AI Alignment From Social Choice Perspectives

DANIEL HALPERN
Google Research
EVI MICHA
University of Southern California
ARIEL D. PROCACCIA
Harvard University
BENJAMIN SCHIFFER
Harvard University
ITAI SHAPIRA
Harvard University
and
SHIRLEY ZHANG
Harvard University

Alignment from human feedback uses human judgments about model outputs to steer the behavior of language models after pretraining. When those judgments reflect conflicting views of desirable behavior, the learned objective becomes an aggregate determination of what the model should prefer. We survey recent work that has studied this aggregation problem through the lens of social choice theory. We illustrate how the social choice perspective helps identify failure modes in the feedback aggregation layer and reveals a broader design space for handling disagreement in explicit and principled ways.

1. INTRODUCTION

AI alignment is the problem of ensuring that artificial intelligence systems act in ways consistent with human intentions, preferences, and normative constraints [1–4]. Achieving alignment requires identifying which outputs and behavioral patterns are desirable or unacceptable, and using those judgments to control and steer the system [5–7].

For open-ended language models, operationalizing these goals presents a difficult challenge. The full range of desirable behaviors cannot be explicitly specified [8]; the underlying human judgments rely on tacit knowledge and subtle tradeoffs between objectives, making them virtually impossible to capture in formal terms [1, 9].

Fortunately, humans are far better at recognizing acceptable behavior than they are at formally articulating it [10, 11], an asymmetry that motivates a shift from explicit specification to learned evaluation [12]. In methods for *alignment from human feedback*, most notably *reinforcement learning from human feedback (RLHF)* [12–16],

Authors are listed alphabetically. Correspondence to Itai Shapira (itaishapira@g.harvard.edu).

Under review for SIGecom Exchanges.

human annotators evaluate concrete model outputs, and a scoring function trained on this feedback is then used to align the model.[1] Typically, annotators express preferences via pairwise comparisons, selecting the better of two responses to a given prompt. Aggregated across users and contexts, the annotations are distilled into a learned reward model: a parametric scoring function trained to predict the desirability of a given output. This reward model is expected to generalize beyond the labeled training data, capturing implicit norms and tradeoffs among competing objectives [20].

The foregoing methods compress human preferences about model behavior into a single, universally applicable scalar reward signal that ostensibly represents human judgment. This approach implicitly assumes a shared, underlying human intuition that can be statistically recovered by querying human evaluators [21–23]. Under this view, annotators are interchangeable. Conflicting preferences are treated as noisy observations of a common ground truth, rather than evidence of value pluralism worth reflecting in the system [22, 24]. While such methods have proven empirically successful on tasks where evaluators largely agree [13, 15, 25], in many contexts, interpretations of "correct" behavior diverge across backgrounds and cultures [26–32].

When annotators inherently disagree, reward modeling goes beyond statistical estimation; it becomes a form of preference aggregation. It collapses conflicting individual judgments into a single collective preference, implicitly fixing tradeoffs among competing values while obscuring the mechanism used to resolve them [33]. Evaluating this choice requires examining the assumed model structure and the aggregation method itself. These questions cannot be resolved within the training objective alone. Instead, they fall squarely within the domain of *social choice theory* [34–41], the mathematical study of how heterogeneous individual preferences are aggregated into collective decisions.

Viewed through this lens, alignment can be analyzed as a formal aggregation pipeline: from elicited judgments to learned objectives to optimized policies. Social choice provides the precise language to express desired properties explicitly, making the normative assumptions embedded in aggregation procedures more transparent and open to mathematical comparison.

In this paper, we survey three closely related types of contributions the social choice perspective has brought to recent alignment research. First, this framing identifies the aggregation rules embedded in widely used methods such as RLHF and DPO [16], showing which assumptions and normative priorities these methods build into their objectives [39, 42]. Second, social choice axioms expose structural failure modes of such rules, identifying conditions under which alignment methods provably violate desirable aggregation properties. Third, social choice provides a broad toolbox of well-studied aggregation methods, informing new approaches that encode different normative objectives through various choices of feedback elicitation, aggregation, and policy optimization.

[1] Constitutional AI [17] and its variants inherit the same framing one level up, as the "constitution" is itself the product of preference aggregation over normative principles [18], and discretion must still be exercised over their conflicts [19].

Yet thinking of alignment from human feedback purely in terms of classical social choice misses two distinctive aspects of alignment that shape much of the research surveyed below. The first is *generalization*. Because the candidate-response space is generated and effectively unbounded, the feedback data cover only a sparse set of comparisons over prompt-response pairs, provided by a small subset of evaluators. The learned preference model must then generalize in three directions: from evaluated responses to unseen responses, from curated prompts to unseen prompts, and from sampled annotators to a broader population of users [43].

The second feature is that alignment is ultimately evaluated on the *downstream policy* [44, 45]. The intermediate learned reward model matters only insofar as optimizing against it produces a policy aligned with human judgments. A reward model that satisfies desirable aggregation properties need not induce a policy that does the same.

The goal, then, is to adapt what social choice has long understood about collective decision-making into the practical machinery of model training. AI alignment, in this light, is not merely the problem of steering AI to follow human values, but rather of establishing fair principles for incorporating the diversity of values people actually hold [3, 46].

The sections that follow trace this adaptation. We begin with the classical aggregation rule implicit in standard reward learning, then examine how that rule changes under the distinctive constraints of alignment, including generated candidate responses, sparse elicitation, parametric generalization, welfare loss under optimization, and policy-level approaches that avoid scalar rewards.

Survey Roadmap. Section 2 introduces the notation used throughout the survey. Section 3 then identifies the aggregation rule implicit in unconstrained Bradley–Terry reward learning, showing that it behaves like the classic *Borda rule*.

The next two sections examine failure modes that arise at different points in the reward-learning pipeline. Section 4 examines clone robustness, an inherited social-choice pathology that is especially natural in language-model settings, where generated candidate sets may contain many nearby strings expressing the same substantive answer, making the learned reward sensitive to how that answer is represented in the sample. Section 5 turns to failures of *unanimity* that arise not from the aggregation rule itself but from the way the reward function is learned from data, in particular from restricting the reward to a limited reward class.

The following sections move from aggregation to the information and policy-level limits of alignment from feedback. Section 6 reframes the analysis around *utilitarian welfare*, quantifying how much utility the optimized policy can lose relative to what would be achievable if the underlying cardinal preferences were observed. It also asks what sparse feedback can identify about the population preference distribution, showing that sparse pairwise feedback can leave unidentified how preferences co-vary across pairs within individuals, while slightly richer elicitation can recover much more of that structure. Section 7 then considers policy-level methods that avoid first compressing pairwise preferences into a scalar reward. It explains how *Nash learning from human feedback* optimizes directly against pairwise preferences and

achieves optimal welfare guarantees. We conclude in Section 8 by discussing related directions at the boundary between social choice and alignment.

2. RLHF PRIMER AND NOTATION

This section fixes the notation used throughout the paper and reviews the standard RLHF pipeline at the level needed for our analysis. We first define prompts, responses, annotators, policies, and reward models, then describe how pairwise preference data are aggregated into a scalar reward via *random utility models* and the Bradley–Terry loss, and finally recall the KL-regularized policy objective used to optimize a language model against the learned reward.

Setup and Notation. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the prompt and response spaces, respectively. A prompt x may encode either a single query or the dialogue history of a multi-turn interaction. Let $\mathcal{A}$ denote the target population of annotators, and let $i \sim \mathcal{A}$ denote a randomly sampled annotator. A stochastic policy π assigns to each prompt $x \in \mathcal{X}$ a distribution over responses, written $\pi(\cdot \mid x) \in \Delta(\mathcal{Y})$. We let π_{base} denote a fixed reference policy, typically a pretrained model that has been supervised fine-tuned. A reward model is a function $r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ assigning a real-valued score to each prompt-response pair. When considering a parameterized family of reward models, we write r_θ for the model indexed by $\theta \in \Theta$. When the prompt x is fixed or clear from context, we write $r(y)$ as shorthand for $r(x, y)$, and similarly write $r_\theta(y)$ for $r_\theta(x, y)$. For analyses restricted to a fixed prompt x, let $\mathcal{Y}_x \subseteq \mathcal{Y}$ denote the set of candidate responses considered for that prompt. Finally, let D denote the distribution over prompts used for post-training and policy evaluation.

Preference Data and Reward Learning. Alignment from human feedback can use several forms of supervision, including demonstrations, scalar ratings, critiques, natural-language feedback, and comparisons [47]. In the reward-modeling stage of standard RLHF pipelines, the data are typically comparison labels: for a prompt x and two candidate responses $y, y' \in \mathcal{Y}$, an annotator indicates which response they prefer [13, 15]. Rankings over a slate of responses are often converted into induced pairwise comparison labels [13, 48]. We treat these labels as samples from a population preference relation and write

$$P_x(y \succ y') \coloneqq \Pr_{i \sim \mathcal{A}}[y \succ_i y'] \in [0, 1].$$

That is, $P_x(y \succ y')$ is the population probability that a randomly drawn annotator prefers y to y' on prompt x. Most methods considered in this survey take only this pairwise preference object as input.[2] At times, we view a sampled annotator $i \sim \mathcal{A}$ as having an unobserved implicit reward function $r_i(x, \cdot)$ over responses, which induces the preference relation $\succ_i$.

Reward learning fits a scalar function r_θ whose induced pairwise probabilities approximate these population probabilities. In standard alignment pipelines [12–15], this is done by fitting a parametric reward model together with model-implied pairwise probabilities of the *random utility model (RUM)* form $P_x^\theta(y \succ y') =$

[2] In social choice, such rules are often called *C2 rules* [49].

$F(r_\theta(x,y) - r_\theta(x,y'))$, where $F : \mathbb{R} \to (0,1)$ is an increasing link function satisfying $F(t) = 1 - F(-t)$ [50–52]. The widely used *Bradley–Terry (BT) model* [53] defines F as the sigmoid function, $\sigma(t) \coloneqq (1+e^{-t})^{-1}$.

Concretely, let $\mathcal{D}$ be a dataset of pairwise comparisons, where each triple $(x, y^+, y^-) \in \mathcal{D}$ has $x \in \mathcal{X}$, $y^+, y^- \in \mathcal{Y}$, and records that an annotator preferred y^+ to y^- on prompt x. For a parametric reward model r_θ, BT training minimizes the following logistic loss:

$$\mathcal{L}_{\mathrm{BT}}(\theta) = \sum_{(x,y^+,y^-)\in\mathcal{D}} \log\Big(1 + \exp\big(-(r_\theta(x,y^+) - r_\theta(x,y^-))\big)\Big). \tag{1}$$

This loss function is the negative log-likelihood of the observed preference data under the assumption that preferences satisfy $P_x(y^+ \succ y^-) = \sigma\left(r(y^+) - r(y^-)\right)$. When this holds, we say that the BT model is *correctly specified*. This corresponds to assuming that preferences are generated by adding independent Gumbel noise to an implicit reward and choosing the response with the larger noisy reward [54, 55]. Different noise structures lead to different objective functions [56].

KL-Regularized RLHF. During post-training, the learned reward function provides scalar feedback to the language model to optimize against. In the classical formulation [14], the policy objective is given by:

$$\max_{\pi}\ \mathbb{E}_{x\sim D}\left[\mathbb{E}_{y\sim\pi(\cdot|x)}\big[r(x,y)\big] - \beta^{-1}\mathrm{KL}\big(\pi(\cdot\mid x)\,\|\,\pi_{\mathrm{base}}(\cdot\mid x)\big)\right]. \tag{2}$$

For later use, define

$$\mathrm{KL}_D(\pi\,\|\,\pi_{\mathrm{base}}) \coloneqq \mathbb{E}_{x\sim D}\big[\mathrm{KL}\big(\pi(\cdot\mid x)\,\|\,\pi_{\mathrm{base}}(\cdot\mid x)\big)\big].$$

Here β is the tilt strength (inverse temperature). The second term keeps the new policy from drifting too far from the reference policy. We interpret β as a tunable parameter for this drift: larger β pushes $\pi(\cdot\mid x)$ more aggressively toward high-reward responses and further away from $\pi_{\mathrm{base}}(\cdot\mid x)$. Under KL-regularized policy optimization, reward gaps determine how probability mass is reallocated across responses. Thus, aggregation errors that appear small at the reward-learning stage can lead to substantially different policy behavior after optimization.

3. REWARD LEARNING AS IMPLICIT AGGREGATION

When human feedback is heterogeneous, different annotators can give conflicting preference labels for the same prompt. Fitting a single reward then requires the objective to reconcile these judgments into one score per response. BT reward learning is usually read as statistical estimation of latent utilities from noisy pairwise comparisons. However, this framing obscures the reconciliation process, treating it as a technical detail rather than a choice of aggregation rule with inherent normative tradeoffs. This section asks what rule BT reward learning implicitly applies, and shows it to be closely connected to a classical voting rule.

The *Borda rule*, also known as Borda count, is named after Jean-Charles de Borda, who (re)introduced it in the 18th century [57, 58]. In our setting, fix a prompt x and let D_x be the distribution over responses used to generate comparison candidates.

Let each annotator induce a preference relation $\succ_i$ over responses in the support of D_x. The *Borda score* of a response y is the population probability that y is preferred to a randomly drawn comparison candidate:[3]

$$\mathrm{Borda}(y) \coloneqq \mathbb{E}_{y' \sim D_x}\big[P_x(y \succ y')\big]. \tag{3}$$

That is, Borda favors responses that maximize their expected pairwise advantage against a randomly drawn candidate.

In practice, the underlying aggregation rule of BT reward learning is obscured by two structural constraints: the learner observes only a sparse, finite sample of comparisons, and it must fit a parametric reward function to generalize across distinct prompts and responses. To isolate the aggregation step, we remove these constraints and consider an idealized setting. Assume the true population pairwise preference probabilities are known for all pairs in the candidate set $\mathcal{Y}_x$, and allow each response $y \in \mathcal{Y}_x$ to receive an independent, arbitrary scalar reward. This unrestricted objective represents the asymptotic limit of empirical BT learning given infinite comparisons and a fully expressive reward class. Under these conditions, the BT objective has a clean social-choice characterization: its optimizer ranks responses exactly by their Borda scores.

Theorem 1 [59]. *Fix a prompt x, and let D_x be the distribution over responses used to generate comparison candidates. Let $r^\star$ be a finite-valued minimizer of the unrestricted prompt-level population analogue of Equation 1, where comparison pairs are drawn independently from D_x. Then, for any $y, y' \in \mathrm{supp}(D_x)$,*

$$\mathrm{Borda}(y) > \mathrm{Borda}(y') \quad \Longleftrightarrow \quad r^\star(y) > r^\star(y')\,.$$

The lineage of this result is long and somewhat fragmented. The identity in Equation 4 appears already in the work of Zermelo [60], was independently rediscovered by Bradley and Terry [53] and Ford [61], and was restated by Daniels [62] and Jech [63], all assuming that the pairwise preference probabilities are generated by a BT model. Anderson et al. [59] state the result in the form closest to ours. Siththaranjan et al. [42] restate it in the RLHF setting.

Proof of Theorem 1. For any y with $D_x(y) > 0$, differentiating the BT objective with respect to $r(y)$ gives

$$\frac{\partial \mathcal{L}}{\partial r(y)}(r) = D_x(y)\mathbb{E}_{z \sim D_x}\Big[\sigma(r(y) - r(z)) - P_x(y \succ z)\Big].$$

Thus first-order optimality at $r^\star$, together with $D_x(y) > 0$, gives

$$\mathrm{Borda}(y) = \mathbb{E}_{z \sim D_x}\Big[\sigma(r^\star(y) - r^\star(z))\Big]. \tag{4}$$

Let $f(t) = \mathbb{E}_{z \sim D_x}[\sigma(t - r^\star(z))]$. By Equation 4, $\mathrm{Borda}(y) = f(r^\star(y))$ for every y with $D_x(y) > 0$. Since $\sigma(t - r^\star(z))$ is strictly increasing in t for every z, the function

[3] If ties are allowed, one can replace $P_x(y \succ y')$ by $P_x(y \succ y') + \frac{1}{2}P_x(y \sim y')$. If D_x assigns positive probability to $y' = y$, this convention also treats self-comparisons as ties.

f is strictly increasing. Hence, for all y, y' with $D_x(y), D_x(y') > 0$,

$$\text{Borda}(y) > \text{Borda}(y') \iff f(r^\star(y)) > f(r^\star(y')) \iff r^\star(y) > r^\star(y')\,. \qquad \square$$

Theorem 1 clarifies what kind of disagreement BT reward learning preserves when it approximates Borda-style aggregation. Intuitively, Borda favors breadth of acceptability rather than depth of support or pairwise dominance [58]. A response can receive a high score because it is consistently acceptable against many candidate responses, even when another response has stronger support in a particular direct comparison. In this sense, Borda selects the consensus or compromise response, the one no annotator subgroup strongly objects to, over a response backed by a narrower majority, and it downranks polarizing responses [64, 65]. The classic critique is that broad mid-strength acceptability can systematically advantage bland responses over distinctive ones that elicit both strong support and strong opposition in the annotator population [66].

After the reward model is learned, post-training converts this score into a change in the model's response distribution, shifting probability mass toward responses with higher Borda scores. How strongly any particular response is upweighted depends on both the Borda score learned from annotators and the base policy π_{base}. To make this concrete, we return to the idealized setting and assume that fine-tuning attains the unparameterized optimum, where the decision variable for each prompt x is the conditional distribution $\pi(\cdot \mid x)$ itself. The maximizer then has the closed-form Boltzmann/Gibbs expression [67, 68] $\pi^\star(y \mid x) \propto \pi_{\text{base}}(y \mid x)\exp\big(\beta\, r(y)\big)$, from which it follows that [54, 69]:

$$\text{Borda}(y) > \text{Borda}(y') \iff \frac{\pi^\star(y \mid x)}{\pi_{\text{base}}(y \mid x)} > \frac{\pi^\star(y' \mid x)}{\pi_{\text{base}}(y' \mid x)}.$$

This implies that the post-training density ratio against the base policy is, up to a monotone transformation, the π_{base}-weighted Borda score of y.

In practice, many high-profile deployments of alignment intentionally flatten P_x into binary preference labels by taking the majority vote, disregarding the soft-label information as noise [15, 17, 70–75]. Applied before BT fitting, this preprocessing induces the *Copeland rule* [76], which ranks responses by the number of pairwise majority contests they win [77, 78].

4. CLONE ROBUSTNESS

Section 3 identified BT reward learning as a Borda-like aggregation rule on a fixed candidate set. This connection suggests that BT reward learning may inherit some of Borda's pathologies. A particularly relevant one for language-model reward learning is sensitivity to near-duplicate candidates. In voting theory, this means that adding candidates nearly identical to an existing candidate can change the rule's outcome [79]. Since Borda is well known to be sensitive to such near-duplicate candidates, *robustness to approximate clones* becomes a natural first benchmark for BT reward learning.

Approximate clones arise structurally in generated language [36, 80]. For a fixed prompt, candidate responses are sampled from a generative model whose probability

mass often concentrates around a small number of semantic and stylistic modes. A single substantive answer can therefore appear in many forms, with differences that are largely surface-level, such as paraphrasing, reordered explanations, or changes in verbosity. The candidate set $\mathcal{Y}_x$ is only a sampled discretization of this broader response space, and the often arbitrary number of variants associated with a given substantive answer reflects the sampling and filtering process as much as human preference. Exact copies can often be extracted and merged before training, but the harder case involves near-duplicates: responses that differ slightly while occupying the same semantic region and expressing the same underlying answer.

Within a near-clone cluster, pairwise comparisons among variants provide little new signal. Annotators may exhibit minor idiosyncratic preferences among them, but these differences are usually not the underlying quality the reward model is meant to capture. A clone-sensitive BT model can nevertheless treat the number of variants in the cluster as meaningful, so adding near-duplicates can change the rewards assigned to individual candidates and the total policy probability assigned to the cluster as a whole. The learned reward, and by extension the policy it induces, can then depend on arbitrary sampling artifacts, such as how densely different answers are represented in $\mathcal{Y}_x$, rather than on human preference [81]. As models generate more fluent paraphrases and subtle stylistic variants of the same substantive answer, this representational multiplicity can become a larger source of reward variation.

The corresponding formal requirement is *robustness to approximate clones*: the learned reward should be stable under the addition of near-duplicates. Adding a response that is nearly identical to an existing one should make the two responses receive nearly the same reward, and should not substantially change the rewards assigned to unrelated responses. Procaccia et al. [82] formalize this requirement using a metric on responses. Let ρ be a metric on $\mathcal{Y}$, where smaller values indicate greater similarity. Fix a prompt x and a finite candidate set $\mathcal{Y}_x \subseteq \mathcal{Y}$. Let $r_{\mathcal{Y}_x}$ denote the reward learned from comparisons over $\mathcal{Y}_x$, for example by minimizing the unrestricted prompt-level population analogue of Equation 1:

Definition 2 Robustness to approximate clones. A reward-learning procedure is *robust to approximate clones* if for every $\delta > 0$ there exists $\varepsilon > 0$ such that the following holds. For any finite candidate set $\mathcal{Y}_x \subseteq \mathcal{Y}$, any $y \in \mathcal{Y}_x$, and any new response y' satisfying $\rho(y, y') \le \varepsilon$, the rewards learned before and after adding y' satisfy

$$\left| r_{\mathcal{Y}_x \cup \{y'\}}(y) - r_{\mathcal{Y}_x \cup \{y'\}}(y') \right| \le \delta$$

and

$$\left| r_{\mathcal{Y}_x \cup \{y'\}}(z) - r_{\mathcal{Y}_x}(z) \right| \le \delta \qquad \forall z \in \mathcal{Y}_x.$$

The first condition says that the original response and its approximate clone should receive nearly the same reward once both are present. The second says that adding the clone should not substantially change the rewards assigned to the pre-existing responses. Together, the two conditions require the reward-learning rule to treat near-duplicates as redundant representations of the same local region.

Using the connection to Borda, Procaccia et al. [82] show that standard BT reward

learning is not robust to approximate clones. They then propose a weighted version of the population BT objective that discounts redundant representatives of the same local region. The idea is to assign each response a uniqueness weight $w(y)$, measuring the normalized mass of points in the response space for which y is the nearest candidate. Responses in crowded regions receive smaller weight, while responses representing larger regions receive larger weight. Applying these weights to pairwise comparisons yields the weighted BT objective in the following theorem.

THEOREM 3 [82]. *Let $\mathcal{S} \subseteq \mathbb{R}^d$ be a compact response space with finite positive volume. Fix a prompt x, a finite candidate set $\mathcal{Y}_x \subseteq \mathcal{S}$, and $\lambda > 0$. Draw a point s uniformly at random from $\mathcal{S}$, and assign it to one of the nearest candidates in $\mathcal{Y}_x$, breaking ties uniformly at random; let $w_{\mathcal{Y}_x}(y)$ denote the resulting weight of candidate y. Define the weighted BT loss*

$$\mathcal{L}^w_\lambda(r) := -\sum_{\substack{y,y' \in \mathcal{Y}_x \\ y \neq y'}} w_{\mathcal{Y}_x}(y) w_{\mathcal{Y}_x}(y') P_x(y \succ y') \log \sigma(r(y) - r(y')) + \frac{\lambda}{2} \sum_{y \in \mathcal{Y}_x} w_{\mathcal{Y}_x}(y) r(y)^2 .$$

If $P_x(y \succ y')$ is Lipschitz continuous in the responses, then the reward-learning algorithm that minimizes $\mathcal{L}^w_\lambda(r)$ is robust to approximate clones under the ℓ_2 metric.

Theorem 3 should be read as a representation-invariance guarantee for the prompt-level aggregation step. Once an embedding and a metric are fixed, near-duplicate responses are treated as surface representatives of the same region, so adding another does not substantially change the learned rewards. The guarantee reduces the need for exact deduplication. What it does not remove is the dependence on the embedding, metric, and reference space $\mathcal{S}$ that fix the weights. In this sense, the invariance holds relative to a fixed geometry, shifting the design burden from detecting duplicates to specifying when responses count as close.

5. PARAMETRIC REWARD LEARNING AND UNANIMITY

Sections 3 and 4 studied BT reward learning under an unconstrained reward class, where each prompt-response pair has its own free scalar score. In that setting, BT reproduces the Borda score and inherits both its guarantees and its pathologies. In practice, the reward must be implemented by a shared parametric function, so the score assigned to one response is tied to the scores assigned to others through the same hypothesis class. This coupling creates a new source of failures beyond the aggregation effects already discussed.

Formally, the learner chooses parameters θ from a parameter space Θ, which induce a reward function $r_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The possible rewards are therefore restricted to the hypothesis class $\mathcal{R}_\Theta = \{r_\theta : \theta \in \Theta\}$. In the standard BT pipeline, θ is fitted on the comparison dataset $\mathcal{D}$ by minimizing the logistic loss in Equation 1, or equivalently by maximizing the BT likelihood.

The simplest failure is a violation of *unanimity*, also known as *Pareto efficiency*. If every annotator strictly prefers response y^+ to response y^-, the learned reward should assign y^+ a higher score. This property is worth studying not because perfect unanimity is likely to occur in practice, but because violating it means ignoring the

strongest possible consensus signal, falling below the baseline set by every standard voting rule in the classical setting.

Definition 4 Unanimity. A reward-learning procedure satisfies *unanimity* if, for any pair of responses $y^+, y^- \in \mathcal{Y}_x$, whenever every annotator prefers y^+ to y^-, the learned reward satisfies $r_\theta(y^+) > r_\theta(y^-)$.

Ge et al. [39] established this failure for a linear hypothesis class; Hollender and Kraiczy [83] showed that it extends to richer parametric reward classes. We follow the linear setting of Ge et al., which admits the cleanest analysis.

In the linear model, each response $y \in \mathcal{Y}_x$ is associated with a known feature vector $\phi(y) \in \mathbb{R}^d$, where d is the embedding dimension, and the reward is linear in those features, $r_\theta(y) \coloneqq \langle \theta, \phi(y) \rangle$ with $\theta \in \mathbb{R}^d$. The learner holds ϕ fixed and fits only θ. This corresponds to one way of building a reward model, in which ϕ is the embedding obtained by removing the final layer of a pretrained language model and the reward is a linear head trained on preference data on top of it. In many settings, the embedding ϕ is itself parameterized and trained alongside the last layer rather than held fixed. The linear model described here is best read as an analytically convenient special case of the restricted reward class $\mathcal{R}_\Theta$.

Throughout this section, we fix a prompt $x \in \mathcal{X}$ and a candidate set $\mathcal{Y}_x \subseteq \mathcal{Y}$. We assume that the learner observes the complete pairwise preferences of a set of annotators $\mathcal{A}$. That is, for every annotator $i \in \mathcal{A}$ and every pair of distinct responses $y, y' \in \mathcal{Y}_x$, the learner observes which of the two responses i prefers.

We assume that the learner fits a loss function to these observed pairwise preferences. Specifically, the analysis in this section concerns the broader family of loss-based reward-learning rules, of which BT is one member. Given any loss function $\ell : \mathbb{R} \to \mathbb{R}$, define

$$\mathcal{L}(\theta; \ell) \coloneqq \sum_{\substack{y^+, y^- \in \mathcal{Y}_x \\ y^+ \neq y^-}} n_{y^+ \succ y^-} \cdot \ell\left(r_\theta(y^-) - r_\theta(y^+)\right),$$

where $n_{y^+ \succ y^-}$ denotes the number of annotators who prefer y^+ to y^-. When $\ell(z) = \log(1 + \exp(z))$, this is the fixed-prompt version of the BT logistic loss in Equation 1; the corresponding population objective replaces the empirical counts $n_{y^+ \succ y^-}$ by the pairwise preference probabilities $P_x(y^+ \succ y^-)$. Hinge loss gives another instance, and more generally the larger the gap $r_\theta(y^-) - r_\theta(y^+)$ on a comparison the rule disagrees with, the larger the penalty. This class of rules is a natural family that contains BT and matches the loss-minimization framing of modern training pipelines.

In the unrestricted setting of Section 3, BT satisfies unanimity as a consequence of its equivalence with Borda. This makes unanimity appear to be a natural property of BT. Yet, the theorem below shows that this guarantee need not survive once the reward is required to generalize through a restricted reward class $\mathcal{R}_\Theta$.

Theorem 5 [39]. *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a loss function satisfying $\inf_z \ell(z) < \ell(0)$, and suppose that ℓ is either nondecreasing and weakly convex, or strictly convex. Then*

there exist a dimension d, a finite candidate set $\mathcal{Y}_x$, feature vectors $\phi(y) \in \mathbb{R}^d$, and a population for which the induced loss-based linear reward rule fails unanimity. In particular, there are responses $y^+, y^- \in \mathcal{Y}_x$ such that every annotator prefers y^+ to y^-, but the learned reward $r_{\theta^\star}$ ranks them in the opposite order: $r_{\theta^\star}(y^-) > r_{\theta^\star}(y^+)$.

At a high level, the loss must find a single set of parameters that applies to all pairwise comparisons at once. Each comparison places two demands on the fit: assigning the preferred response the higher score, and increasing confidence through a larger margin $r(y^+) - r(y^-)$. In the unrestricted setting, these demands can be satisfied independently, since the relevant reward differences can be adjusted without forcing a change in unrelated comparisons. Under a restricted hypothesis class, they interact because all margins are induced by the same trainable parameters.

The loss then chooses a global fit across all observed comparisons, allocating score separation across comparison directions. Some directions matter more for this fit than others, either because many comparisons point along them or because their feature differences have a larger effect on the loss. A unanimous comparison can still be associated with a weak direction in this global problem. The learned model can then fit the dominant comparison directions while assigning the wrong ordering to the unanimous pair.

6. WELFARE LOSS AND SPARSE ELICITATION

One can view alignment from human feedback as a limited-information approximation procedure. For each prompt, the pipeline aims to select or induce the model behavior that is best for a population of annotators, while observing only a partial description of their preferences. The true objective, however, may depend on unobserved information, such as the preference intensities represented by cardinal values. Here, we examine the loss incurred due to this lack of information, starting with whether there is, in fact, a loss.

Identifiability. Fix a prompt x and a candidate set $\mathcal{Y}_x \subseteq \mathcal{Y}$. Suppose an annotator $i \sim \mathcal{A}$ has a reward function $r_i(\cdot)$ over these responses. A standard goal is to maximize *utilitarian welfare*, defined as the expected reward across annotators. Formally, given a reward profile $\mathbf{r} = (r_i)_{i \in \mathcal{A}}$, the welfare of a response y is

$$W_x^{\mathbf{r}}(y) \coloneqq \mathbb{E}_{i \sim \mathcal{A}} \left[r_i(y) \right].$$

For a policy $\pi(\cdot \mid x) \in \Delta(\mathcal{Y}_x)$, we can similarly define its expected welfare as

$$W_x^{\mathbf{r}}(\pi) \coloneqq \mathbb{E}_{y \sim \pi(\cdot|x)} \left[W_x^{\mathbf{r}}(y) \right].$$

If we had access to these cardinal utilities, we could train our policy to directly maximize welfare. In practice, we only observe pairwise comparisons. Assuming annotators respond according to a link function F, we learn only the aggregate win probabilities:

$$P_x(y \succ y') = \mathbb{E}_{i \sim \mathcal{A}} \left[F\left(r_i(y) - r_i(y') \right) \right].$$

This formulation captures BT responses, where choices are made with probabilities proportional to $\exp(r_i(y))$, as well as deterministic responses, where annotators

simply select the higher-reward option.

Unfortunately, unless F is linear, it is impossible to reliably identify welfare-maximizing candidates from these probabilities alone, even for a single prompt x with just two responses, y and y'. Because a non-linear link function F distorts the magnitude of reward differences, one can construct two distinct reward profiles, $\mathbf{r}$ and $\mathbf{r}'$, that yield identical expected win probabilities but differ in aggregate welfare. Specifically, there exist $\mathbf{r}$ and $\mathbf{r}'$ such that:

$$\mathbb{E}_{i\sim\mathcal{A}}\left[F\left(r_i(y)-r_i(y')\right)\right]=\mathbb{E}_{i\sim\mathcal{A}}\left[F\left(r'_i(y)-r'_i(y')\right)\right]$$

yet

$$\mathbb{E}_{i\sim\mathcal{A}}\left[r_i(y)-r_i(y')\right]\neq\mathbb{E}_{i\sim\mathcal{A}}\left[r'_i(y)-r'_i(y')\right].$$

Consequently, identical pairwise observation data can arise from populations with fundamentally different welfare-maximizing preferences.

Distortion. While *exact* identification of welfare-maximizing candidates is impossible, a learned policy may still achieve approximately high welfare using only the observed comparisons. This guarantee is formalized in social choice through the concept of *distortion* [84, 85].

Assume reward functions are nonnegative and normalized with $0 \le r_i(y) \le 1$ for each annotator i and response y. Let $\mathcal{R}_x(P_x)$ denote the set of normalized reward profiles consistent with the pairwise preference object P_x at prompt x. For a policy $\pi(\cdot \mid x) \in \Delta(\mathcal{Y}_x)$, its utilitarian distortion [84, 86] at x is:

$$\mathrm{Dist}_x(\pi;P_x) \coloneqq \sup_{\mathbf{r}\in\mathcal{R}_x(P_x)} \frac{\max_{\pi'\in\Delta(\mathcal{Y}_x)} W_x^{\mathbf{r}}(\pi')}{W_x^{\mathbf{r}}(\pi)}.$$

Conditional on P_x, this quantity measures the worst-case welfare loss that π can incur over normalized reward profiles that could have induced P_x,[4] where the welfare loss is defined by the ratio above. Since compatibility is defined by the elicited feedback, while the chosen policy is determined by the aggregation rule and policy class, the distortion reflects all three components of the decision-making process together.

Theorem 6 [87]. *Fix a prompt x and a finite response set $\mathcal{Y}_x$ with $|\mathcal{Y}_x| \ge 3$. Suppose annotator comparisons are well specified by a BT model with inverse-temperature parameter η, so that for each annotator $i \in \mathcal{A}$,*

$$\Pr[y \succ_i y' \mid x] = \sigma\left(\eta \cdot (r_i(y) - r_i(y'))\right).$$

In other words, the link function is $F(t) \coloneqq \sigma(\eta \cdot t)$. For each pairwise preference object P_x, let $\pi_0(P_x)$ be the policy that BT reward learning returns once the KL penalty is removed; by Theorem 1, it places all mass on the Borda winner under P_x. Then

$$(1-o(1))\,\eta \;\le\; \sup_{P_x} \mathrm{Dist}_x\left(\pi_0(P_x);P_x\right) \;\le\; O(\eta^2),$$

[4] When $W_x^{\mathbf{r}}(\pi) = 0$, the corresponding ratio is interpreted as ∞.

where P_x ranges over pairwise preference objects generated by a normalized reward profile $\mathbf{r}$ *and a comparison-pair sampling distribution. The upper bound holds for every such P_x; the lower bound is the rate as $\eta \to \infty$.*

The lower bound should be read against the minimax lower bound for the same information model. Gölz et al. [87] show that, when each annotator contributes a single comparison, every rule mapping the comparisons to a policy incurs distortion at least $\left(\frac{1}{2} + o(1)\right)\eta$ on some reward profile. Strikingly, this lower bound is achieved, up to lower-order terms, by a framework known as Nash learning from human feedback. We formally define this method and demonstrate its optimal distortion guarantees in Section 7 (see Theorem 7).

Richer Elicitation. Theorem 6 demonstrates that approximating welfare is possible, though imperfect, at least under BT responses. However, utilitarian welfare is not the only valid objective in social choice. Consider two hypothetical responses y and y': y provides a utility of 2 to all annotators, while y' provides a utility of 3 to half the annotators and 1 to the other half. Utilitarian welfare treats these outcomes as equivalent. By contrast, more egalitarian choices, such as *Nash welfare* — the product of utilities — would strictly prefer y, favoring broad approval over polarizing outcomes that benefit one group at the expense of another. How, then, can we optimize for these alternative choices?

Chidambaram et al. [88] and Ge et al. [89] consider linear social choice models (similar to the one described in Section 5) where responses follow BT and deterministic link functions, respectively. Both show that even in these restricted settings, a single pairwise comparison per annotator cannot reveal which of the two responses a more egalitarian objective should prefer. Without preference intensities, a single comparison cannot separate a universally indifferent population from one that is evenly split between strong opposing preferences; in both, each candidate is preferred equally often across the population. Thus, there is no hope of optimizing more egalitarian welfare functions that would strictly prefer broad-appeal candidates. On the other hand, under mild structural conditions, eliciting two comparisons per annotator yields enough information to completely identify the voter type distribution. This identification enables direct optimization for *any* desired social welfare function.

Similarly, Cherapanamjeri et al. [90] demonstrate in a latent-utility model that moving from pairwise choices to best-of-three queries provides essentially complete identifiability under appropriate structural conditions.

Richer information can also come from passively recorded signals. For example, the time taken to provide a comparison label can carry information about preference intensities that binary labels discard [91, 92].

While these models rely on different structural assumptions, each points to the same fact that slightly richer elicitation can, in principle, make far more welfare objectives optimizable than pairwise comparison allows.

7. DIRECT ALIGNMENT FROM PAIRWISE PREFERENCES

The preceding sections analyzed feedback pipelines that infer a scalar reward from preference data and then optimize a policy against it. Such a reward projects pairwise comparisons among responses onto a single ordered axis. This section asks what information is lost when the aggregate preference relation is not transitive, and how alignment objectives can more directly retain this pairwise structure.

Cyclic aggregate preferences can arise from heterogeneous feedback even when each annotator is internally consistent. Averaging across annotators may produce a collective majority relation with a *Condorcet cycle*, where y_a is preferred to y_b, y_b to y_c, and y_c to y_a [58]. Additional data may estimate the cycle more accurately but cannot make it representable by a scalar reward. In such a case, there is no *Condorcet winner*, meaning no response defeats every other response by majority comparison. Any deterministic target must break the cycle somewhere, motivating alignment objectives that reason over the pairwise relation itself.

Maximal Lotteries. One way to avoid arbitrarily breaking cycles is to select a distribution over responses rather than a single response. Two such distributions can be compared by the expected majority margin between them, which is well-defined whether or not a Condorcet winner exists. This idea is formalized by *maximal lotteries*, a solution concept first introduced by Kreweras [93], developed systematically by Fishburn [94],[5] and later applied in many settings, including AI evaluation [97, 98].

Fix a prompt x and a finite candidate set $\mathcal{Y}_x \subseteq \mathcal{Y}$. Let

$$M_x(y, y') \coloneqq P_x(y \succ y') - P_x(y' \succ y)$$

denote the *majority margin* — the net pairwise preference for y over y' in the population. A maximal lottery is a distribution $p^\star \in \Delta(\mathcal{Y}_x)$ whose expected margin against every response distribution $q \in \Delta(\mathcal{Y}_x)$ is nonnegative:

$$\mathbb{E}_{y\sim p^\star, y'\sim q}[M_x(y, y')] \geq 0.$$

Since M_x is skew-symmetric, the maximal lotteries are exactly the maximin strategies of the zero-sum game with payoff matrix M_x:[6]

$$p^\star \in \arg\max_{p\in\Delta(\mathcal{Y}_x)} \min_{q\in\Delta(\mathcal{Y}_x)} p^\top M_x q = \arg\max_{p\in\Delta(\mathcal{Y}_x)} \min_{q\in\Delta(\mathcal{Y}_x)} \mathbb{E}_{y\sim p,\, y'\sim q}\left[M_x(y, y')\right].$$

Randomization lets the maximal lottery represent disagreement directly. When the majority relation is cyclic, it can place mass on several mutually contesting responses, so the target itself records the unresolved disagreement. When a strict Condorcet winner exists, no mixing is needed and the unique maximal lottery places all mass

[5] The same solution has been rediscovered independently under other names, including the "game theory method" in voting [95] and the von Neumann winner in contextual dueling bandits [96].
[6] Existence follows from von Neumann's minimax theorem [99]. For a generic majority-margin matrix, the maximal lottery is unique, although degenerate ties can yield multiple maximal lotteries. It can be computed efficiently using the standard linear program for the row player's maximin strategy.

on that response. Maximal lotteries also satisfy other desirable axiomatic properties [100]. In particular, they are invariant to *exact* clones (duplicate responses). However, the introduction of *approximate* clones — as defined in Section 4 — can affect the lottery, even if the approximation is arbitrarily precise.

Nash Learning from Human Feedback. *Nash Learning from Human Feedback (NLHF)* [41, 101] applies the same maximin idea directly at the level of policies. Rather than compressing comparisons into a scalar reward and then maximizing it, NLHF keeps the pairwise preference model as the payoff of a two-player game between policies, and trains the policy toward the equilibrium of that game.

The population preference P_x induces a preference between two policies by comparing the responses they generate,

$$P_x(\pi \succ \pi') \coloneqq \mathbb{E}_{y\sim\pi(\cdot|x),\, y'\sim\pi'(\cdot|x)}[P_x(y \succ y')]\,.$$

Thus, $P_x(\pi \succ \pi')$ is the probability that π produces a response preferred to one produced by π', at prompt x. We write

$$P(\pi \succ \pi') \coloneqq \mathbb{E}_{x\sim D}[P_x(\pi \succ \pi')]$$

for the corresponding comparison after averaging over prompts drawn from D. This induces a preference relation over policies: policy π is preferred to policy π' when $P(\pi \succ \pi') > \frac{1}{2}$. The new target of alignment is to find a policy $\pi^\star$ that is preferred to any alternative policy π'. Equivalently, NLHF defines a maximin policy in the induced preference game:

$$\pi^\star \in \arg\max_{\pi} \min_{\pi'} P(\pi \succ \pi').$$

The pair $(\pi^\star, \pi^\star)$ is a Nash equilibrium of this game, which is the sense in which the policy is "Nash." For a fixed prompt with finite candidate support and without regularization, $\pi^\star$ coincides with the maximal lottery of that prompt's margin game.

Because the response space is open-ended and the policy class is parametric, the equilibrium is not computed by the finite linear program for maximal lotteries. It is instead approached by policy optimization against opponents generated from the policy's own iterates. This self-play template, optimizing a policy against its own iterates toward the equilibrium, underlies several recent preference-optimization methods [102–107].

As with RLHF, NLHF can regularize the learned policy toward a reference policy π_{base} with a KL penalty. The regularized preference between policies is

$$P_\tau(\pi \succ \pi') \;=\; P(\pi \succ \pi') - \tau\,\mathrm{KL}_D(\pi \,\|\, \pi_{\text{base}}) + \tau\,\mathrm{KL}_D(\pi' \,\|\, \pi_{\text{base}}),$$

and the training target is the Nash equilibrium of P_τ, which exists and is unique [101]. Viewing each player's payoff separately, a policy receives its preference payoff minus a KL penalty for its own divergence from π_{base}.[7] Once this penalty is present, the correspondence with the maximal lottery is only partial: deployed NLHF converges

[7] The equivalent payoff form is $R(\pi;\pi') = P(\pi \succ \pi') - \tau\,\mathrm{KL}_D(\pi \,\|\, \pi_{\text{base}})$, with the symmetric expression for π'. No feasibility constraint of the form $\mathrm{KL}_D(\pi \,\|\, \pi_{\text{base}}) \le \varepsilon$ is imposed.

to the regularized Nash equilibrium of P_τ, which recovers the maximal lottery only as $\tau \to 0$.

Welfare Interpretation. The welfare-loss perspective of Section 6 gives a precise sense in which the Nash objective is optimal. The following result bounds the worst-case utilitarian distortion of the maximal lottery under anonymous pairwise feedback.

THEOREM 7 [87]. *Suppose annotator comparisons are generated from normalized rewards by the BT model in Theorem 6, with inverse-temperature η. Let $\pi_x^{\mathrm{NLHF}}(P_x)$ be any maximal lottery of the margin game M_x. Then*

$$\sup_{P_x} \mathrm{Dist}_x\left(\pi_x^{\mathrm{NLHF}}(P_x); P_x\right) \le \left(\frac{1}{2} + o(1)\right)\eta,$$

where P_x ranges over pairwise preference objects generated by normalized reward profiles and comparison-pair sampling distributions, and the $o(1)$ term is as $\eta \to \infty$.

Together with the minimax lower bound of Section 6, Theorem 7 shows NLHF attains the smallest possible worst-case distortion, up to lower-order terms, among rules observing anonymous BT pairwise feedback.

8. DISCUSSION

Taken together, the sections above give a social-choice account of alignment from human feedback as a sequence of design choices. Human judgments must be elicited, aggregated into a model, generalized beyond the observed data, and translated into policy behavior. Each step can affect how disagreement is represented in the final system. We conclude by discussing several related research threads that extend this pipeline view beyond the formal results surveyed earlier.

Preference Collapse. Fine-tuning a base policy toward a learned reward signal can narrow the policy's behavior even when the reward itself supports a broader range of responses. With stronger optimization, probability mass can concentrate around a small set of reward-favored outputs, producing *preference collapse* [108], where majority views are further amplified [109, 110] and response diversity is reduced [111–113]. This is one way in which the downstream policy can sharpen the aggregation choices made by the learned reward: biases in the preference data, once encoded in the reward signal, may be amplified by subsequent optimization [114]. As a result, the learned policy may fail to represent the full range of normative considerations expressed across populations, domains, and interaction contexts [115–117].

Multiple Policies. A growing number of proposals advocate for *pluralistic alignment*: modeling multiple perspectives in parallel so as to better capture the breadth of human judgments [21, 30, 118]. Work in this direction differs in where it relaxes the standard single-reward pipeline. Some approaches retain a single learned reward while limiting how aggressively the policy optimizes against it, thereby reducing the policy-level amplification of the aggregate objective [108, 117, 119, 120]. Others enrich the reward model so that it can encode preference heterogeneity more

directly [121–123].

A more direct approach represents divergent preferences by training multiple reward models, each matched to a distinct segment of the population [118, 124–126]. The typical approach clusters annotators, by demographics or inferred preference patterns, and fits a group-specific reward to each cluster. Halpern et al. [22] avoid this intermediate clustering step by fitting a compact reward ensemble whose aggregate pairwise choices match the observed preference P_x, so each component induces a coherent policy while the mixture preserves disagreement.

Once such an ensemble is learned, its components can be deployed in several ways to serve pluralistic goals [21]. For example, the system can present multiple outputs as an Overton-style slate [127], distill them into a consensus statement [128–130], select the policy best matched to a user's stated preference, or sample from the policy mixture. Over repeated use, mixture sampling preserves population-level diversity and counters the tendency of aligned models to recycle a small set of high-reward responses [110, 112, 131], which can be particularly valuable in creative workflows.

Personalization. At the opposite extreme, the alignment objective can be personalized to individual annotators, treating each user as carrying a unique reward function [132]. Since fitting a separate model per user is infeasible under sparse per-user data, these methods place each user's reward in a shared low-dimensional space, recovering a per-user latent variable [133] or a weighting over common reward components [134, 135] from a handful of preferences. Lightweight user embeddings or low-rank adapters then specialize a shared model to the individual. Even so, personalizing to each user risks narrowing their information diet and reinforcing prior views, prompting calls to bound how far personalization should extend [116, 136].

AI for Social Choice and Democracy. The work surveyed in this paper is primarily *social choice for AI*: it uses social choice theory to diagnose how alignment pipelines aggregate heterogeneous feedback, and to design learned rewards or policies that handle disagreement more explicitly. A complementary agenda runs in the reverse direction, asking how AI systems can support collective decision-making, deliberation, and democratic representation. One direction uses AI to help groups deliberate and converge on shared positions, through *Generative Social Choice* [129, 130], deliberation mediators such as the Habermas Machine [137], formal accounts of common ground [128, 138], and platforms for large-scale opinion aggregation [139]. More broadly, progress in social choice for AI and AI for social choice may prove mutually reinforcing, as advances in one direction generate both the concepts and the tools needed to advance the other.

REFERENCES

[1] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable Agent Alignment via Reward Modeling: A Research Direction, 2018. arXiv:1811.07871. 1

[2] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Hua Xu, Aidan O'Gara,

Kwan Ng, Brian Tse, Jie Fu, Stephen McAleer, Yanfeng Wang, Mingchuan Yang, Yunhuai Liu, Yizhou Wang, Song-Chun Zhu, Yike Guo, Yaodong Yang, and Wen Gao. AI Alignment: A Contemporary Survey. *ACM Computing Surveys*, 58(5):132:1–132:38, 2025. 1

[3] Iason Gabriel. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437, 2020. 1, 3

[4] Stuart J. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Allen Lane, 2019. 1

[5] Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. Alignment for Advanced Machine Learning Systems. In S. Matthew Liao, editor, *Ethics of Artificial Intelligence*, pages 342–382. Oxford University Press, 2020. 1

[6] Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. Alignment of Language Agents, 2021. arXiv:2103.14659. 1

[7] Iason Gabriel and Geoff Keeling. A Matter of Principle? AI Alignment as the Fair Treatment of Claims. *Philosophical Studies*, 182(7):1951–1973, 2025. 1

[8] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved Problems in ML Safety, 2021. arXiv:2109.13916. 1

[9] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J. K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*, 2023. 1

[10] Andrew P. Clark, Kate L. Howard, Andy T. Woods, Ian S. Penton-Voak, and Christof Neumann. Why Rate When You Could Compare? Using the "EloChoice" Package to Assess Pairwise Comparisons of Perceived Physical Strength. *PLOS ONE*, 13(1):e0190393, 2018. 1

[11] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-Armed Dueling Bandits Problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012. 1

[12] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In *The Thirty-first Annual Conference on Neural Information Processing Systems*, 2017. 1, 4

[13] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training Language Models to Follow Instructions with Human Feedback.

In *The Thirty-sixth Annual Conference on Neural Information Processing Systems*, 2022. 1, 2, 4

[14] 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, 2020. arXiv:1909.08593. 1, 4, 5

[15] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to Summarize with Human Feedback. In *The Thirty-fourth Annual Conference on Neural Information Processing Systems*, 2020. 1, 2, 4, 7

[16] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. In *The Thirty-seventh Annual Conference on Neural Information Processing Systems*, 2023. 1, 2

[17] 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. 2, 7

[18] Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective Constitutional AI: Aligning a Language Model with Public Input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024. 2

[19] Maarten Buyl, Hadi Khalaf, Claudio Mayrink Verdun, Lucas Monteiro Paes, Caio Cesar Vieira Machado, and Flavio du Pin Calmon. AI Alignment at Your Discretion. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2025. 2

[20] Abigail Z. Jacobs and Hanna Wallach. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021. 2

[21] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A Roadmap to Pluralistic Alignment. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. 2, 16, 17

[22] Daniel Halpern, Evi Micha, Ariel D. Procaccia, and Itai Shapira. Pairwise Calibrated Rewards for Pluralistic Alignment. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2, 17

[23] Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond Preferences in AI Alignment. *Philosophical Studies*, 182(7):1813–1863, 2025. 2

[24] Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, 2024. 2

[25] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, 2022. arXiv:2204.05862. 2

[26] Jiahao Yuan, Zixiang Di, Shangzixin Zhao, Zhiqing Cui, Hanqing Wang, Guisong Yang, and Usman Naseem. Cultural Palette: Pluralising Culture Alignment via Multi-Agent Palette, 2024. arXiv:2412.11167. 2

[27] Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. Cultural Incongruencies in Artificial Intelligence. In *First Workshop on Cultures in AI/AI in Culture, NeurIPS 2022*, 2022. 2

[28] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric J. Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Chenyu Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwani, Yoshua Bengio, Danqi Chen, Philip Torr, Samuel Albanie, Tegan Maharaj, Jakob Nicolaus Foerster, Florian Tramèr, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational Challenges in Assuring Alignment and Safety of Large Language Models. *Transactions on Machine Learning Research*, 2024. 2

[29] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Michael Bean, Katerina Margatina, Rafael Mosquera, Juan Manuel Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 2

[30] Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular Pluralism: Pluralistic Alignment via

Multi-LLM Collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. 2, 16

[31] Carter Blair, Kate Larson, and Edith Law. Reflective Verbal Reward Design for Pluralistic Alignment. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 2025. 2

[32] Michael J.Q. Zhang, Zhilin Wang, Jena D. Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. Diverging Preferences: When Do Annotators Disagree and Do Models Know? In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. 2

[33] Kanad Shrikar Pardeshi, Itai Shapira, Ariel D. Procaccia, and Aarti Singh. Learning Social Welfare Functions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[34] Amartya Sen. The Possibility of Social Choice. *American Economic Review*, 89(3):349–378, 1999. 2

[35] Amartya Sen. Social Choice Theory. In Kenneth J. Arrow and Michael D. Intriligator, editors, *Handbook of Mathematical Economics*, volume 3, pages 1073–1181. Elsevier, 1986. 2

[36] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. Position: Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. 2, 7

[37] Jessica Dai and Eve Fleisig. Mapping Social Choice Theory to RLHF. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024. 2

[38] Abhilash Mishra. AI Alignment and Social Choice: Fundamental Limitations and Policy Implications, 2023. arXiv:2310.16048. 2

[39] Luise Ge, Daniel Halpern, Evi Micha, Ariel D. Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. Axioms for AI Alignment from Human Feedback. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2, 10

[40] Parand A. Alamdari, Soroush Ebadian, and Ariel D. Procaccia. Policy Aggregation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[41] Roberto-Rafael Maura-Rivero, Marc Lanctot, Francesco Visin, and Kate Larson. Jackpot! Alignment as a Maximal Lottery, 2025. arXiv:2501.19266. 2, 15

[42] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 6

[43] Leo Gao, John Schulman, and Jacob Hilton. Scaling Laws for Reward Model Overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. 3

[44] Xueru Wen, Jie Lou, Yaojie Lu, Hongyu Lin, Xing Yu, Xinyu Lu, Ben He, Xianpei Han, Debing Zhang, and Le Sun. Rethinking Reward Model Evaluation: Are We Barking Up the Wrong Tree? In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[45] Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios Nikolas Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to Evaluate Reward Models for RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[46] Iason Gabriel and Vafa Ghazavi. The Challenge of Value Alignment: From Fairer Algorithms to AI Safety. In Carissa Véliz, editor, *Oxford Handbook of Digital Ethics*, pages 336–355. Oxford University Press, 2023. 3

[47] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A Survey of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*, 2025. 4

[48] Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled Reinforcement Learning with Human Feedback from Pairwise or K-Wise Comparisons. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. 4

[49] Peter C. Fishburn. Condorcet Social Choice Functions. *SIAM Journal on Applied Mathematics*, 33(3):469–489, 1977. 4

[50] Louis L. Thurstone. A Law of Comparative Judgment. *Psychological Review*, 34(4):273–286, 1927. 5

[51] R. Duncan Luce. *Individual Choice Behavior*. John Wiley, Oxford, England, 1959. 5

[52] Daniel McFadden. Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers in Econometrics*, pages 105–142. Academic Press, New York, 1974. 5

[53] Ralph Allan Bradley and Milton E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345, 1952. 5, 6

[54] Sang T. Truong, Andreas Haupt, and Sanmi Koyejo. *Machine Learning from Human Preferences*. Stanford University, 2025. 5, 7

[55] John I. Yellott. The Relationship Between Luce's Choice Axiom, Thurstone's Theory of Comparative Judgment, and the Double Exponential Distribution. *Journal of Mathematical Psychology*, 15(2):109–144, 1977. 5

[56] W. Bradley Knox, Stephane Hatgis-Kessell, Serena Booth, Scott Niekum, Peter Stone, and Alessandro G. Allievi. Models of Human Preference for Learning Reward Functions. *Transactions on Machine Learning Research*, 2023. 5

[57] Jean-Charles de Borda. Mémoire sur les élections au scrutin. In *Mémoires de l'Académie Royale des Sciences année 1781*, pages 657–665. l'Imprimerie Royale, Paris, 1784. 5

[58] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia. *Handbook of Computational Social Choice*. Cambridge University Press, 2016. 5, 7, 14

[59] Lowell Bruce Anderson, Helena Dandurova, James E. Falk, and Lana Yeganova. Relationships Between Borda Voting and Zermelo Ranking. *Social Choice and Welfare*, 32(3):355–365, 2009. 6

[60] Ernst Zermelo. Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29:436–460, 1929. 6

[61] Lester R. Ford. Solution of a Ranking Problem from Binary Comparisons. *The American Mathematical Monthly*, 64(8):28–33, 1957. 6

[62] Henry E. Daniels. Round-Robin Tournament Scores. *Biometrika*, 56(2): 295–299, 1969. 6

[63] Thomas Jech. The Ranking of Incomplete Tournaments: A Mathematician's Guide to Popular Sports. *The American Mathematical Monthly*, 90(4):246–266, 1983. 6

[64] Eric Maskin. Borda's Rule and Arrow's Independence Condition. *Journal of Political Economy*, 133(2):385–420, 2025. 7

[65] Benjamin Reilly. Social Choice in the South Seas: Electoral Innovation and the Borda Count in the Pacific Island Countries. *International Political Science Review*, 23(4):355–372, 2002. 7

[66] Michel L. Balinski and Rida Laraki. *Majority Judgment: Measuring, Ranking, and Electing*. MIT Press, 2010. 7

[67] Emanuel Todorov. Linearly-Solvable Markov Decision Problems. In *The Twentieth Annual Conference on Neural Information Processing Systems*, 2006. 7

[68] Jan Peters, Katharina Mülling, and Yasemin Altun. Relative Entropy Policy Search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1): 1607–1612, 2010. 7

[69] Ali Shirali, Arash Nasr-Esfahany, Abdullah Omar Alomar, Parsa Mirtaheri, Rediet Abebe, and Ariel D. Procaccia. Direct Alignment with Heterogeneous Preferences. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 7

[70] Hannah Rose Kirk, Andrew M. Bean, Bertie Vidgen, Paul Röttger, and Scott A. Hale. The Past, Present and Better Future of Feedback Learning in Large Language Models for Subjective Human Preferences and Values. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 7

[71] Angelica Chen, Sadhika Malladi, Lily H. Zhang, Xinyi Chen, Qiuyi Zhang, Rajesh Ranganath, and Kyunghyun Cho. Preference Learning Algorithms Do Not Learn Preference Rankings. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 7

[72] Kyuyoung Kim, Ah Jeong Seo, Hao Liu, Jinwoo Shin, and Kimin Lee. Margin Matching Preference Optimization: Enhanced Model Alignment with Granular Feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. 7

[73] Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. The History and Risks of Reinforcement Learning and Human Feedback, 2023. arXiv:2310.13595. 7

[74] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. On Releasing Annotator-Level Labels and Information in Datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, 2021. 7

[75] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned, 2022. arXiv:2209.07858. 7

[76] Arthur H. Copeland. A Reasonable Social Welfare Function. Mimeographed notes from the Seminar on Applications of Mathematics to the Social Sciences, University of Michigan, Ann Arbor, 1951. 7

[77] Jiancong Xiao, Zhekun Shi, Kaizhao Liu, Qi Long, and Weijie J. Su. Theoretical Tensions in RLHF: Reconciling Empirical Success with Inconsistencies in Social Choice Theory, 2025. arXiv:2506.12350. 7

[78] Zhiyu An, Duaa Nakshbandi, and Wan Du. Differential Voting: Loss Functions for Axiomatically Diverse Aggregation of Heterogeneous Preferences, 2026. arXiv:2601.18824. 7

[79] T. Nicolaus Tideman. Independence of Clones as a Criterion for Voting Rules. *Social Choice and Welfare*, 4(3):185–206, 1987. 7

[80] Ratip Emin Berker, Sílvia Casacuberta Puig, Isaac Robinson, and Christopher Ong. Obvious Independence of Clones. In *Artificial Intelligence for Research and Democracy*, 2024. 7

[81] Damien Berriaud and Roger Wattenhofer. Clone-Robust Weights in Metric Spaces: Handling Redundancy Bias for Benchmark Aggregation. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, 2026. 8

[82] Ariel D. Procaccia, Benjamin Schiffer, and Shirley Zhang. Clone-Robust AI Alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. 8, 9

[83] Alexandros Hollender and Sonja Kraiczy. Enforcing Axioms for AI Alignment Under Loss-Based Rules. In *The Fourteenth International Conference on Learning Representations*, 2026. 10

[84] Ariel D. Procaccia and Jeffrey S. Rosenschein. The Distortion of Cardinal Preferences in Voting. In *Cooperative Information Agents X*, 2006. 12

[85] Craig Boutilier, Ioannis Caragiannis, Simi Haber, Tyler Lu, Ariel D. Procaccia, and Or Sheffet. Optimal Social Choice Functions: A Utilitarian View. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, 2012. 12

[86] Elliot Anshelevich, Aris Filos-Ratsikas, Nisarg Shah, and Alexandros A. Voudouris. Distortion in Social Choice Problems: The First 15 Years and Beyond. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 2021. 12

[87] Paul Gölz, Nika Haghtalab, and Kunhe Yang. Distortion of AI Alignment: Does Preference Optimization Optimize for Preferences? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 12, 13, 16

[88] Keertana Chidambaram, Karthik Vinay Seetharaman, and Vasilis Syrgkanis. Direct Preference Optimization with Unobserved Preference Heterogeneity: The Necessity of Ternary Preferences. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. 13

[89] Luise Ge, Daniel Halpern, Gregory Kehne, and Yevgeniy Vorobeychik. Linear Social Choice with Few Queries: A Moment-Based Approach, 2026. arXiv:2603.19510. 13

[90] Yeshwanth Cherapanamjeri, Constantinos Daskalakis, Gabriele Farina, and Sobhan Mohammadpour. Learning Correlated Reward Models: Statistical Barriers and Opportunities. In *The Fourteenth International Conference on Learning Representations*, 2026. 13

[91] Federico Echenique, Alireza Fallah, and Michael I. Jordan. A General Framework for Estimating Preferences Using Response Time Data. In *Proceedings of the 27th ACM Conference on Economics and Computation*, 2026. 13

[92] Federico Echenique, Alireza Fallah, Baihe Huang, and Michael I. Jordan. Response Time Enhances Alignment with Heterogeneous Preferences, 2026. arXiv:2605.06987. 13

[93] Germain Kreweras. Aggregation of Preference Orderings. In *Mathematics and Social Sciences I: Proceedings of the Seminars of Menthon-Saint-Bernard*, 1965. 14

[94] Peter C. Fishburn. Probabilistic Social Choice Based on Simple Voting Comparisons. *The Review of Economic Studies*, 51(4):683–692, 1984. 14

[95] Ronald L. Rivest and Emily Shen. An Optimal Single-Winner Preferential Voting System Based on Game Theory. In *Proceedings of the Third International Workshop on Computational Social Choice (COMSOC-2010)*, 2010. 14

[96] Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual Dueling Bandits. In *Proceedings of The 28th Conference on Learning Theory*, 2015. 14

[97] Marc Lanctot, Kate Larson, Yoram Bachrach, Luke Marris, Zun Li, Avishkar Bhoopchand, Thomas Anthony, Brian Tanner, and Anna Koop. Evaluating Agents Using Social Choice Theory, 2025. arXiv:2312.03121. 14

[98] Hadi Khalaf, Flavio Calmon, Daniel Halpern, Ariel D. Procaccia, Itai Shapira, and Serena Lutong Wang. Robust AI Evaluation Through Maximal Lotteries.

In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. 14

[99] John von Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928. 14

[100] Florian Brandl, Felix Brandt, and Hans Georg Seedig. Consistent Probabilistic Social Choice. *Econometrica*, 84(5):1839–1880, 2016. 15

[101] Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. Nash Learning from Human Feedback. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. 15

[102] Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A Minimaximalist Approach to Reinforcement Learning from Human Feedback. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. 15

[103] Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-Play Preference Optimization for Language Model Alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. 15

[104] Daniele Calandriello, Zhaohan Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, Rishabh Joshi, Zeyu Zheng, and Bilal Piot. Human Alignment of Large Language Models Through Online Preference Optimisation. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. 15

[105] Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct Nash Optimization: Teaching Language Models to Self-Improve with General Preferences, 2024. arXiv:2404.03715. 15

[106] Daniil Tiapkin, Daniele Calandriello, Denis Belomestny, Eric Moulines, Alexey Naumov, Kashif Rasul, Michal Valko, and Pierre Menard. Proximal Point Nash Learning from Human Feedback, 2025. arXiv:2505.19731. 15

[107] Benjamin Heymann. Adaptive Preference Aggregation, 2025. arXiv:2503.10215. 15

[108] Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J. Su. On the Algorithmic Bias of Aligning Large Language Models with RLHF: Preference Collapse and Matching Regularization. *Journal of the American Statistical Association*, 120(552):2154–2164, 2025. 16

[109] Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. Evaluating the Diversity and Quality of LLM Generated Content. In *Second Conference on Language Modeling*, 2025. 16

[110] Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the Effects of RLHF on LLM Generalisation and Diversity. In *The Twelfth International Conference on Learning Representations*, 2024. 16, 17

[111] Thom Lake, Eunsol Choi, and Greg Durrett. From Distributional to Overton Pluralism: Investigating Large Language Model Alignment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025. 16

[112] Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. A Distributional Approach to Controlled Text Generation. In *The Ninth International Conference on Learning Representations*, 2021. 16, 17

[113] Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models, 2025. arXiv:2505.22617. 16

[114] Itai Shapira, Gerdus Benade, and Ariel D. Procaccia. How RLHF Amplifies Sycophancy. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. 16

[115] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning*, 2023. 16

[116] 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, 2023. arXiv:2303.05453. 16, 17

[117] Stewart Slocum, Asher Parker-Sartori, and Dylan Hadfield-Menell. Diverse Preference Learning for Capabilities and Alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. 16

[118] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with Diverse Human Preferences. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. 16, 17

[119] Haoxian Chen, Hanyang Zhao, Henry Lam, David Yao, and Wenpin Tang. Mallows-DPO: Fine-Tune Your LLM with Preference Dispersions. In *Pluralistic Alignment Workshop at NeurIPS 2024*, 2024. 16

[120] Anthony GX-Chen, Jatin Prakash, Jeff Guo, Rob Fergus, and Rajesh Ranganath. KL-Regularized Reinforcement Learning for Generative Modelling Is Designed to Mode Collapse. In *The Fourteenth International Conference on Learning Representations*, 2026. 16

[121] Ilgee Hong, Zichong Li, Alexander Bukharin, Yixiao Li, Haoming Jiang, Tianbao Yang, and Tuo Zhao. Adaptive Preference Scaling for Reinforcement Learning with Human Feedback. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 17

[122] Jiashuo Wang, Haozhao Wang, Shichao Sun, and Wenjie Li. Aligning Language Models with Human Preferences via a Bayesian Approach. In *The Thirty-*

seventh Annual Conference on Neural Information Processing Systems, 2023. 17

[123] Nuoya Xiong and Aarti Singh. Projection Optimization: A General Framework for Multi-Objective and Multi-Group RLHF. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. 17

[124] Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman E. Ozdaglar. RLHF from Heterogeneous Feedback via Personalization and Preference Aggregation. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*, 2024. 17

[125] Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. PAL: Sample-Efficient Personalized Reward Modeling for Pluralistic Alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. 17

[126] Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel A. Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. Value Profiles for Encoding Human Variation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. 17

[127] Elinor Poole-Dayan, Jiayi Wu, Taylor Sorensen, Jiaxin Pei, and Michiel A. Bakker. Benchmarking Overton Pluralism in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. 17

[128] Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-Tuning Language Models to Find Agreement Among Humans with Diverse Preferences. In *The Thirty-sixth Annual Conference on Neural Information Processing Systems*, 2022. 17

[129] Sara Fish, Paul Gölz, David Parkes, Ariel Procaccia, Gili Rusak, Itai Shapira, and Manuel Wuthrich. Generative Social Choice. *Journal of the ACM*, 73(2): 11:1–11:52, 2026. 17

[130] Niclas Boehmer, Sara Fish, and Ariel D. Procaccia. Generative Social Choice: The Next Generation. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. 17

[131] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red Teaming Language Models with Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. 17

[132] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 17

[133] Gihoon Kim and Euntai Kim. Swap-Guided Preference Learning for Personalized Reinforcement Learning from Human Feedback. In *The Fourteenth International Conference on Learning Representations*, 2026. 17

[134] Andre Barreto, Vincent Dumoulin, Yiran Mao, Mark Rowland, Nicolas Perez-Nieves, Bobak Shahriari, Yann Dauphin, Doina Precup, and Hugo Larochelle. Capturing Individual Human Preferences with Reward Features. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 17

[135] Avinandan Bose, Zhihan Xiong, Yuejie Chi, Simon Shaolei Du, Lin Xiao, and Maryam Fazel. LoRe: Personalizing LLMs via Low-Rank Reward Modeling. In *Second Conference on Language Modeling*, 2025. 17

[136] Hannah Rose Kirk, Liu Leqi, Fanzhi Zeng, Henry Davidson, Bertie Vidgen, Christopher Summerfield, and Scott A. Hale. PRISM-X: Experiments on Personalised Fine-Tuning with Human and Simulated Users, 2026. arXiv:2605.13307. 17

[137] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. AI Can Help Humans Find Common Ground in Democratic Deliberation. *Science*, 386(6719):eadq2852, 2024. 17

[138] Jay Chooi, Paul Gölz, Ariel D. Procaccia, Benjamin Schiffer, and Shirley Zhang. Finding Common Ground in a Sea of Alternatives, 2026. arXiv:2603.16751. 17

[139] Aviv Ovadya. Generative CI Through Collective Response Systems, 2023. arXiv:2302.00672. 17