithms for selecting citizens' assemblies. *Nature*, 596:548–552, 2021.

Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P. Dickerson, and Vincent Conitzer. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, 283(103261), 2020.

Rachel Freedman, Rohin Shah, and Anca Dragan. Choice set misspecification in reward inference. *arXiv preprint arXiv:2101.07691*, 2021.

Rachel Freedman, Justin Svegliato, Kyle Wray, and Stuart Russell. Active teacher selection for reinforcement learning from human feedback. *arXiv preprint arXiv:2310.15288*, 2023.

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

D. Ganguli 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-inp retrieved 2024-01-31.

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

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

Joshua Greene, Francesca Rossi, John Tasioulas, Kristen Brent Venable, and Brian C. Williams. Embedding ethical principles in collective decision support systems. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 4147–4151, Phoenix, AZ, USA, 2016.

Davide Grossi and Gabriella Pigozzi. *Judgment Aggregation: A Primer*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer Cham, 1 edition, 2022.

Daniel Halpern, Gregory Kehne, Ariel D. Procaccia, Jamie Tucker-Foltz, and Manuel Wüthrich. Representation with incomplete votes. In Brian Williams, Yiling Chen, and Jennifer Neville, editors, *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*, pages 5657–5664. AAAI Press, 2023.

Wesley H. Holliday and Eric Pacuit. Strategic voting under uncertainty about the voting method. In Larry S. Moss, editor, *Theoretical Aspects of Rationality and Knowledge: Proceedings of the 2019 Conference (TARK 2019)*, volume 297 of *Electronic Proceedings in Theoretical Computer Science*, pages 252–272. EPTCS, 2019.

Wesley H. Holliday and Eric Pacuit. Arrow's decisive coalitions. *Social Choice and Welfare*, 54:463–505, 2020.

Joey Hong, Kush Bhatia, and Anca Dragan. On the sensitivity of reward inference to misspecified human models. *arXiv preprint arXiv:2212.04717*, 2022.

Daniel Jarrett, Miruna Pislar, Michiel A Bakker, Michael Henry Tessler, Raphael Koster, Jan Balaguer, Romuald Elie, Christopher Summerfield, and Andrea Tacchetti. Language agents as digital representatives in collective decision-making. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

Jerry S. Kelly. *Social Choice Theory: An Introduction*. Springer, Berlin, 1988.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 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.

Martin Lackner and Piotr Skowron. *Multi-Winner Voting with Approval Preferences*. SpringerBriefs in Intelligent Systems. Springer Cham, 2023.

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

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

Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. The history and risks of reinforcement learning and human feedback. *arXiv preprint arXiv:2310.13595*, 2023.

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

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

Jean-François Laslier and M. Remzi Sanver, editors. *Handbook on Approval Voting*. Studies in Choice and Welfare. Springer Berlin Heidelberg, 1 edition, 2010.

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

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

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

Duncan C McElfresh, Lok Chan, Kenzie Doyle, Walter Sinnott-Armstrong, Vincent Conitzer, Jana Schaich Borg, and John P Dickerson. Indecision modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5975–5983, 2021.

Iain McLean and Arnold Urken, editors. *Classics of Social Choice*. The University of Michigan Press, Ann Arbor, 1995.

Xiangrui Meng. 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.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. More human than human: measuring chatgpt political bias. *Public Choice*, 198:3–23, 2023.

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

Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Samir Okasha. 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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Eric Pacuit. Voting methods. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2019 edition, 2019.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International Conference on Machine Learning*, pages 26837–26867. PMLR, 2023.

Alois Paulin. An overview of ten years of liquid democracy research. In *Proceedings of the 21st Annual International Conference on Digital Government Research*, pages 116–121, 2020.

Uwe Peters and Mary Carman. Cultural bias in explainable ai research: A systematic analysis. *J. Artif. Int. Res.*, 79, mar 2024.

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

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*, 2021.

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

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

Tuomas Sandholm and Craig Boutilier. Preference elicitation in combinatorial auctions. In Peter Cramton, Yoav Shoham, and Richard Steinberg, editors, *Combinatorial Auctions*, chapter 10, pages 233–263. MIT Press, 2006.

Mark Satterthwaite. 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.

Thomas Schwartz. *Cycles and Social Choice: The True and Unabridged Story of a Most Protean Paradox*. Cambridge University Press, 3 2018.

Rohin Shah, Pedro Freire, Neel Alex, Rachel Freedman, Dmitrii Krasheninnikov, Lawrence Chan, Michael D Dennis, Pieter Abbeel, Anca Dragan, and Stuart Russell. Benefits of assistance over reward learning. In *NeurIPS Workshop on Cooperative AI*, 2020.

Archit Sharma, Sedrick Keh, Eric Mitchell, Chelsea Finn, Kushal Arora, and Thomas Kollar. A critical evaluation of ai feedback for aligning large language models, 2024.

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

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Alan D. Taylor. *Social Choice and the Mathematics of Manipulation*. Cambridge University Press, Cambridge, 2005.

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

Victor Uc-Cetina, Nicolas Navarro-Guerrero, Anabel Martin-Gonzalez, Cornelius Weber, and Stefan Wermter. Survey on reinforcement learning for language processing. *Artificial Intelligence Review*, 56(2):1543–1575, 2023.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *arXiv preprint arXiv:2306.01693*, 2023.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

William S. Zwicker. Introduction to the theory of voting. In Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia, editors, *Handbook of Computational Social Choice*, pages 23–56. Cambridge University Press, New York, 2016.

Figures

4	4	9	4	2	
A	A	B	C	C	Borda Count: CBA
B	C	C	A	B	Instant Runoff: ABC
C	B	A	B	A	Ranked Pairs: BCA

Figure 2: Individual rankings on the left (4 voters submit the ranking ABC, 4 submit ACB, etc.) lead to different aggregated rankings on the right, depending on the aggregation rule. Borda Count gives an alternative 0 points for each voter who ranks it last, 1 point for each voter who ranks it second, and 2 points for each voter who ranks it first; alternatives are then ordered by descending score. Instant Runoff ranks C last since C has the fewest first-place rankings; then, after removing C from all voters' rankings, B has the fewest first-place rankings, so B is in second and A is in 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; but the smallest majority margin of victory is for A over B, so we reverse this majority preference, yielding BCA.

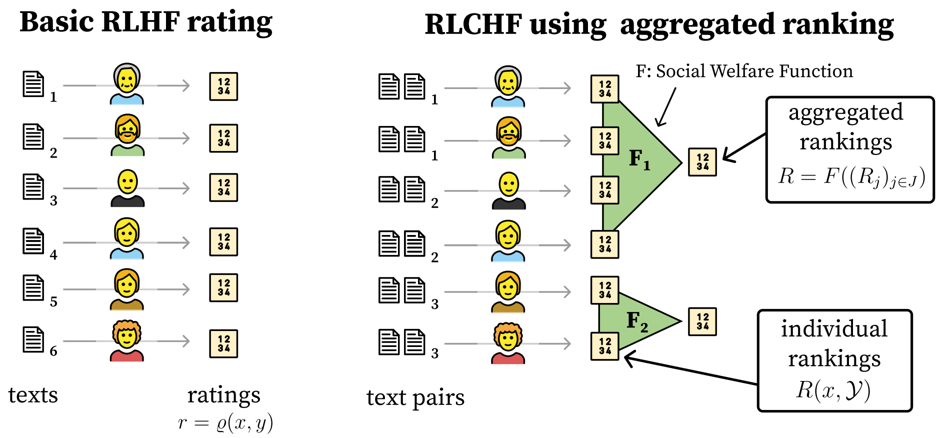

Figure 3: **RLCHF using aggregated rankings**. The core addition to the standard RLHF process is the call-out of an explicit social welfare function, F, which determines how preferences are aggregated.

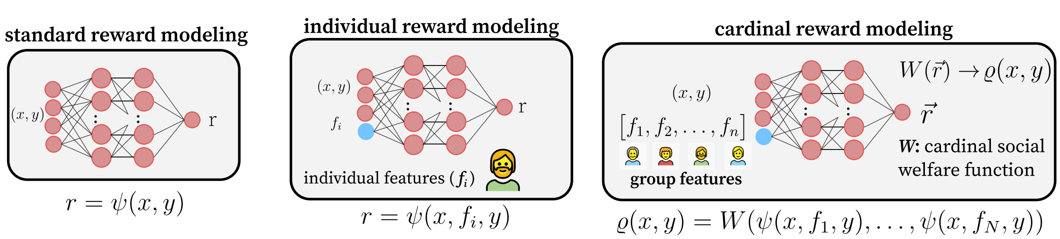

Figure 4: **RLCHF using evaluator features and aggregated ranks**. We show how an individuals' features can be used as an additional input to reward models within the RLHF process.