

Direct Alignment with Heterogeneous Preferences

Ali Shirali^{*†1}, Arash Nasr-Esfahany^{*2}, Abdullah Alomar^{2,3}, Parsa Mirtaheri⁴,
Rediet Abebe^{‡5}, and Ariel Procaccia^{‡5}

¹University of California Berkeley

²Massachusetts Institute of Technology

³Ikigai Labs

⁴University of California San Diego

⁵Harvard University

February 16, 2025

Abstract

Alignment with human preferences is commonly framed using universal reward functions, even though human preferences are inherently heterogeneous. We examine the limits of this homogeneity assumption and characterize settings where data from annotators can mitigate these challenges. We show that aligning to heterogeneous preferences with a single policy is best achieved using average reward. However, this requires annotator information. We examine improvements under different information settings, focusing on direct alignment methods. We find that minimal information can yield first-order improvements, while full feedback from each user type leads to consistent learning of the optimal policy. Surprisingly, however, no sample-efficient consistent direct loss exists in this latter setting. These results reveal a fundamental tension between consistency and sample efficiency in direct policy alignment.

1 Introduction

Human rewards and preferences are heterogeneous [1, 2, 3, 4, 5]. Despite this, learning from preference data often bypasses this insight, relying on what we dub the *preference homogeneity assumption*. This tension in assumptions is readily apparent in standard human-AI alignment methods—such as reinforcement learning from human feedback (RLHF) [6, 7, 8] and direct preference optimization (DPO) [9]—which assume a single reward function captures the interests of the entire population.

We examine the limits of the preference homogeneity assumption when individuals belong to *user types*, each characterized by a specific reward function. Recent work has shown that in this setting, the homogeneity assumption can lead to unexpected behavior [10, 11, 12]. One challenge is that, under this assumption, learning from human preferences becomes *unrealizable*, as a single reward function cannot capture the complexity of population preferences with multiple reward functions [13, 14]. Both RLHF and DPO rely on maximum likelihood estimation (MLE) to optimize the reward or policy. Unrealizability implies their likelihood functions cannot fully represent the underlying preference data distribution, resulting in a nontrivial optimal MLE solution. From another perspective, learning a universal reward or policy from a heterogeneous population inherently involves an *aggregation* of diverse interests, and this aggregation is nontrivial.

In the quest for a single policy that accommodates a heterogeneous population with multiple user types, we show that the only universal reward yielding a well-defined alignment problem is an affine

*Equal contribution

†Work done while visiting Harvard

‡Equal advising

aggregation of the reward functions across user types, with the average reward as a natural choice. However, standard methods like DPO do not maximize this user-weighted average reward. Building on insights by Siththaranjan, Laidlaw, and Hadfield-Menell [15], we show that DPO implicitly maximizes Borda count, which comes with unexpected drawbacks, e.g., the optimal solution depends on how alternative responses are sampled, even for infinite data.

We observe that learning the average reward over user types—or equivalently, a policy that maximizes it—from anonymous data is impossible. Focusing on *direct alignment methods*, which avoid explicit reward modeling, we study the benefits of using annotator data for a range of information settings. We show that improving DPO with a first-order correction to its objective is possible with minimal annotator information. Specifically, we design an approximate direct alignment method when each preference data point is paired with another one labeled by the same user.

On the other hand, we find that there are limits to what is possible even with significant annotator information. In particular, we propose a consistent loss function for direct alignment when we have feedback on each data point from each user type. But this loss is sample-inefficient, using only data where all annotator types agree. Surprisingly, we prove that no consistent loss uses the rest of the data.

In summary, the homogeneity assumption leads to undesirable outcomes when aligning a single AI agent to diverse preferences. Our analysis shows that there is a limited class of reward aggregation that results in a valid objective for alignment, with average reward over user types emerging as the natural candidate. This requires some annotator data, though small amounts of data can yield significant improvements. Our findings, however, uncover a fundamental tension between consistency and sample efficiency in direct alignment. To achieve both sample efficiency and consistency, we must forgo the benefits of direct optimization and instead train individualized reward models, which inevitably incurs significant training and storage costs.

2 Preliminaries

In the *alignment problem*, we consider a setting where a *reward* function r^* evaluates responses to queries. Formally, $r^*([\mathbf{x}, \mathbf{y}])$ is the reward value of responding \mathbf{y} to a query \mathbf{x} .

The alignment problem involves designing a *policy*, which chooses high-reward responses. Let π denote a policy, defining a probability distribution over possible responses: i.e., given a query \mathbf{x} , π returns \mathbf{y} with probability $\pi(\mathbf{y} | \mathbf{x})$. Commonly, we start with a reference policy π_{ref} , which serves as a prior over \mathbf{y} [16]. The goal is then to find a new policy π that, for every \mathbf{x} , maximizes

$$\mathbb{E}_{\mathbf{y} \sim \pi(\cdot | \mathbf{x})} [r^*([\mathbf{x}, \mathbf{y}])] - \beta D_{\text{KL}}(\pi(\cdot | \mathbf{x}); \pi_{\text{ref}}(\cdot | \mathbf{x})). \quad (1)$$

We denote the optimal policy by π^* . In practice, π_{ref} is often a pretrained language model, and the regularization parameter β controls deviation from it. Eq. (1) often includes $\mathbb{E}_{\mathbf{x}}$, which is important in practice but does not affect π^* in theory.

When r^* is explicitly known, we can directly apply RL to maximize Eq. (1). In many real-world settings, however, we do not know r^* and must estimate it. In such cases, we can collect human feedback to infer the reward function, after which we can use RL to optimize the policy, commonly known as RLHF. While RLHF is widely used, tuning this approach can be challenging due to the inherent complexities of RL. Recently, *direct alignment with preferences* has gained popularity as an alternative approach [17, 9, 18]. Unlike RLHF, direct alignment methods bypass explicit reward modeling to instead train a policy directly from human feedback.

Preference Model. Both direct alignment and RLHF rely on a model of human preference to relate reward values with observed preference data. Consider the case where responses $\mathbf{y}_1, \mathbf{y}_2$ are generated for a given query \mathbf{x} . We express the probability that \mathbf{y}_2 is preferred to \mathbf{y}_1 as

$$\Pr(\mathbf{y}_2 \succ \mathbf{y}_1 | \mathbf{x}; r^*) = \sigma(r^*([\mathbf{x}, \mathbf{y}_2]) - r^*([\mathbf{x}, \mathbf{y}_1])), \quad (2)$$

where σ is a non-decreasing function in $[0, 1]$ [19, 20, 21]. A widely-used choice for σ is the sigmoid function corresponding to the well-known Bradley-Terry (BT) model [22].

Direct Preference Optimization. Among direct alignment methods, DPO has emerged as the most widely used approach. It leverages a closed-form solution to Eq. (1), which allows it to link any reward directly to its optimal policy. Thereby, rather than explicitly estimating the reward, DPO optimizes a policy whose *induced reward* best explains the observed preferences. We derive this connection below:

First, maximizing Eq. (1) has a well-known solution [23]. The optimal policy π^* takes the form:

$$\pi^*(\mathbf{y} | \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \pi_{\text{ref}}(\mathbf{y} | \mathbf{x}) \cdot \exp\left(\frac{1}{\beta} r^*([\mathbf{x}, \mathbf{y}])\right). \quad (3)$$

Here, $Z(\mathbf{x}) = \sum_{\mathbf{y}'} \pi_{\text{ref}}(\mathbf{y}' | \mathbf{x}) \cdot \exp(\frac{1}{\beta} r^*([\mathbf{x}, \mathbf{y}']))$ is the partition function. Eq. (3) establishes a direct relationship between policy ratios and reward differences:

$$r^*(\mathbf{y}_2) - r^*(\mathbf{y}_1) = \beta \log \frac{\pi^*(\mathbf{y}_2)}{\pi_{\text{ref}}(\mathbf{y}_2)} - \beta \log \frac{\pi^*(\mathbf{y}_1)}{\pi_{\text{ref}}(\mathbf{y}_1)}. \quad (4)$$

Henceforth, we omit \mathbf{x} when we can do so without ambiguity. This equation shows that the difference in rewards between two responses is fully captured by the difference in their policy ratios. Using this formulation, we can define the *induced reward* of a policy π by $\beta \log \frac{\pi(\mathbf{y})}{\pi_{\text{ref}}(\mathbf{y})}$. The induced reward of π is the reward for which π is the optimal policy.

The difference in rewards of \mathbf{y}_1 and \mathbf{y}_2 is sufficient to express the likelihood of $\mathbf{y}_2 \succ \mathbf{y}_1$ in Eq. (2). Using Eq. (4), we can therefore write the likelihood as a function of π^* :

$$\Pr(\mathbf{y}_2 \succ \mathbf{y}_1 | \pi^*) = \sigma\left(\beta \log \frac{\pi^*(\mathbf{y}_2)}{\pi_{\text{ref}}(\mathbf{y}_2)} - \beta \log \frac{\pi^*(\mathbf{y}_1)}{\pi_{\text{ref}}(\mathbf{y}_1)}\right). \quad (5)$$

For any policy π , we can define $\Pr(\mathbf{y}_2 \succ \mathbf{y}_1 | \pi)$ similarly. We can then estimate π^* using MLE: Given a dataset \mathcal{D} with query and response pairs $(\mathbf{x}, \mathbf{y}_l, \mathbf{y}_w)$, where $\mathbf{y}_w \succ \mathbf{y}_l$, DPO finds π^* by maximizing the log-likelihood

$$\sum_{(\mathbf{x}, \mathbf{y}_l, \mathbf{y}_w) \in \mathcal{D}} \log \Pr(\mathbf{y}_w \succ \mathbf{y}_l | \mathbf{x}; \pi), \quad (6)$$

or equivalently, minimizing the cross-entropy loss. Under mild assumptions, MLE is a consistent estimator of π^* .

3 Problem Formulation

The alignment problem is traditionally framed under a preference homogeneity assumption, where a single reward is presumed to capture all individual interests. In practice, people’s preferences can differ significantly. To better capture real-world settings, we formalize preference heterogeneity by allowing reward functions to vary across *user types*.

Heterogeneous Preferences. The influential study of “individual choice behavior” by Luce [20] and other foundational works on human decision-making in mathematical psychology such as Shepard [24] focus on *individual* preference models. Luce [20] uses an axiomatic approach to establish the existence of a value function for each individual that, once normalized, explains the individual’s choice probabilities. The widely used BT is one such example.

In practice, we often cannot observe individuals’ identities. Therefore, standard approaches in preference modeling use a single reward function across the entire population. This homogeneity

assumption makes preference learning *unrealizable*: Even if a specific model family can explain preferences of every individual, we cannot ensure that a model from the same family explains population-level choices. For example, we cannot represent a mixture of BT models with a single BT (we prove this in Proposition B.1 for completeness).

To account for heterogeneity, we need to define individual rewards. However, learning at scale with this level of granularity is impractical, especially when working with finite data. Hence, we group individuals into multiple *user types*, denoted by \mathcal{U} . Individuals with the same type have similar rewards, but this need not hold across types.

For a given user of type $u \in \mathcal{U}$, we denote the corresponding reward function by $r^*(\cdot; u)$. This function assigns a scalar reward $r^*([\mathbf{x}, \mathbf{y}]; u)$ for every response \mathbf{y} to a given query \mathbf{x} . We model the preferences of an individual of type u with

$$\Pr(\mathbf{y}_2 \succ \mathbf{y}_1 \mid r^*, u) = \sigma(r^*(\mathbf{y}_2; u) - r^*(\mathbf{y}_1; u)), \quad (7)$$

The population-level preferences are therefore given by:

$$\Pr(\mathbf{y}_2 \succ \mathbf{y}_1 \mid r^*) = \mathbb{E}_u \left[\sigma(r^*(\mathbf{y}_2; u) - r^*(\mathbf{y}_1; u)) \right]. \quad (8)$$

The Extended Alignment Problem. Our goal is to find a single policy that can effectively accommodate a heterogeneous population. This is essential when user types are not observable during inference. Furthermore, a universal policy may be preferable when personalization comes with significant drawbacks: e.g., cases where prioritizing a broadly accepted notion of truth or safety is more important than catering to individual preferences [25, 26].

Deriving a universal policy requires aggregation of diverse rewards. As we show next, an affine combination is the only form of aggregation that guarantees a well-defined problem, i.e., a problem that yields the same optimal policy for every reward that is consistent with the preference data.

Proposition 3.1. *Consider an aggregation $f : \mathbb{R}^{\mathcal{U}} \rightarrow \mathbb{R}$. If $f(\{r(\mathbf{y}; u)\}_{u \in \mathcal{U}})$ induces the same ordering over \mathbf{y} for every reward r consistent with the preferences distribution, then under weak regulatory assumptions, f must be affine.*

See proof on page 20. This result rules out many commonly-used aggregations, such as Max-Min [27] or Nash social welfare [28]. The expected reward across user types emerges as a natural choice here. Any other affine combination would weigh people unequally, which requires strong justifications and is rare in practice.

To summarize, our objective is to maximize

$$\mathbb{E}_{\mathbf{y} \sim \pi(\cdot \mid \mathbf{x})} [\mathbb{E}_u [r^*([\mathbf{x}, \mathbf{y}]; u)]] - \beta D_{\text{KL}}(\pi(\cdot \mid \mathbf{x}); \pi_{\text{ref}}(\cdot \mid \mathbf{x})) \quad (9)$$

for every prompt \mathbf{x} . With this extended framework in mind, we next discuss why standard approaches like RLHF or DPO do not necessarily yield the optimal policy.

4 Implications of Homogeneity Assumption

With heterogeneous preferences, standard RLHF or DPO cannot yield the optimal policy π^* that maximizes Eq. (9). If they did, it would also be possible to learn the user-weighted average reward as the induced reward of π^* . However, as we show in Proposition 5.1 and was previously observed by Siththaranjan, Laidlaw, and Hadfield-Menell [15] and Procaccia, Schiffer, and Zhang [29], learning the expected reward from anonymous preferences is impossible.

To explain DPO’s failure in finding π^* , we extend its derivation to the heterogeneous setting in Section 4.1. This analysis lays the foundations to account for heterogeneity in DPO later on. In Section 4.2, we show that DPO’s policy aligns with Borda count and, in Section 4.3, highlight its limitations. While our analysis focuses on DPO, similar insights extend to RLHF by substituting the policy with its induced reward.

4.1 Objective is Not the Expected Reward

We follow DPO’s derivation from Section 2 but under heterogeneity. We show the closed-form connection between π^* and r^* is no longer sufficient to express the likelihood function. Beginning with Eq. (9), the optimal policy is

$$\pi^*(\mathbf{y}) = \frac{1}{Z(\mathbf{x})} \pi_{\text{ref}}(\mathbf{y}) \cdot \exp\left(\frac{1}{\beta} \mathbb{E}_u[r^*(\mathbf{y}; u)]\right). \quad (10)$$

Define $\Delta r^*(\mathbf{y}_1, \mathbf{y}_2; u) := r^*(\mathbf{y}_2; u) - r^*(\mathbf{y}_1; u)$. The policy ratios of π^* are related to the expected difference in rewards:

$$\mathbb{E}_u[\Delta r^*(\mathbf{y}_1, \mathbf{y}_2; u)] = \beta \log \frac{\pi^*(\mathbf{y}_2)}{\pi_{\text{ref}}(\mathbf{y}_2)} - \beta \log \frac{\pi^*(\mathbf{y}_1)}{\pi_{\text{ref}}(\mathbf{y}_1)}. \quad (11)$$

In the homogeneous case, Δr^* was sufficient to describe the likelihood of $\mathbf{y}_2 \succ \mathbf{y}_1$. However, with heterogeneous preferences, $\mathbb{E}_u[\Delta r^*]$ alone does not suffice to write the likelihood function in Eq. (8). It is only under the approximation

$$\mathbb{E}_u[\sigma(\Delta r^*(\mathbf{y}_1, \mathbf{y}_2; u))] \approx \sigma\left(\mathbb{E}_u[\Delta r^*(\mathbf{y}_1, \mathbf{y}_2; u)]\right) \quad (12)$$

that we can write $\Pr(\mathbf{y}_2 \succ \mathbf{y}_1 \mid r^*)$ in terms of policy ratios as in Eq. (5), and minimize DPO’s loss to find π^* .

4.2 Ordinal Consistency with Borda Count

If DPO were the answer, what would the question be? We partially answer this question by an adaptation of a result from Siththaranjan, Laidlaw, and Hadfield-Menell [15] which we restate for completeness. First, define Borda count as follows.

Definition 4.1 (Normalized Borda count). For a prompt \mathbf{x} , let $\mathcal{D}(\cdot \mid \mathbf{x})$ denote the distribution of alternative responses sampled for \mathbf{x} . The Normalized Borda Count (NBC) of \mathbf{y} at \mathbf{x} is the probability that an annotator with a random type prefers \mathbf{y} over an alternative response $\mathbf{y}' \sim \mathcal{D}(\cdot \mid \mathbf{x})$:

$$\text{NBC}(\mathbf{y} \mid \mathbf{x}) := \mathbb{E}_{\mathbf{y}' \sim \mathcal{D}(\cdot \mid \mathbf{x})} \left[\Pr(\mathbf{y} \succ \mathbf{y}' \mid \mathbf{x}; r^*) \right]. \quad (13)$$

We next show that DPO’s policy ratios are ordinally consistent with the normalized Borda count.

Proposition 4.2. *Suppose responses to \mathbf{x} in the preference dataset are drawn from $\mathcal{D}(\cdot \mid \mathbf{x})$. In the limit of many data points, DPO’s induced reward, or equivalently $\frac{\pi_{\text{DPO}}(\cdot \mid \mathbf{x})}{\pi_{\text{ref}}(\cdot \mid \mathbf{x})}$, has the same ordering over responses as $\text{NBC}(\cdot \mid \mathbf{x})$.¹*

See [proof](#) on page 20. Proposition 4.2 also applies to the homogeneous setting. In this case, however, NBC aligns with r^* . It is worth mentioning that DPO is not the only method consistent with NBC; identity preference optimization (IPO) [30] uses NBC as its objective. We next highlight key differences between NBC and the user-weighted expected reward along with DPO’s drawbacks in practice.

¹We can view $\text{NBC}(\mathbf{y} \mid \mathbf{x})$ as an aggregation of rewards at \mathbf{y} . One can verify that NBC meets the order consistency condition of Proposition 3.1. However, it uses the reward value at $\mathbf{y}' \neq \mathbf{y}$ to define the aggregated reward at \mathbf{y} and thus does not fall under Proposition 3.1. In fact, this interdependency causes the issues we discuss Section 4.3.

4.3 Practical Drawbacks

In case of heterogeneous preferences, Borda count can significantly diverge from the user-weighted expected reward. This is studied under *distortion* in social choice problems [31]. Notably, NBC in Eq. (13) depends on \mathcal{D} . Therefore, although data collection is irrelevant in defining the optimal policy, it does affect π_{DPO} .

Next, we illustrate two key differences between π^* and π_{DPO} using examples. Unless otherwise stated, we assume $\mathcal{D}(\cdot | \mathbf{x})$ and $\pi_{\text{ref}}(\cdot | \mathbf{x})$ are uniform, and annotators follow BT. Refer to Appendix A for further drawbacks of DPO (minority suppression and IIA violation).

Sensitivity to Preference Dataset Distribution. Suppose $\mathcal{U} = \{A, B\}$ and types are equally represented. Given three possible responses, type A prefers \mathbf{y}_1 but type B prefers \mathbf{y}_2 :

$$\begin{aligned} r^*(\mathbf{y}_1; A) &= 6, r^*(\mathbf{y}_2; A) = 1, r^*(\mathbf{y}_3; A) = 4, \\ r^*(\mathbf{y}_1; B) &= 3, r^*(\mathbf{y}_2; B) = 9, r^*(\mathbf{y}_3; B) = 4. \end{aligned}$$

One can verify when $\mathcal{D}(\mathbf{y}_1) = \mathcal{D}(\mathbf{y}_2)$, increasing $\mathcal{D}(\mathbf{y}_3)$ from 0.02 to 0.04 changes π_{DPO} 's preference from \mathbf{y}_2 to \mathbf{y}_1 .

DPO's policy is also sensitive to the preference model. Consider a variation of BT with a temperature of 2: $\sigma_2(z) := (1 + \exp(-z/2))^{-1}$. For the same users and uniform sampling of alternatives, increasing the temperature from 1 to 2 flips π_{DPO} 's ranking over \mathbf{y}_1 and \mathbf{y}_2 while the preference model has no effect on π^* .

We have to emphasize that the dependence of NBC, and consequently π_{DPO} , on the dataset sampling distribution \mathcal{D} is not due to finite-sample limitations or insufficient offline dataset support. This issue persists even with complete data coverage and in the limit of infinite data.

Mediocrity Promotion. Consider the task of summarization. Suppose $\mathcal{U} = \{A, B, C\}$ and types are equally represented. Type A (B) strongly favors longer (shorter) summaries while type C slightly prefers medium-length ones:

$$\begin{aligned} r^*(\text{short}; A) &= 0, r^*(\text{med}; A) = 1, r^*(\text{long}; A) = 4, \\ r^*(\text{short}; B) &= 4, r^*(\text{med}; B) = 1, r^*(\text{long}; B) = 0, \\ r^*(\text{short}; C) &= 0, r^*(\text{med}; C) = 1, r^*(\text{long}; C) = 0. \end{aligned}$$

In this case, $\pi^*(\text{short}) = \pi^*(\text{long}) > \pi^*(\text{med})$, however, $\text{NBC}(\text{short}) = \text{NBC}(\text{long}) < \text{NBC}(\text{med})$. DPO prefers medium-length summaries not strongly favored by any type.

Real-World Examples. The examples above are not contrived; in real-world cases, NBC can produce rankings different from π^* and is sensitive to dataset distribution as extensively studied under distortion of social choice rules. To show this with a real example, we use Pew Research Center surveys and analyze a question to 5101 participants: "The next time you purchase a vehicle, how likely are you to consider purchasing an electric vehicle?" (options from A: very likely to D: not at all likely). We discuss how we select this question in Appendix D. Responses come from groups of different political leanings: Republican (45%), Democratic (48%), and Neither/refused (7%).

Assuming the Luce-Shepard model [24] (see Eq. (27)), we estimate the reward for each group to calculate NBC and a user-weighted average reward. To find NBC, we use two distributions for alternatives: a uniform distribution \mathcal{D}_U and a slightly altered distribution \mathcal{D}_a with 0.2 total variation distance (TV) from \mathcal{D}_U . As shown in Fig. 1, NBC (with \mathcal{D}_U) ranks option C first despite its mediocrity: it is the second or third preference for the three groups (see Fig. 8). In contrast, the user-weighted average reward favors D: the first and second preference for Republicans and the no-lean groups, respectively. Notably, altering \mathcal{D}_U to \mathcal{D}_a flips NBC's top ranking, highlighting NBC's sensitivity to dataset distribution. Similar discrepancies appear in other Pew surveys (see Appendix D).

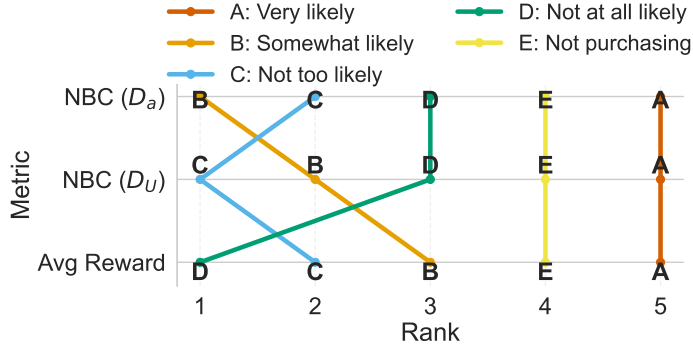


Figure 1: “The next time you purchase a vehicle, how likely are you to seriously consider purchasing an EV?” NBC ranking differs from the user-weighted average reward and is sensitive to the dataset distribution.

5 Approximate Direct Alignment with Minimal Annotator Information

The failure of standard alignment methods to find the optimal policy π^* raises the question of whether it is even possible to design a method that identifies π^* . The answer is no without annotator information. To show this, it suffices to prove that the ranking based on the user-weighted average reward is not learnable. This implies that π^* is also not learnable since its induced reward corresponds to this average reward by definition. We defer the formal definition of learnability to Definition B.2 in the appendix. Based on this definition, we prove the following impossibility:

Proposition 5.1. *If there are at least two alternatives and two user types with a continuous preference model, the ranking based on the user-weighted expected reward is not learnable without annotator information.*

See [proof](#) on page 21. Siththaranjan, Laidlaw, and Hadfield-Menell [15] (Theorem 3.4) presented a version of Proposition 5.1 in case of two alternatives and two types with $\sigma(\Delta r) = \mathbb{1}\{\Delta r > 0\}$. Procaccia, Schiffer, and Zhang [29] (Theorem 2.2) generalized this to BT. Proposition 5.1 presents a fresh perspective by generalizing the impossibility to any continuous preference model and presenting multiple proof strategies, including one that draws on a robust version of Arrow’s theorem [32].

To circumvent the impossibility in Proposition 5.1, we must either relax the requirement of exactly identifying π^* or collect some information from the annotators. This section focuses on the former, and the latter is the subject of Section 6. Next, we introduce an approximate alignment objective, along with the required information and algorithms to solve it.

5.1 First-Order Approximation

The approximation in Eq. (12) is equivalent to using a zeroth-order Taylor expansion of $\sigma(\cdot)$ around the average reward to calculate the likelihood function. To improve it, we extend DPO by incorporating an additional non-zero term from the expansion, which we call *first-order corrected DPO*. The derivation is as follows. Expanding $\sigma(\Delta r^*(\mathbf{y}_1, \mathbf{y}_2; u))$ around $\Delta \bar{r}^*(\mathbf{y}_1, \mathbf{y}_2) := \mathbb{E}_u[\Delta r^*(\mathbf{y}_1, \mathbf{y}_2; u)]$ up to the second order gives the below approximation for the likelihood:

$$\begin{aligned} \mathbb{E}_u \left[\sigma(\Delta r^*(\mathbf{y}_1, \mathbf{y}_2; u)) \right] &\approx \sigma(\Delta \bar{r}^*(\mathbf{y}_1, \mathbf{y}_2)) \\ &+ \frac{1}{2} \sigma''(\Delta \bar{r}^*(\mathbf{y}_1, \mathbf{y}_2)) \cdot \text{Var}_u[\Delta r^*(\mathbf{y}_1, \mathbf{y}_2; u)]. \end{aligned} \quad (14)$$

Note that Eq. (12) is loose when σ is nonlinear and preferences have high variance. We improve likelihood approximation by incorporating the variance term from Eq. (14). To calculate Eq. (14), we can substitute $\Delta\bar{r}^*$ in by the difference in log policy ratios (Eq. (11)). We then need to estimate the variance term. Section 5.3 offers a variance estimator.

Once the variance is estimated by a function $V(\mathbf{y}_1, \mathbf{y}_2)$, first-order corrected DPO estimates $\Pr(\mathbf{y}_2 \succ \mathbf{y}_1 \mid \pi)$ using

$$\sigma(h(\mathbf{y}_1, \mathbf{y}_2; \pi)) + \frac{\alpha}{2} \sigma''(h(\mathbf{y}_1, \mathbf{y}_2; \pi)) \cdot V(\mathbf{y}_1, \mathbf{y}_2).$$

Here, $\alpha > 0$ determines the strength of correction, and

$$h(\mathbf{y}_1, \mathbf{y}_2; \pi) := \beta \log \frac{\pi(\mathbf{y}_2)}{\pi_{\text{ref}}(\mathbf{y}_2)} - \beta \log \frac{\pi(\mathbf{y}_1)}{\pi_{\text{ref}}(\mathbf{y}_1)} \quad (15)$$

denotes the difference of π 's induced rewards. Given $\Pr(\mathbf{y}_2 \succ \mathbf{y}_1 \mid \pi)$ and a preference dataset \mathcal{D} , we can maximize the log-likelihood similar to Eq. (6). For numerical stability, we use a stable logarithm $\widetilde{\log}(z) := \log(\max\{z, \epsilon\})$ in computations. Note that our theory suggests $\alpha = 1$ while DPO uses $\alpha = 0$. Our empirical findings in Section 7.2 show that larger values of α improve the effectiveness of the correction. We next discuss the estimation of $V(\mathbf{y}_1, \mathbf{y}_2)$.

5.2 Impossibility without Annotator Information

If we limit our algorithms to M-estimators, which encompass most practical learning methods, consistent estimation of the variance term is impossible with anonymous data:

Proposition 5.2. *There is no M-estimator that can estimate $V(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2) := \text{Var}_u[\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u)]$ consistently without annotator information.*

See proof on page 23. While Proposition 5.1 already implies that we cannot learn π^* without annotation information, Proposition 5.2 goes further, showing that even improving DPO with a first-order approximation is practically impossible. Next, we show how minimal annotation information can overcome this impossibility.

5.3 Using Paired Preferences

We can get around Proposition 5.2 by collecting additional information on annotators. Specifically, it suffices to have a dataset \mathcal{D} of pairs of preferences in the form $\{(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, o), (\mathbf{x}', \mathbf{y}'_1, \mathbf{y}'_2, o')\}$ where $o = \mathbb{1}\{\mathbf{y}_2 \succ \mathbf{y}_1\}$, $o' = \mathbb{1}\{\mathbf{y}'_2 \succ \mathbf{y}'_1\}$ are labeled by the *same person*. Using \mathcal{D} , we can train a *joint likelihood model* $J(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{x}', \mathbf{y}'_1, \mathbf{y}'_2)$ by minimizing cross-entropy between J and $(o \cdot o')$ as the label. The joint likelihood model consistently estimates

$$\mathbb{E}_u \left[\sigma(\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u)) \cdot \sigma(\Delta r^*(\mathbf{x}', \mathbf{y}'_1, \mathbf{y}'_2; u)) \right].$$

This is in fact sufficient to estimate the variance term:

Lemma 5.3. *Using J_1 and J_2 as shorthands for $J(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$ and $J(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{x}, \mathbf{y}_2, \mathbf{y}_1)$, we can use the following to estimate the variance term:*

$$V(\mathbf{y}_1, \mathbf{y}_2) = \frac{J_1 - (J_1 + J_2)^2}{\sigma'(\Delta\bar{r}^*(\mathbf{y}_1, \mathbf{y}_2))^2}. \quad (16)$$

See proof on page 24. Note that we can substitute $\Delta\bar{r}^*(\mathbf{y}_1, \mathbf{y}_2)$ in terms of log policy ratios from Eq. (11). Thus, we have all the elements to calculate V in Eq. (16). This completes our derivation of first-order corrected DPO.

6 Direct Alignment with Maximum Annotator Information

Recall that learning the optimal policy from anonymous data is impossible, and an approximate improvement to DPO requires only minimal information about the annotations. But what if we collect richer data? Can we design a direct alignment method that consistently learns the optimal policy π^* ? To explore this, we consider a dataset where every sample is labeled by representatives of all user types. We show that consistent direct alignment is possible using this dataset.

Suppose user types are in a finite set \mathcal{U} and equally represented. This assumption makes our negative results stronger. Consider a rich data collection: for every context and candidate pairs $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$, we collect one preference data point from each user type. Let $\mathbf{o} \in \{0, 1\}^{|\mathcal{U}|}$ be the vector that indicates preferences where $o_u = 1$ if user of type $u \in \mathcal{U}$ has preferred \mathbf{y}_2 , and 0 otherwise. Given such a dataset \mathcal{D} with context, candidates, and preferences represented as $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{o})$, our goal is to design a loss function

$$\mathcal{L}(\mathcal{D}; \pi) = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{o}) \in \mathcal{D}} l(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{o}; \pi) \quad (17)$$

such that $\arg \min_{\pi} \mathcal{L}(\mathcal{D}; \pi)$ is a consistent estimator of π^* . Designing such a loss is, in fact, possible. For instance, suppose we only look into the agreement cases in \mathcal{D} where \mathbf{o} is either all one or zero. Conditioned on agreement, we will show that the probability of $\mathbf{y}_2 \succ \mathbf{y}_1$ is proportional to $\exp(\sum_u \Delta r^*(\mathbf{y}_1, \mathbf{y}_2; u))$. We can write this likelihood in terms π^* directly as we have a correspondence between π^* and the difference in user-weighted average rewards (see Eq. (11)). We formally show this possibility through a temperature-adjusted DPO:

Proposition 6.1. *Defining l in Eq. (17) as follows results in a consistent estimation of the optimal policy when preferences follow the BT model:*

$$l(\mathbf{y}_1, \mathbf{y}_2, \mathbf{o}; \pi) = \begin{cases} -\log \sigma(|\mathcal{U}| \cdot h(\mathbf{y}_1, \mathbf{y}_2; \pi)), & \mathbf{o} = \vec{\mathbf{1}}, \\ -\log \sigma(|\mathcal{U}| \cdot h(\mathbf{y}_2, \mathbf{y}_1; \pi)), & \mathbf{o} = \vec{\mathbf{0}}, \\ 0 & o.w. \end{cases}$$

Here, h is the difference of π 's induced rewards (Eq. (15)).

See proof on page 24. Consistent loss function is not unique. We give another example in Proposition B.3, and a systematic way to find such losses in Lemma C.2. In both examples, loss functions reduce to the standard DPO loss when $|\mathcal{U}| = 1$.

While the loss function in Proposition 6.1 benefits from consistency, it only uses samples where all user types have agreed. In other words, it discards a sample with any disagreement. A natural question arises: Can we design a loss function that uses all data, including those with disagreement, while maintaining consistency? Surprisingly, the answer is no:

Theorem 6.2. *Suppose l in Eq. (17) only depends on $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$ through π and π_{ref} . If there are more than three types of user and the preferences follow BT, any loss \mathcal{L} that allows a consistent estimation of the optimal policy discards samples with disagreement, i.e., those with $\mathbf{o} \notin \{\mathbf{0}, \mathbf{1}\}$.*

See proof on page 25. This theorem highlights a tension: To improve efficiency, one must compromise either consistency or direct optimization. The approximate direct alignment method proposed in Section 5 exemplifies forgoing consistency. Next, we discuss an alternative that favors consistency.

An Indirect Practical Solution: Averaging Personalized Rewards. The tradeoff between sample efficiency and consistency arises from the requirement for direct optimization. To regain sample efficiency, we may relax the requirement for direct alignment by training reward models while still avoiding RL. Specifically, we can learn personalized reward models $r(\cdot; u)$ for different user types $u \in \mathcal{U}$, calculate a user-weighted expected reward, and use it to relabel a preference dataset. A dataset labeled

with this average reward makes any direct alignment method applicable and avoids RL. It is both consistent and sample-efficient when personalized reward learning is feasible, but comes at the cost of additional training for each user type and memory to store multiple models. Relabeling followed by DPO is found effective in practice [33].

7 Experiments

We provide empirical evidence for our claims throughout the paper. Section 7.1 extends our sensitivity example in Section 4.3 to a real-world preference dataset. In Section 7.2, we simulate DPO and our proposed improvements in a synthetic, small-scale environment where we can visualize and compare the resulting policies. Finally, we scale this experiment in Section 7.3 by fine-tuning large language models, illustrating the extent of improvement over DPO. Our code for reproducing the experimental results is publicly available at <https://github.com/arashne/dahp>.

7.1 NBC Sensitivity to Sampling Distribution

Recall from Section 4.2 that the common practice of alignment assuming homogeneity results in ordinal consistency with NBC. Here, we analyze the sensitivity of NBC to the distribution of pairwise preference datasets in real-world cases by using Pew surveys [34], the same dataset used in Section 4.3. Specifically, we address two questions: (i) Across the questions in the Pew surveys, how often would NBC rankings change when the sampling distribution of alternatives varies (while retaining support over all alternatives)? (ii) How much must the sampling distribution deviate from uniform to alter NBC rankings?

To answer these questions, we first estimate the reward of each option in each question (see Section 4.3 and Appendix D for further details). Given the rewards, we can calculate NBC under any sampling distribution using Eq. (13).

For question (i), among 1519 questions from 19 Pew surveys, NBC rankings change due to changing the sampling distribution from uniform in 20% of cases (306 questions), with the preferred choice changing in 136 cases. For question (ii), we find that a modest change in the sampling distribution suffices; in half the cases, a total variation (TV) distance of less than 0.23 from uniform alters the rankings. The cumulative distribution function (CDF) of the minimum TV distances required to change NBC rankings is shown in Fig. 2. Further experimental details are in Appendix D.

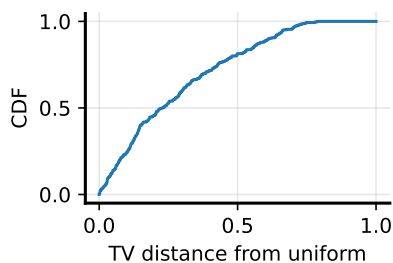


Figure 2: CDF of the minimum TV distances (from uniform) required to change the NBC order in the Pew surveys. A change of 0.23 is sufficient to change the order in half the questions.

7.2 Synthetic Experiments

We generalize the discrete environment of Xu et al. [35] with a heterogeneous population. This environment enables us to visualize the differences between DPO’s policy and the optimal policy, as well as to evaluate the effectiveness of applying a first-order correction (Section 5) and using a consistent loss function (Section 6).

Environment. A prompt x can take a value from 1 to n . There are also only n possible responses to each x . The reward for responding y to x for a type u is $r^*([x, y]; u) = R_u(\text{dist}(x + u, y))$, where dist is a circular distance, and R_u is a linearly decreasing function floored at zero. In this experiment, we set $n = 40$ and consider three equally represented types: $\mathcal{U} = \{-10, 0, 10\}$. We assume BT annotators.

Since the reward (and thus the policies) depends only on $y - x$, we can reduce everything to a 1D representation by setting $\delta := y - x$ and averaging over x . For example, for a policy π , define a 1D policy $\pi(\delta) := \frac{1}{n} \sum_{x \in [n]} \pi(x + \delta | x)$, and similarly a 1D reward $r^*(\delta; u)$. We also compute standard errors of these 1D representations across x . Fig. 3 shows our choice of rewards as well as the expected reward across user types.

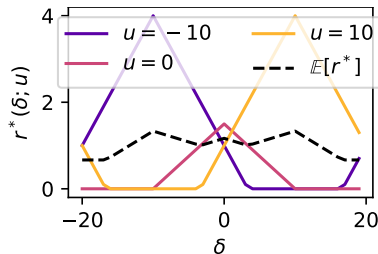


Figure 3: Rewards in the synthetic experiments

Policies. For a uniform π_{ref} and $\beta = 1$, Eq. (10) implies $\pi^*(y | x) \propto \exp(\frac{1}{3} \sum_{u \in \mathcal{U}} r^*([x, y]; u))$. We generate a large dataset of preferences under uniform context and alternative distributions and use the Adam optimizer to minimize the loss for different methods. For the first-order correction of DPO, we additionally train a joint likelihood model J to estimate the variance term V from Eq. (16). We use the loss from Proposition 6.1 as our choice for the consistent loss. This loss effectively utilizes less data compared to other methods since it throws away data points with disagreement. Refer to the accompanying code for more details.

Results. Fig. 4(a) presents π_{DPO} along with π^* and NBC. Unlike the optimal policy which prefers δ around -10 and 10 , DPO prefers $\delta \approx 0$. To a large extent, DPO’s policy is ordinally consistent with NBC.

Fig. 4(b) shows that increasing correction strength α brings the corrected DPO policy closer to π^* . In particular, at $\alpha = 1$, the corrected DPO already favors alternatives with $\delta \in \{-10, 10\}$, consistent with π^* . Furthermore, increasing α makes these alternatives even more favorable. In the full-information setting, Fig. 4(c) shows that minimizing the consistent loss largely leads to π^* . Note that minor deviations from theoretical derivations are likely due to limited data and imperfect optimization in these experiments.

7.3 Semi-Synthetic Experiments

To demonstrate our findings in a realistic setup, we use LoRA [36] to fine-tune Llama-3-8B [37] for both reward learning and direct alignment on two relabeled variations of the HH-RLHF dataset [38], which contains user prompts with pairs of chatbot responses. To simulate heterogeneous preferences, we define three user types with distinct length-based rewards: the first type prefers long, the second type prefers short, and the third type prefers mid-length prompt response combinations (see Appendix E for details). Recall from Section 3 that we argued for the average reward across user types as the proper objective for alignment with heterogeneous preferences. Therefore, we use agreement with the ground-truth average reward on the test set as the success metric.

First, we use vanilla reward learning and DPO with the homogeneity assumption on an anonymous preference dataset where every sample is labeled with a random user type. As Fig. 5 shows in blue,

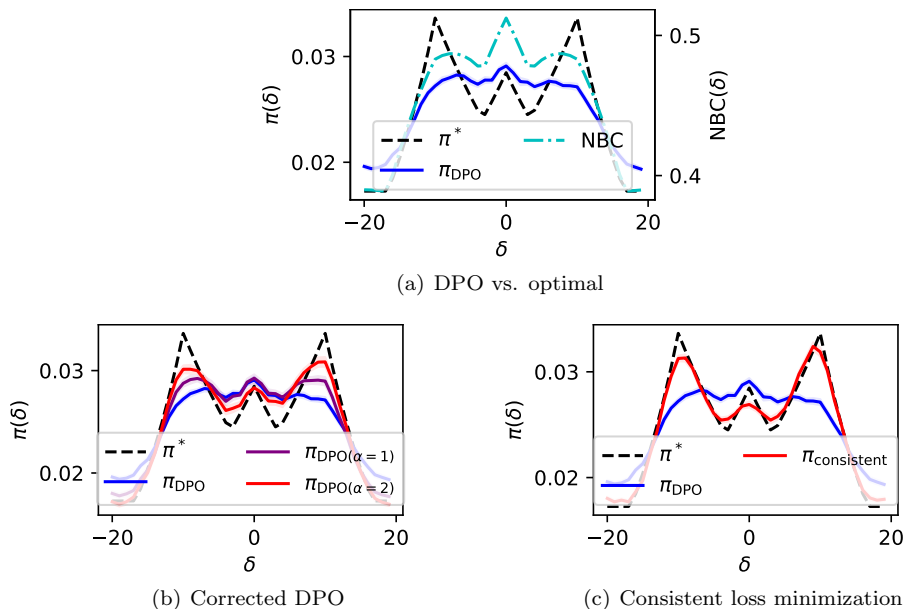


Figure 4: Policies explicitly accounting for heterogeneity are more consistent with the average reward across types in a synthetic setup.

their induced order agrees with the average reward in 89.6% and 67.4% of the test cases, respectively. Next, we use the loss function in Proposition 6.1 on a dataset with maximum annotator information (Section 6). To create this dataset, for every response pair to a prompt, we sample the preferences of the three user types until a consensus is reached, using the agreed-upon preference as the label. As Fig. 5 shows in red, the learned reward and the induced reward by the learned policy agree with the average reward in 93.9% and 71.7% of the test cases. In summary, explicitly accounting for heterogeneity increases the agreement with the average reward across user types by 4.3%—an additional 368 test cases—for both reward learning and direct alignment.

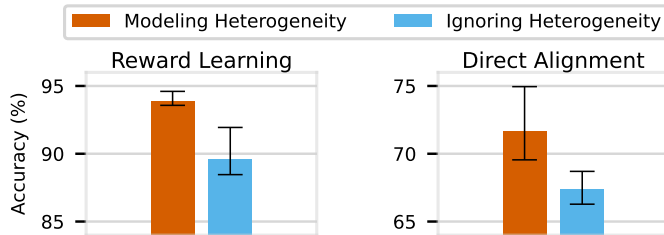


Figure 5: In the presence of preference labels from every user type, our proposed loss function produces reward models (left) and aligned policies (right) that are more consistent with the average reward across user types, compared to typical approaches that overlook heterogeneity. Bars show the mean, and whiskers denote the second and third quartiles across five random seeds.

8 Related Work

The challenge of handling heterogeneous preferences in alignment has been recognized as a significant problem in alignment research [39, 40, 11, 12]. This problem has attracted considerable attention from

researchers in the field. Here, we highlight a few representative works that address key directions in tackling this challenge.

Analysis of DPO. Our study of how standard preference learning methods, such as DPO, behave in the presence of heterogeneous preferences was inspired by Siththaranjan, Laidlaw, and Hadfield-Menell [15]’s result, which shows that RLHF aggregates preferences according to a well-known voting rule called Borda count. Chakraborty et al. [27] highlights the impossibility of aligning with a singular reward model in RLHF by providing a lower bound on the gap between the optimal policy and a subpopulation’s optimal policy. Dumoulin et al. [13] adopts a density estimation perspective on learning from human feedback to illustrate the challenges of preference learning from a population of annotators with diverse viewpoints. Rosset et al. [41] and Gao et al. [42] point out the limitations of point-wise reward models in expressing complex, intransitive preferences that may arise due to the aggregation of diverse preferences. Additionally, frameworks that generalize DPO and unify different alignment methods have been proposed to analyze current approaches and explore possible alternatives [43, 44, 30, 45].

Policy Personalization. Many works in the literature have proposed personalization as a solution to the problem of pluralistic alignment. Poddar et al. [46] propose a latent variable formulation of the problem and learn rewards and policies conditioned on it. Chen et al. [47] use an ideal point model for preferences and learn latent spaces representing different preferences. Mapping user information to user representations, Li, Lipton, and Leqi [48] perform personalized DPO to jointly learn a user model and a personalized language model. Balepur et al. [49] use abductive reasoning to infer user personas and train models to tailor responses accordingly. Lee et al. [50] explore the possibility of steering a language model to align with a user’s intentions through system messages. Dang et al. [51] extend personalized alignment to text-to-image diffusion models. Jang et al. [52] perform personalized alignment by decomposing preferences into multiple dimensions. Lau et al. [53] dynamically adapt the model to individual preferences using in-context learning.

Preference Aggregation. Closely aligned with our goal of serving the entire population with a single policy, several works have explored ways to aggregate diverse preferences. The rich literature on social choice theory has proven to be a valuable source of inspiration for studying existing preference learning approaches and proposing new ones [10, 54, 55, 11, 56]. Drawing insights from social choice theory, robustness to approximate clones has been proposed as a desirable property of RLHF algorithms, which current methods lack [29]. The Minimax Winner, a concept in preference aggregation, has inspired the use of the proportion of wins as the reward for a particular trajectory to align a model through self-play [57]. The impact of heterogeneity on strategic behavior in feedback and its effects on aggregation are also explored in Park et al. [14], which further examines the use of different social welfare functions for preference aggregation.

Methods. Solutions proposed to address different formulations of the problem span a wide range of methods. Siththaranjan, Laidlaw, and Hadfield-Menell [15] estimate a distribution of scores for alternatives to account for heterogeneity as hidden context. Chidambaram, Seetharaman, and Syrgkanis [58] propose an Expectation-Maximization (EM) version of DPO to minimize a notion of worst-case regret. Multi-objective reinforcement learning [59, 52] and its direct optimization variant [60] have also been proposed to align with diverse preferences. Wang et al. [61] train a multi-objective reward model to capture diverse preferences. Zhong et al. [62] use meta-learning to learn diverse preferences and aggregate them using different social welfare functions. Li et al. [63] design an optimal-transport-based loss to calibrate their model with the categorical distribution of preferences. Producing a Pareto front of models has also been explored as a solution. Boldi et al. [64] employ an iterative process to select solutions, while Rame et al. [65] interpolate the weights of independent networks linearly to achieve a Pareto-optimal generalization across preferences.

Empirical Observations. Empirical studies of alignment methods have had a significant impact on the study of preference learning. Zhang et al. [5] demonstrate the Bradley-Terry model’s failure to distinguish between unanimous agreement among annotators and the majority opinion in cases of diverging user preferences. Chen et al. [66] show that RLHF and DPO struggle to improve ranking accuracy. Zeng et al. [67] study the role of model size and data size in the impact of diversified human preferences. Bansal, Dang, and Grover [68] demonstrate the significant influence of feedback protocol choice on alignment evaluation. Santurkar et al. [69] explore the opinions reflected by a language model, while Bakker et al. [70] investigate a language model’s ability to generate consensus statements by training it to predict individual preferences. Jiang et al. [71] propose individualistic alignment to predict an individual’s values, and Zollo et al. [72] introduce the PersonalLLM benchmark to measure a model’s adaptation to a particular user’s preferences.

9 Discussion and Conclusion

Aligning a single policy to the average reward across user types in a heterogeneous population requires collecting annotator information. This can range from minimal information such as linking two instances labeled by the same annotator, to richer information like using questionnaires to infer annotator types. We improved DPO using the former and introduced a consistent loss when annotators from all user types label every data point. With additional assumptions, unsupervised methods might be able to identify annotator types from anonymous datasets [73]. Further research should explore the additional structures that, when used during data collection, can help with identifiability.

Our results revealed a tension between consistency and sample efficiency in direct alignment. Thus, an alternative approach—individual reward training and aggregation—may be more practical for addressing heterogeneity when individual rewards are identifiable.

We use the average reward across user types, as the natural choice among aggregations that define a well-defined alignment problem from pairwise preferences. However, the choice of aggregation is inherently a social and policy question rather than purely a technical one. In certain contexts, the policymaker might prefer to give higher weight to disadvantaged people to address issues such as inequality. Additionally, in some cases, the very existence of a reward function may be questionable, requiring objectives to be defined in terms of choice probabilities rather than rewards.

We believe that trained policies should not be used to elicit or represent aggregate preferences, even when reward aggregation is appropriate and estimation is consistent. While such policies may capture certain patterns in user behavior, they do not necessarily reflect the underlying interests or values of the population. In other words, we view the resulting policy as a functional tool for decision-making rather than a true representation of users’ collective interests.

In summary, while preference heterogeneity is well recognized in mathematical psychology, standard methods often bypass this complexity. As we showed, accounting for heterogeneity, even when the goal remains the same as in the homogeneous setting—to derive a single policy—can render common techniques inefficient or inapplicable. Understanding these limitations calls for new approaches that explicitly incorporate heterogeneity while effectively balancing efficiency, consistency, and practicality.

Acknowledgment

We thank Mohammad Alizadeh, Pouya Hamadianian, Moritz Hardt, Erfan Jahanparast, and Itai Shapira for discussions and feedback on earlier versions of this paper. Abebe was partially supported by the Andrew Carnegie Fellowship Program. Procaccia was partially supported by the National Science Foundation under grants IIS-2147187 and IIS-2229881; by the Office of Naval Research under grants N00014-24-1-2704 and N00014-25-1-2153; and by a grant from the Cooperative AI Foundation.

References

- [1] Lora Aroyo and Chris Welty. “Truth is a lie: Crowd truth and the seven myths of human annotation”. In: *AI Magazine* 36.1 (2015), pp. 15–24.
- [2] Ellie Pavlick and Tom Kwiatkowski. “Inherent disagreements in human textual inferences”. In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 677–694.
- [3] Remi Denton et al. “Whose ground truth? Accounting for individual and collective identities underlying dataset annotation”. In: *arXiv preprint arXiv:2112.04554* (2021).
- [4] Marta Sandri et al. “Why Don’t You Do It Right? Analysing Annotators’ Disagreement in Subjective Tasks”. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2428–2441. DOI: [10.18653/v1/2023.eacl-main.178](https://doi.org/10.18653/v1/2023.eacl-main.178). URL: <https://aclanthology.org/2023.eacl-main.178/>.
- [5] Michael JQ Zhang et al. “Diverging Preferences: When do Annotators Disagree and do Models Know?”. In: *arXiv preprint arXiv:2410.14632* (2024).
- [6] Daniel M Ziegler et al. “Fine-tuning language models from human preferences”. In: *arXiv preprint arXiv:1909.08593* (2019).
- [7] Nisan Stiennon et al. “Learning to summarize with human feedback”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3008–3021.
- [8] Long Ouyang et al. “Training language models to follow instructions with human feedback”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27730–27744.
- [9] Rafael Rafailov et al. “Direct preference optimization: Your language model is secretly a reward model”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [10] Vincent Conitzer et al. “Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback”. In: *arXiv preprint arXiv:2404.10271* (2024).
- [11] Luise Ge et al. “Axioms for AI Alignment from Human Feedback”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [12] Taylor Sorensen et al. “Position: A Roadmap to Pluralistic Alignment”. In: *Forty-first International Conference on Machine Learning*. 2024. URL: <https://openreview.net/forum?id=gQpBnRHwxM>.
- [13] Vincent Dumoulin et al. “A density estimation perspective on learning from pairwise human preferences”. In: *arXiv preprint arXiv:2311.14115* (2023).
- [14] Chanwoo Park et al. “RLHF from heterogeneous feedback via personalization and preference aggregation”. In: *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*. 2024.
- [15] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. “Distributional preference learning: Understanding and accounting for hidden context in RLHF”. In: *arXiv preprint arXiv:2312.08358* (2023).
- [16] Tomasz Korbak, Ethan Perez, and Christopher Buckley. “RL with KL penalties is better viewed as Bayesian inference”. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. 2022, pp. 1083–1091.
- [17] Yao Zhao et al. “SLiC-HF: Sequence likelihood calibration with human feedback”. In: *arXiv preprint arXiv:2305.10425* (2023).
- [18] Shangmin Guo et al. “Direct language model alignment from online AI feedback”. In: *arXiv preprint arXiv:2402.04792* (2024).
- [19] LL Thurstone. “A law of comparative judgment.” In: *Psychological Review* 34.4 (1927), p. 273.

- [20] R Duncan Luce. *Individual choice behavior*. Vol. 4. Wiley New York, 1959.
- [21] John I Yellott Jr. “The relationship between Luce’s choice axiom, Thurstone’s theory of comparative judgment, and the double exponential distribution”. In: *Journal of Mathematical Psychology* 15.2 (1977), pp. 109–144.
- [22] Ralph Allan Bradley and Milton E Terry. “Rank analysis of incomplete block designs: I. The method of paired comparisons”. In: *Biometrika* 39.3/4 (1952), pp. 324–345.
- [23] Brian D Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University, 2010.
- [24] Roger N. Shepard. “Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space”. In: *Psychometrika* 22.4 (Dec. 1957), pp. 325–345. ISSN: 1860-0980. DOI: [10.1007/BF02288967](https://doi.org/10.1007/BF02288967). URL: <https://doi.org/10.1007/BF02288967>.
- [25] Lucas Monteiro Paes et al. “On the epistemic limits of personalized prediction”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 1979–1991.
- [26] Hannah Rose Kirk et al. “The benefits, risks and bounds of personalizing the alignment of large language models to individuals”. In: *Nature Machine Intelligence* (2024), pp. 1–10.
- [27] Souradip Chakraborty et al. “MaxMin-RLHF: Towards equitable alignment of large language models with diverse human preferences”. In: *arXiv preprint arXiv:2402.08925* (2024).
- [28] Mamoru Kaneko and Kenjiro Nakamura. “The Nash social welfare function”. In: *Econometrica: Journal of the Econometric Society* (1979), pp. 423–435.
- [29] Ariel D. Procaccia, Benjamin Schiffer, and Shirley Zhang. “Clone-Robust AI Alignment”. In: *arXiv preprint arXiv:2501.09254* (2025).
- [30] Mohammad Gheshlaghi Azar et al. “A general theoretical paradigm to understand learning from human preferences”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024, pp. 4447–4455.
- [31] Elliot Anshelevich et al. “Distortion in social choice problems: The first 15 years and beyond”. In: *arXiv preprint arXiv:2103.00911* (2021).
- [32] Ehud Friedgut, Gil Kalai, and Assaf Naor. “Boolean functions whose Fourier transform is concentrated on the first two levels”. In: *Advances in Applied Mathematics* 29.3 (2002), pp. 427–437.
- [33] Evan Frick et al. “How to Evaluate Reward Models for RLHF”. In: *arXiv preprint arXiv:2410.14872* (2024).
- [34] Pew Research Center. *About Pew Research Center*. Accessed: 2025-01-20. 2025. URL: <https://www.pewresearch.org/about/>.
- [35] Shusheng Xu et al. “Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study”. In: *Forty-first International Conference on Machine Learning*. 2024.
- [36] Edward J Hu et al. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *International Conference on Learning Representations*. 2022. URL: <https://openreview.net/forum?id=nZeVKeeFYf9>.
- [37] AI@Meta. “Llama 3 Model Card”. In: (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- [38] Yuntao Bai et al. “Training a helpful and harmless assistant with reinforcement learning from human feedback”. In: *arXiv preprint arXiv:2204.05862* (2022).
- [39] Usman Anwar et al. “Foundational Challenges in Assuring Alignment and Safety of Large Language Models”. In: *Transactions on Machine Learning Research* (2024). Survey Certification, Expert Certification. ISSN: 2835-8856. URL: <https://openreview.net/forum?id=oVTkOs8Pka>.

- [40] Stephen Casper et al. “Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback”. In: *Transactions on Machine Learning Research* (2023). Survey Certification, Featured Certification. ISSN: 2835-8856. URL: <https://openreview.net/forum?id=bx24KpJ4Eb>.
- [41] Corby Rosset et al. “Direct Nash optimization: Teaching language models to self-improve with general preferences”. In: *arXiv preprint arXiv:2404.03715* (2024).
- [42] Zhaolin Gao et al. “Rebel: Reinforcement learning via regressing relative rewards”. In: *arXiv preprint arXiv:2404.16767* (2024).
- [43] Haoxian Chen et al. “Mallows-DPO: Fine-Tune Your LLM with Preference Dispersions”. In: *arXiv preprint arXiv:2405.14953* (2024).
- [44] Yunhao Tang et al. “Generalized Preference Optimization: A Unified Approach to Offline Alignment”. In: *Forty-first International Conference on Machine Learning*. 2024. URL: <https://openreview.net/forum?id=gu3nacA9AH>.
- [45] Yu Meng, Mengzhou Xia, and Danqi Chen. “SimPO: Simple Preference Optimization with a Reference-Free Reward”. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024. URL: <https://openreview.net/forum?id=3Tzcot1LKb>.
- [46] Sriyash Poddar et al. “Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning”. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024. URL: <https://openreview.net/forum?id=gRG6SzbW9p>.
- [47] Daiwei Chen et al. “PAL: Pluralistic Alignment Framework for Learning from Heterogeneous Preferences”. In: *arXiv preprint arXiv:2406.08469* (2024).
- [48] Xinyu Li, Zachary C Lipton, and Liu Leqi. “Personalized language modeling from personalized human feedback”. In: *arXiv preprint arXiv:2402.05133* (2024).
- [49] Nishant Balepur et al. *Whose Boat Does it Float? Improving Personalization in Preference Tuning via Inferred User Personas*. 2025. arXiv: 2501.11549 [cs.CL]. URL: <https://arxiv.org/abs/2501.11549>.
- [50] Seongyun Lee et al. “Aligning to Thousands of Preferences via System Message Generalization”. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024. URL: <https://openreview.net/forum?id=recsheq7e8>.
- [51] Meihua Dang et al. “Personalized Preference Fine-tuning of Diffusion Models”. In: *arXiv preprint arXiv:2501.06655* (2025).
- [52] Joel Jang et al. “Personalized soups: Personalized large language model alignment via post-hoc parameter merging”. In: *arXiv preprint arXiv:2310.11564* (2023).
- [53] Allison Lau et al. “Personalized Adaptation via In-Context Preference Learning”. In: *arXiv preprint arXiv:2410.14001* (2024).
- [54] Tianyi Qiu. “Representative Social Choice: From Learning Theory to AI Alignment”. In: *arXiv preprint arXiv:2410.23953* (2024).
- [55] Parand A Alamdari, Soroush Ebadian, and Ariel D Procaccia. “Policy aggregation”. In: *Advanced in Neural Information Processing Systems* 36 (2024).
- [56] Jessica Dai and Eve Fleisig. “Mapping social choice theory to RLHF”. In: *arXiv preprint arXiv:2404.13038* (2024).
- [57] Gokul Swamy et al. “A Minimaximalist Approach to Reinforcement Learning from Human Feedback”. In: *Forty-first International Conference on Machine Learning*. 2024. URL: <https://openreview.net/forum?id=5kVgd2MwMY>.
- [58] Keertana Chidambaram, Karthik Vinay Seetharaman, and Vasilis Syrgkanis. *Direct Preference Optimization With Unobserved Preference Heterogeneity*. 2024. URL: <https://openreview.net/forum?id=NQZNNUsutn>.

- [59] Hadassah Harland et al. “Adaptive Alignment: Dynamic Preference Adjustments via Multi-Objective Reinforcement Learning for Pluralistic AI”. In: *arXiv preprint arXiv:2410.23630* (2024).
- [60] Zhanhui Zhou et al. “Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization”. In: *Findings of the Association for Computational Linguistics ACL 2024*. 2024, pp. 10586–10613.
- [61] Haoxiang Wang et al. “Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards”. In: *arXiv preprint arXiv:2402.18571* (2024).
- [62] Huiying Zhong et al. “Provable multi-party reinforcement learning with diverse human feedback”. In: *arXiv preprint arXiv:2403.05006* (2024).
- [63] Dexun Li et al. “Aligning crowd feedback via distributional preference reward modeling”. In: *arXiv preprint arXiv:2402.09764* (2024).
- [64] Ryan Boldi et al. “Pareto-optimal learning from preferences with hidden context”. In: *arXiv preprint arXiv:2406.15599* (2024).
- [65] Alexandre Rame et al. “Rewarded soups: Towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [66] Angelica Chen et al. “Preference Learning Algorithms Do Not Learn Preference Rankings”. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024. URL: <https://openreview.net/forum?id=YkJ5BuEXdD>.
- [67] Dun Zeng et al. *On Diversified Preferences of Large Language Model Alignment*. 2024. arXiv: [2312.07401](https://arxiv.org/abs/2312.07401) [cs.AI]. URL: <https://arxiv.org/abs/2312.07401>.
- [68] Hritik Bansal, John Dang, and Aditya Grover. “Peering Through Preferences: Unraveling Feedback Acquisition for Aligning Large Language Models”. In: *The Twelfth International Conference on Learning Representations*. 2024. URL: <https://openreview.net/forum?id=dKl6lMwbCy>.
- [69] Shibani Santurkar et al. “Whose opinions do language models reflect?” In: *International Conference on Machine Learning*. PMLR. 2023, pp. 29971–30004.
- [70] Michiel Bakker et al. “Fine-tuning language models to find agreement among humans with diverse preferences”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 38176–38189.
- [71] Liwei Jiang et al. “Can Language Models Reason about Individualistic Human Values and Preferences?” In: *arXiv preprint arXiv:2410.03868* (2024).
- [72] Thomas P. Zollo et al. *PersonalLLM: Tailoring LLMs to Individual Preferences*. 2024. arXiv: [2409.20296](https://arxiv.org/abs/2409.20296) [cs.LG]. URL: <https://arxiv.org/abs/2409.20296>.
- [73] Xiaomin Zhang et al. “On the identifiability of mixtures of ranking models”. In: *arXiv preprint arXiv:2201.13132* (2022).
- [74] Ilya Loshchilov and Frank Hutter. “Decoupled Weight Decay Regularization”. In: *International Conference on Learning Representations*. 2019. URL: <https://openreview.net/forum?id=Bkg6RiCqY7>.

A Additional Drawbacks of DPO under Heterogeneity

Violating Independence of Irrelevant Alternatives (IIA). Suppose $\mathcal{U} = \{A, B\}$ and types are equally represented. Given two possible responses \mathbf{y}_1 and \mathbf{y}_2 , type A prefers \mathbf{y}_1 but type B prefers \mathbf{y}_2 :

$$\begin{aligned} r^*(\mathbf{y}_1; A) &= 6, r^*(\mathbf{y}_2; A) = 1, \\ r^*(\mathbf{y}_1; B) &= 3, r^*(\mathbf{y}_2; B) = 9. \end{aligned}$$

A direct calculation shows $\mathbb{E}_u[r^*(\mathbf{y}_2)] > \mathbb{E}_u[r^*(\mathbf{y}_1)]$ and $\text{NBC}(\mathbf{y}_2) > \text{NBC}(\mathbf{y}_1)$. So, both π^* and π_{DPO} prefer \mathbf{y}_2 . Let's consider another possible response \mathbf{y}_3 which is not the preferred response for any user type:

$$r^*(\mathbf{y}_3; A) = r^*(\mathbf{y}_3; B) = 2.$$

While π^* still prefers \mathbf{y}_2 to \mathbf{y}_1 , now $\text{NBC}(\mathbf{y}_1) \approx 0.62 > \text{NBC}(\mathbf{y}_2) \approx 0.55$, so, introducing an irrelevant alternative can alter DPO's ranking over existing alternatives.

Tyranny of Majority. Suppose $\mathcal{U} = \{A, B\}$ with type A shaping 90% of the population. Given two responses $\mathbf{y}_1, \mathbf{y}_2$, type A slightly favors \mathbf{y}_2 but type B finds \mathbf{y}_2 offensive:

$$\begin{aligned} r^*(\mathbf{y}_1; A) &= 0.5, r^*(\mathbf{y}_2; A) = 1, \\ r^*(\mathbf{y}_1; B) &= 0.5, r^*(\mathbf{y}_2; B) = -10. \end{aligned}$$

In this case, π^* prefers \mathbf{y}_1 even though type B is a minority. In contrast, we have $\text{NBC}(\mathbf{y}_1) \approx 0.47, \text{NBC}(\mathbf{y}_2) \approx 0.53$, which implies that the majority dominates in DPO.

B Additional Statements

Proposition B.1. *There exists a mixture of BTs that a single BT cannot represent.*

See [proof](#) on page 25.

Definition B.2 (Learnability). Denote by $\mathcal{D}_{r,\sigma}$ an i.i.d. sampled pairwise preference dataset labeled by random users with reward r and preference model σ . Let $\bar{r}(\mathbf{y}) := \mathbb{E}_u[r(\mathbf{y}; u)]$. We say that the ranking based on \bar{r} is (weakly) learnable if, for some $\epsilon > 0$, there exists an algorithm with a bounded sample complexity m , such that for every reward r , when given a dataset $\mathcal{D}_{r,\sigma}$ of size $|\mathcal{D}_{r,\sigma}| \geq m(\epsilon, \bar{r})$, it outputs a ranking consistent with \bar{r} with a probability at least ϵ above the chance level.

Proposition B.3. *Defining l in Eq. (17) as follows results in a consistent estimation of the optimal policy when preferences follow the BT model:*

$$l(\mathbf{y}_1, \mathbf{y}_2, \mathbf{o}; \pi) = \begin{cases} -\sigma(h(\mathbf{y}_1, \mathbf{y}_2; \pi)) - I(\sigma(h(\mathbf{y}_1, \mathbf{y}_2; \pi))), & \mathbf{o} = \mathbf{1}, \\ -\sigma(h(\mathbf{y}_2, \mathbf{y}_1; \pi)) - I(\sigma(h(\mathbf{y}_2, \mathbf{y}_1; \pi))), & \mathbf{o} = \mathbf{0}, \\ 0 & o.w. \end{cases}$$

Here, we define $I(\theta) := \int_1^\theta (\frac{1}{\theta'} - 1)^{|\mathcal{U}|} d\theta'$, and h is the difference of π 's induced rewards (Eq. (15)).

See [proof](#) on page 26.

C Missing Proofs

Proposition 3.1. Consider an aggregation $f : \mathbb{R}^{\mathcal{U}} \rightarrow \mathbb{R}$. If $f(\{r(\mathbf{y}; u)\}_{u \in \mathcal{U}})$ induces the same ordering over \mathbf{y} for every reward r consistent with the preferences distribution, then under weak regulatory assumptions, f must be affine.

Proof of Proposition 3.1. First of all, if a reward function $r^*(\mathbf{y}; u)$ can explain the preferences of a user type u , any other reward function $r(\mathbf{y}; u) := r^*(\mathbf{y}; u) + c(u)$ induces the same preference distribution:

$$\Pr(\mathbf{y}_2 \succ \mathbf{y}_1 \mid r; u) = \sigma(r(\mathbf{y}_2; u) - r(\mathbf{y}_1; u)) = \sigma(r^*(\mathbf{y}_2; u) - r^*(\mathbf{y}_1; u)) = \Pr(\mathbf{y}_2 \succ \mathbf{y}_1 \mid r^*; u).$$

Therefore, r^* is identifiable up to a bias term that can depend on the context and user type.

Consider a reward aggregation function $f : \mathbb{R}^{\mathcal{U}} \rightarrow \mathbb{R}$. Denoting all the rewards from different user types by a vector $\mathbf{r}(\mathbf{y}) \in \mathbb{R}^{\mathcal{U}}$, the aggregation $f(\mathbf{r}(\mathbf{y}))$ should induce the same ranking for every \mathbf{r} consistent with (possibly infinite) preference data. Our above argument then implies that $f(\mathbf{r}^*(\mathbf{y}) + \mathbf{c})$ should induce a consistent ranking for every $\mathbf{c} \in \mathbb{R}^{\mathcal{U}}$. For a sufficiently large space of alternatives, where $\mathbf{r}^*(\mathbf{y})$ can take any value within a closed interval of \mathbb{R} , this is possible only if there exists a function $\psi : \mathbb{R}^{\mathcal{U}} \rightarrow \mathbb{R}$ such that

$$f(\mathbf{r}_2 + \mathbf{c}) - f(\mathbf{r}_1 + \mathbf{c}) = \psi(\mathbf{r}_2 - \mathbf{r}_1),$$

for every \mathbf{c} , \mathbf{r}_1 , and \mathbf{r}_2 in $\mathbb{R}^{\mathcal{U}}$. Choosing $\mathbf{c} = -\mathbf{r}_1$, this implies $f(\mathbf{r}) = f(\mathbf{0}) + \psi(\mathbf{r})$ for every \mathbf{r} . Therefore, we have the following Cauchy functional equation for ψ :

$$\psi(\mathbf{r} + \mathbf{\Delta}) = \psi(\mathbf{r}) + \psi(\mathbf{\Delta}).$$

Under weak regularity conditions such as the monotonicity or continuity of f , it is well-known that ψ has to be a linear function. This implies that f has to be an affine function which completes the proof. \square

Proposition 4.2. Suppose responses to \mathbf{x} in the preference dataset are drawn from $\mathcal{D}(\cdot \mid \mathbf{x})$. In the limit of many data points, DPO’s induced reward, or equivalently $\frac{\pi_{\text{DPO}}(\cdot \mid \mathbf{x})}{\pi_{\text{ref}}(\cdot \mid \mathbf{x})}$, has the same ordering over responses as $\text{NBC}(\cdot \mid \mathbf{x})$.²

Proof of Proposition 4.2. We start from DPO’s objective in Eq. (6). For notational simplicity, we assume $\pi(\mathbf{y} \mid \mathbf{x})$ already contains a normalization by $\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})$. In the limit of many data points, we can rewrite DPO’s objective as the minimization of a cross-entropy loss

$$\begin{aligned} \mathcal{L}_{\text{DPO}}(\pi) := & -\mathbb{E}_{\mathbf{x}, \mathbf{y}, \mathbf{y}'} \left[\bar{\sigma}(\Delta r^*(\mathbf{x}, \mathbf{y}', \mathbf{y})) \cdot \log \sigma \left(\beta \log \frac{\pi(\mathbf{y} \mid \mathbf{x})}{\pi(\mathbf{y}' \mid \mathbf{x})} \right) \right. \\ & \left. + \left(1 - \bar{\sigma}(\Delta r^*(\mathbf{x}, \mathbf{y}', \mathbf{y})) \right) \cdot \log \left(1 - \sigma \left(\beta \log \frac{\pi(\mathbf{y} \mid \mathbf{x})}{\pi(\mathbf{y}' \mid \mathbf{x})} \right) \right) \right], \end{aligned}$$

where $\bar{\sigma}(\Delta r^*(\mathbf{x}, \mathbf{y}', \mathbf{y}))$ is shorthand for $\Pr(\mathbf{y} \succ \mathbf{y}' \mid \mathbf{x}; r^*) = \mathbb{E}_u[\sigma(r^*([\mathbf{x}, \mathbf{y}]; u) - r^*([\mathbf{x}, \mathbf{y}']; u))]$. The minimizer of \mathcal{L}_{DPO} should meet the first-order condition: $\frac{\partial \mathcal{L}_{\text{DPO}}}{\partial \pi(\mathbf{y} \mid \mathbf{x})} = 0$, for every \mathbf{x} and \mathbf{y} . Then, a direct calculation shows that the optimal policy π^* meets

$$\mathbb{E}_{\mathbf{y}' \sim \mathcal{D}(\cdot \mid \mathbf{x})} \left[\sigma \left(\beta \log \frac{\pi^*(\mathbf{y} \mid \mathbf{x})}{\pi^*(\mathbf{y}' \mid \mathbf{x})} \right) \right] - \mathbb{E}_{\mathbf{y}' \sim \mathcal{D}(\cdot \mid \mathbf{x})} \left[\bar{\sigma}(\Delta r^*(\mathbf{x}, \mathbf{y}', \mathbf{y})) \right] = 0. \quad (18)$$

Recognize that the second term is $\text{NBC}(\mathbf{y} \mid \mathbf{x})$:

$$\text{NBC}(\mathbf{y} \mid \mathbf{x}) = \mathbb{E}_{\mathbf{y}' \sim \mathcal{D}(\cdot \mid \mathbf{x})} \left[\mathbb{E}_u \left[\sigma(r^*([\mathbf{x}, \mathbf{y}]; u) - r^*([\mathbf{x}, \mathbf{y}']; u)) \right] \right].$$

²We can view $\text{NBC}(\mathbf{y} \mid \mathbf{x})$ as an aggregation of rewards at \mathbf{y} . One can verify that NBC meets the order consistency condition of Proposition 3.1. However, it uses the reward value at $\mathbf{y}' \neq \mathbf{y}$ to define the aggregated reward at \mathbf{y} and thus does not fall under Proposition 3.1. In fact, this interdependency causes the issues we discuss Section 4.3.

In the absence of heterogeneity, we have $\bar{\sigma} = \sigma$, so setting $\beta \log \pi^*(\mathbf{y} \mid \mathbf{x}) = r^*([\mathbf{x}, \mathbf{y}]) + C(\mathbf{x})$ for a normalizing C would solve Eq. (18). In general, we are not aware of any closed-form solution. However, we can still infer the ordering the optimal policy induces from Eq. (18): Since the first term is increasing in $\pi^*(\mathbf{y} \mid \mathbf{x})$, the optimal policy will be monotone in $\text{NBC}(\mathbf{y} \mid \mathbf{x})$. This completes the proof. \square

Proposition 5.1. *If there are at least two alternatives and two user types with a continuous preference model, the ranking based on the user-weighted expected reward is not learnable without annotator information.*

Proof of Proposition 5.1. We give three related proof strategies. The first strategy works for every preference model. The second strategy draws on a connection to a robust version of Arrow’s impossibility theorem [32]. The third strategy is inspired by Procaccia, Schiffer, and Zhang [29]. We start with the notation and definitions specific to this proof.

Notation and Definitions. Consider a fixed prompt \mathbf{x} with a set of possible responses \mathcal{Y} . Let R denote a *complete ranking* over \mathcal{Y} , where $\mathbf{y}_2 R \mathbf{y}_1$ indicates whether $\mathbf{y}_2 \succ \mathbf{y}_1$ or vice versa. A *profile* refers to a set of complete rankings. For a heterogeneous reward function $r(\mathbf{y}; u)$ and a prior \mathcal{P} over user types \mathcal{U} , let $R_{\bar{r}}$ be the ranking according to $\bar{r}(\mathbf{y}) := \mathbb{E}_{u \sim \mathcal{P}}[r(\mathbf{y}; u)]$.

A *pairwise preference dataset* \mathcal{D} consists of tuples $(\mathbf{y}_1, \mathbf{y}_2, o)$, where $o := \mathbb{1}\{\mathbf{y}_2 \succ \mathbf{y}_1\}$. We assume that \mathbf{y}_1 and \mathbf{y}_2 are i.i.d. draws. When a random user with a reward function r labels each instance in the dataset, we denote the resulting dataset by \mathcal{D}_r . A *pairwise learning algorithm* \mathcal{A} produces a complete ranking over \mathcal{Y} based on the pairwise preference dataset \mathcal{D} .

Proof Strategy 1. Suppose there exists an algorithm \mathcal{A} such that for some \mathcal{Y} with $|\mathcal{Y}| \geq 2$, for any reward function r and any preference dataset \mathcal{D}_r with $|\mathcal{D}_r| \geq n_{\bar{r}}$, it outputs $R_{\bar{r}}$ on \mathcal{Y} with a probability of at least $\frac{1}{|\mathcal{Y}|!} + \epsilon$.

Suppose r is a heterogeneous reward function that its expectation induces a complete ranking $R_{\bar{r}}$ with no tie. Define a new heterogeneous reward function r_γ as follows. Consider a new user type $0 \notin \mathcal{U}$. For some $\gamma > 1$, let $r_\gamma(\mathbf{y}; u) = \gamma r(\mathbf{y}; u)$ when $u \neq 0$, and $r_\gamma(\mathbf{y}; 0) = 0$. Define a new user distribution $\mathcal{P}_\gamma(u) := (1 - \frac{1}{\gamma}) \mathbb{1}\{u = 0\} + \frac{1}{\gamma} \mathcal{P}(u)$. It is straightforward to verify $\bar{r}_\gamma := \mathbb{E}_{u \sim \mathcal{P}_\gamma}[r_\gamma(\mathbf{y}; u)] = \bar{r}$. Therefore, with high probability, \mathcal{A} outputs $R_{\bar{r}_\gamma} = R_{\bar{r}}$ from \mathcal{D}_{r_γ} for every $\gamma > 1$:

$$\Pr(\mathcal{A}(\mathcal{D}_{r_\gamma}) = R_{\bar{r}}) \geq \frac{1}{|\mathcal{Y}|!} + \epsilon.$$

As we increase γ , for any continuous preference model σ , the pairwise preference dataset \mathcal{D}_{r_γ} approaches a uniform preference dataset $\mathcal{D}_{\text{unif}}$ labeled mostly by an indifferent annotator of type $u = 0$. So, we have

$$\Pr(\mathcal{A}(\mathcal{D}_{\text{unif}}) = R_{\bar{r}}) \geq \frac{1}{|\mathcal{Y}|!} + \epsilon. \tag{19}$$

This is true for every r . For different choices of r , agreements with $R_{\bar{r}}$ are disjoint events. Since there are $|\mathcal{Y}|!$ different rankings overall, the pigeonholed principle implies $\epsilon = 0$.

Proof Strategy 2. The proof is by contradiction. Suppose there exists an algorithm \mathcal{A} that for any reward function r and any preference dataset \mathcal{D}_r with $|\mathcal{D}_r| \geq n_{\bar{r}}$, it outputs $R_{\bar{r}}$ with a probability of at least $1 - \epsilon$. We follow Friedgut, Kalai, and Naor [32] and define a social choice function as a function that yields an asymmetric relation on the alternatives given a profile. A social choice is *rational* if it is an order relation on the alternatives, and is *neutral* if it is invariant under permutations of alternatives.

Let $\mathcal{Y}_3 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3\}$ be an arbitrary subset of \mathcal{Y} with size 3. Next, we construct a neutral social choice function f acting on \mathcal{Y}_3 that is independent of irrelevant alternatives (IIA). Here is how we

design f : Let P_r be a profile of size $n \geq n_{\bar{r}}$ at the input, where every $R \in P_r$ is an i.i.d. draw from a Plackett–Luce (PL) ranking model with $\exp(r)$ as the weight of the alternatives. Note that the marginal distribution induced by PL on any two alternatives follows BT. Create three pairwise preference datasets $\mathcal{D}_{r,12}$, $\mathcal{D}_{r,23}$, and $\mathcal{D}_{r,13}$ where $\mathcal{D}_{r,ij} = \{\mathbf{y}_i R \mathbf{y}_j \mid R \in P_r\}$. The social function f applies \mathcal{A} to every dataset to obtain a relation over \mathcal{Y}_3 . By construction, f is neutral and IIA.

By assumption, for every internal dataset $\mathcal{D}_{r,ij}$ we have $\Pr(\mathcal{A}(\mathcal{D}_{r,ij}) = R_{r,ij}) \geq 1 - \epsilon$, where $R_{r,ij}$ is the projection of $R_{\bar{r}}$ to only two alternatives \mathbf{y}_i and \mathbf{y}_j . Using union bound, we have

$$\Pr(f(P_r) = R_{\bar{r}}) \geq 1 - 3\epsilon.$$

Similar to strategy 1, we can define a new reward function r_γ with $\bar{r}_\gamma = \bar{r}$ such that the above holds for every r_γ with $\gamma > 1$. Then, by increasing γ , the profile P_{r_γ} approaches a uniformly distributed profile P_{unif} . In this case, since $R_{\bar{r}}$ is an order relation, we have $\Pr(f \text{ is rational}) \geq 1 - 3\epsilon$. Then Theorem 1.3 of Friedgut, Kalai, and Naor [32] implies that for some global constant K ,

$$\Pr(f \text{ is dictatorship}) \geq 1 - 3K\epsilon.$$

On the other hand, we know that the order that $R_{\bar{r}}$ induces for different r is not a dictatorship, so, we have

$$\Pr(f \text{ is dictatorship}) < 3\epsilon.$$

Putting these together, we obtain a lower bound on ϵ :

$$\epsilon \geq \frac{1}{3(K+1)} > 0.$$

The rest of the proof is similar for the second and third strategies. We can use a boosting argument to show that from any weak pairwise learner, we can obtain an arbitrarily strong weak learner corresponding at the cost of collecting a larger dataset. We show this when for the weak learner $\epsilon < \frac{1}{2}$ but a weaker condition of choosing $R_{\bar{r}}$ better than chance level is sufficient for our argument. Consider a pairwise preference dataset \mathcal{D}_r of size $m n_{\bar{r}}$. Partition \mathcal{D}_r into m equal-size datasets. Let R_i be the output of \mathcal{A} on the i^{th} dataset. Since samples in \mathcal{D}_r are independently generated, R_i s err independently. Construct a meta-algorithm \mathcal{A}_{maj} that outputs the majority winner of R_i s. A standard Hoeffding bound implies

$$\Pr(\mathcal{A}_{\text{maj}}(\mathcal{D}_r) \neq R_{\bar{r}}) \leq \exp(-2(1/2 - \epsilon)^2 m).$$

A simple calculation then shows that for any arbitrarily small $\epsilon' > 0$, by choosing $m = O(\log(\frac{1}{\epsilon'}) (\frac{1}{2} - \epsilon)^{-2})$ the majority-winner algorithm agrees with $R_{\bar{r}}$ with probability of at least $1 - \epsilon'$. This contradicts the lower bound we established earlier and completes the proof.

Proof Strategy 3. The proof is by contradiction and is inspired by Procaccia, Schiffer, and Zhang [29]. This proof requires at least four different user types. Suppose there are two equally represented user types $\mathcal{U} = \{A, B\}$ who follow BT. For some arbitrary response $\mathbf{y}_0 \in \mathcal{Y}$ and $0 < \tau < \frac{1}{3}$, consider the following reward function:

$$r_\tau(\mathbf{y}; u) = \begin{cases} 0, & \mathbf{y} \neq \mathbf{y}_0, \\ \sigma^{-1}(\frac{2}{3} + \tau) = \log \frac{\frac{2}{3} - \tau}{\frac{1}{3} + \tau}, & \mathbf{y} = \mathbf{y}_0, u = A, \\ \sigma^{-1}(\frac{2}{3} - \tau) = \log \frac{\frac{2}{3} + \tau}{\frac{1}{3} - \tau}, & \mathbf{y} = \mathbf{y}_0, u = B. \end{cases} \quad (20)$$

One can see $\Pr(\mathbf{y}_0 \succ \mathbf{y} \mid r_\tau) = \frac{2}{3}$ for every $\mathbf{y} \neq \mathbf{y}_0$. Therefore, r_τ induces the same pairwise preference distribution for every τ . On the other hand, $\bar{r}_\tau(\mathbf{y}_0) := \mathbb{E}_u[r_\tau(\mathbf{y}_0; u)] = \log \frac{\frac{4}{3} - \tau^2}{\frac{1}{3} - \tau^2} > 0$ is an increasing function of τ .

Consider two arbitrary τ_1 and τ_2 such that $0 < \tau_1 < \tau_2 < \frac{1}{3}$. Sample a pairwise preference dataset as follows. Draw a random u and a random permutation ρ over \mathcal{Y} . If ρ is not identity, ask an annotator of type u with reward $r_{\tau_1}(\cdot; u)$ to label this sample and permute the ranking with ρ . If ρ is identity, ask an annotator of type u with reward $r_{\tau_2}(\cdot; u)$ to label. This sampling is equivalent to sampling from $2 * |\mathcal{Y}|!$ different user types. By symmetry, this pairwise preference dataset is distributionally equivalent to a dataset $\mathcal{D}_{\text{unif}}$ with indifferent preferences. However, our construction implies that \mathbf{y}_0 has the highest expected reward and other alternatives have similar rewards:

$$\bar{r}(\mathbf{y}) = \frac{1}{|\mathcal{Y}|} \begin{cases} \bar{r}_{\tau_1}(\mathbf{y}), & \mathbf{y} \neq \mathbf{y}_0, \\ \bar{r}_{\tau_2}(\mathbf{y}), & \mathbf{y} = \mathbf{y}_0. \end{cases}$$

Denote the ranking based on \bar{r} above by R_0 .

Similar to the second strategy, suppose there exists an algorithm \mathcal{A} that for any reward function r and any preference dataset \mathcal{D}_r with $|\mathcal{D}_r| \geq n_{\bar{r}}$, it outputs $R_{\bar{r}}$ with a probability of at least $1 - \epsilon$. Collect a preference dataset as explained above with at least $n_{\bar{r}}$ samples. Then, by assumption,

$$\Pr(\mathcal{A}(\mathcal{D}_{\text{unif}}) = R_0) \geq 1 - \epsilon.$$

Note that our choice of \mathbf{y}_0 could be any of the alternatives. Therefore, the above should be true when any of the alternatives has the highest expected reward. This implies a lower bound on ϵ :

$$\epsilon \geq 1 - \frac{1}{|\mathcal{Y}|} > 0.$$

The rest of the proof is similar to the second strategy. \square

Proposition 5.2. *There is no M-estimator that can estimate $V(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2) := \text{Var}_u[\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u)]$ consistently without annotator information.*

Proof of Proposition 5.2. Consider a dataset \mathcal{D} of context, candidate pairs, and preference represented as $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, o)$, where $o = \mathbb{1}\{\mathbf{y}_2 \succ \mathbf{y}_1\}$, and $\mathbf{y}_1, \mathbf{y}_2$ are independently drawn. Then consider an M-estimator

$$\arg \min_V \sum_{(\mathbf{x}, \mathbf{y}_l, \mathbf{y}_w) \in \mathcal{D}} \rho(\mathbf{x}, \mathbf{y}_l, \mathbf{y}_w; V) = \sum_{(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, o) \in \mathcal{D}} o \cdot \rho(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; V) + (1 - o) \cdot \rho(\mathbf{x}, \mathbf{y}_2, \mathbf{y}_1; V).$$

Under preference model of Eq. (2), for a reward r , we have $\mathbb{E}[o \mid \mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u] = \sigma(\Delta r(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u))$. In the limit of a large dataset, the M-estimator solves

$$\begin{aligned} & \arg \min_V \mathbb{E}_{\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, o, u} \left[o \cdot \rho(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; V) + (1 - o) \cdot \rho(\mathbf{x}, \mathbf{y}_2, \mathbf{y}_1; V) \right] \\ &= \mathbb{E}_{\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, u} \left[\sigma(\Delta r(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u)) \cdot \rho(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; V) + \sigma(\Delta r(\mathbf{x}, \mathbf{y}_2, \mathbf{y}_1; u)) \cdot \rho(\mathbf{x}, \mathbf{y}_2, \mathbf{y}_1; V) \right] \quad (V \text{ has no } o) \\ &= \mathbb{E}_{\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2} \left[\mathbb{E}_u[\sigma(\Delta r(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u))] \cdot \rho(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; V) + \mathbb{E}_u[\sigma(\Delta r(\mathbf{x}, \mathbf{y}_2, \mathbf{y}_1; u))] \cdot \rho(\mathbf{x}, \mathbf{y}_2, \mathbf{y}_1; V) \right] \quad (V \text{ has no } u) \\ &= 2 \mathbb{E}_{\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2} \left[\mathbb{E}_u[\sigma(\Delta r(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u))] \cdot \rho(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; V) \right]. \quad (\mathbf{y}_1, \mathbf{y}_2 \text{ are i.i.d}) \end{aligned}$$

Here, we used the fact that V does not depend on u . We also relied on the assumption that \mathbf{y}_1 and \mathbf{y}_2 are identically and independently distributed. The above equation suggests that regardless of how ρ is designed, $V(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$ can only depend on u 's distribution through $\mathbb{E}_u[\sigma(\Delta r(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u))]$. Therefore, no consistent M-estimator can generally estimate $\text{Var}_u[\Delta r(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u)]$ even with the availability of infinite preference data. \square

Lemma 5.3. Using J_1 and J_2 as shorthands for $J(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$ and $J(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{x}, \mathbf{y}_2, \mathbf{y}_1)$, we can use the following to estimate the variance term:

$$V(\mathbf{y}_1, \mathbf{y}_2) = \frac{J_1 - (J_1 + J_2)^2}{\sigma'(\Delta\bar{r}^*(\mathbf{y}_1, \mathbf{y}_2))^2}. \quad (16)$$

Proof of Lemma 5.3. First of all, J can give us the likelihood itself:

$$\begin{aligned} J(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{x}, \mathbf{y}_1, \mathbf{y}_2) + J(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{x}, \mathbf{y}_2, \mathbf{y}_1) &= \mathbb{E}_u \left[\sigma(\Delta r^*(\mathbf{y}_1, \mathbf{y}_2; u))^2 + \sigma(\Delta r^*(\mathbf{y}_1, \mathbf{y}_2; u)) \cdot \sigma(\Delta r^*(\mathbf{y}_2, \mathbf{y}_1; u)) \right] \\ &= \mathbb{E}_u \left[\sigma(\Delta r^*(\mathbf{y}_1, \mathbf{y}_2; u)) \right]. \end{aligned}$$

Here, we used the property $\sigma(\Delta r^*(\mathbf{y}_2, \mathbf{y}_1; u)) = 1 - \sigma(\Delta r^*(\mathbf{y}_1, \mathbf{y}_2; u))$. We also dropped \mathbf{x} from the notation for simplicity. Since J can give us both the first and second moments, we can use it to find $\text{Var}_u[\sigma(\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u))]$ as follows:

$$\begin{aligned} \text{Var}_u[\sigma(\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u))] &= \mathbb{E}_u[\sigma(\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u))^2] - \mathbb{E}_u[\sigma(\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u))]^2 \\ &= J(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{x}, \mathbf{y}_1, \mathbf{y}_2) - (J(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{x}, \mathbf{y}_1, \mathbf{y}_2) + J(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{x}, \mathbf{y}_2, \mathbf{y}_1))^2. \end{aligned}$$

In the last piece of the proof, we connect $\text{Var}_u[\sigma(\Delta r^*)]$ with $\text{Var}_u[\Delta r^*]$. The Taylor expansion of $\sigma(\Delta r^*)$ around $\Delta\bar{r}^* := \mathbb{E}_u[\Delta r^*]$ gives

$$\begin{aligned} \text{Var}_u[\sigma(\Delta r^*)] &= \text{Var}_u \left[\sigma(\Delta\bar{r}^*) + \sigma'(\Delta\bar{r}^*) \cdot (\Delta r^* - \Delta\bar{r}^*) + O((\Delta r^* - \Delta\bar{r}^*)^2) \right] \\ &= \sigma'(\Delta\bar{r}^*)^2 \cdot \text{Var}_u[\Delta r^*] + O\left(\mathbb{E}_u[(\Delta r^* - \Delta\bar{r}^*)^3]\right). \end{aligned}$$

We can neglect the third-order term in calculations as first-order correction uses up to $O\left(\mathbb{E}_u[(\Delta r^* - \Delta\bar{r}^*)^2]\right)$ in its approximation. \square

Proposition 6.1. Defining l in Eq. (17) as follows results in a consistent estimation of the optimal policy when preferences follow the BT model:

$$l(\mathbf{y}_1, \mathbf{y}_2, \boldsymbol{\sigma}; \pi) = \begin{cases} -\log \sigma(|\mathcal{U}| \cdot h(\mathbf{y}_1, \mathbf{y}_2; \pi)), & \boldsymbol{\sigma} = \vec{\mathbf{1}}, \\ -\log \sigma(|\mathcal{U}| \cdot h(\mathbf{y}_2, \mathbf{y}_1; \pi)), & \boldsymbol{\sigma} = \vec{\mathbf{0}}, \\ 0 & o.w. \end{cases}$$

Here, h is the difference of π 's induced rewards (Eq. (15)).

Proof of Proposition 6.1. The proof follows similar steps as the derivation of DPO. First of all, conditioned on agreement, the likelihood of observing $\mathbf{y}_2 \succ \mathbf{y}_1$ under the BT model is

$$\begin{aligned} \Pr(\mathbf{y}_2 \succ \mathbf{y}_1 \mid r^*, \text{agreement}) &= \frac{\Pi_u \sigma(\Delta r^*(\mathbf{y}_1, \mathbf{y}_2; u))}{\Pi_u \sigma(\Delta r^*(\mathbf{y}_1, \mathbf{y}_2; u)) + \Pi_u \sigma(\Delta r^*(\mathbf{y}_2, \mathbf{y}_1; u))} \\ &= \frac{\exp\left(\sum_u r^*(\mathbf{y}_2; u)\right)}{\exp\left(\sum_u r^*(\mathbf{y}_1; u)\right) + \exp\left(\sum_u r^*(\mathbf{y}_2; u)\right)} \\ &= \sigma\left(|\mathcal{U}| \cdot \mathbb{E}_u[\Delta r^*(\mathbf{y}_1, \mathbf{y}_2)]\right). \end{aligned}$$

On the other hand, Eq. (11) allows us to write $\mathbb{E}_u[\Delta r^*(\mathbf{y}_1, \mathbf{y}_2)]$ with difference π 's induced rewards, i.e., h :

$$\Pr(\mathbf{y}_2 \succ \mathbf{y}_1 \mid \pi^*, \text{agreement}) = \sigma(h(\mathbf{y}_1, \mathbf{y}_2; \pi^*)).$$

We can define the likelihood in this way for every policy π . Then, the proposed loss function is equivalent to maximizing log-likelihood, which under mild conditions is a consistent estimator for π^* . \square

Theorem 6.2. Suppose l in Eq. (17) only depends on $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$ through π and π_{ref} . If there are more than three types of user and the preferences follow BT, any loss \mathcal{L} that allows a consistent estimation of the optimal policy discards samples with disagreement, i.e., those with $\mathbf{o} \notin \{\mathbf{0}, \mathbf{1}\}$.

Proof of Theorem 6.2. The proof involves three steps: First, the next lemma shows that any loss function l in Eq. (17) with the desired consistency property can only depend on π through the ratio $\frac{\pi(\mathbf{y}_2|\mathbf{x})}{\pi(\mathbf{y}_1|\mathbf{x})}$.

Lemma C.1. Suppose l in Eq. (17) only depends on $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$ through π and π_{ref} . Then, for any l that gives a consistent estimation of the optimal policy in Eq. (10), there exists an equivalent loss \tilde{l} such that

$$l(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{o}; \pi) = \tilde{l}(\mathbf{o}; h(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; \pi)),$$

where h is defined in Eq. (15).

See [proof](#) on page 27.

In the second step, we further limit the search space of \tilde{l} (as introduced by Lemma C.1) to those that meet certain first- and second-order conditions:

Lemma C.2. Any loss \tilde{l} as in Lemma C.1 that leads to a consistent estimation of the optimal policy meets

$$\begin{aligned} \sum_{\mathbf{o} \in \{0,1\}^{\mathcal{U}}} \frac{\partial \tilde{l}}{\partial \theta}(\mathbf{o}; \theta^*(\mathbf{z})) \cdot \chi_{\mathbf{o}}(\mathbf{z}) &= 0, \\ \sum_{\mathbf{o} \in \{0,1\}^{\mathcal{U}}} \frac{\partial^2 \tilde{l}}{\partial \theta^2}(\mathbf{o}; \theta^*(\mathbf{z})) \cdot \chi_{\mathbf{o}}(\mathbf{z}) &\geq 0, \end{aligned}$$

for every $\mathbf{z} \in [0, 1]^{\mathcal{U}}$. Here, we define

$$\chi_{\mathbf{o}}(\mathbf{z}) := \prod_{u \in \mathcal{U}} z_u^{o_u} (1 - z_u)^{1 - o_u}.$$

and

$$\theta^*(\mathbf{z}) := \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sigma^{-1}(z_u).$$

See [proof](#) on page 28.

Finally, we show that when preferences follow the BT model and there are more than three user types, all $\tilde{l}(\mathbf{o}; \theta)$ terms corresponding to $\mathbf{o} \notin \{\mathbf{0}, \mathbf{1}\}$ do not depend on θ . Therefore, these terms do not depend on π and can be removed from the loss function, thereby completing the proof.

Lemma C.3. If $|\mathcal{U}| > 3$, for any loss \tilde{l} that meets the first-order condition of Lemma C.2, we have $\frac{\partial \tilde{l}}{\partial \theta}(\mathbf{o}; \theta) = 0$ for every $\mathbf{o} \notin \{\mathbf{0}, \mathbf{1}\}$.

See [proof](#) on page 29. □

Proposition B.1. There exists a mixture of BTs that a single BT cannot represent.

Proof of Proposition B.1. Suppose the pairwise comparison distribution over a set of alternatives $(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \dots)$ satisfies the Bradley-Terry (BT) model; i.e. $\Pr(\mathbf{y}_i \succ \mathbf{y}_j) = \sigma(r^*(\mathbf{y}_i) - r^*(\mathbf{y}_j))$. Then:

$$\begin{aligned} \Pr(\mathbf{y}_1 \succ \mathbf{y}_2) \Pr(\mathbf{y}_2 \succ \mathbf{y}_3) \Pr(\mathbf{y}_3 \succ \mathbf{y}_1) &= \frac{\prod_{i=1}^3 \exp(r^*(\mathbf{y}_i))}{\prod_{i=1}^3 (\exp(r^*(\mathbf{y}_i)) + \exp(r^*(\mathbf{y}_{(i+1) \bmod 3+1})))} \\ &= \Pr(\mathbf{y}_1 \succ \mathbf{y}_3) \Pr(\mathbf{y}_3 \succ \mathbf{y}_2) \Pr(\mathbf{y}_2 \succ \mathbf{y}_1). \end{aligned}$$

Now, consider two BT models corresponding to u_1 and u_2 , with a uniform mixture over them. For the mixture:

$$\Pr(\mathbf{y}_i \succ \mathbf{y}_j) = \frac{\Pr(\mathbf{y}_i \succ \mathbf{y}_j \mid u_1) + \Pr(\mathbf{y}_i \succ \mathbf{y}_j \mid u_2)}{2}.$$

The probability of cyclic preferences in one direction is given by

$$\Pr(\mathbf{y}_1 \succ \mathbf{y}_2) \Pr(\mathbf{y}_2 \succ \mathbf{y}_3) \Pr(\mathbf{y}_3 \succ \mathbf{y}_1) = \frac{\sum_{s \in \{1,2\}^3} \prod_{i=1}^3 \Pr(\mathbf{y}_i \succ \mathbf{y}_{(i+1) \bmod 3+1} \mid u_{s_i})}{8},$$

which is not necessarily equal to the probability of the cyclic preferences in the reverse direction:

$$\Pr(\mathbf{y}_1 \succ \mathbf{y}_3) \Pr(\mathbf{y}_3 \succ \mathbf{y}_2) \Pr(\mathbf{y}_2 \succ \mathbf{y}_1) = \frac{\sum_{s \in \{1,2\}^3} \prod_{i=1}^3 \Pr(\mathbf{y}_{(i+1) \bmod 3+1} \succ \mathbf{y}_i \mid u_{s_i})}{8}.$$

To verify this, consider specific examples such as $\Pr(\mathbf{y}_i \succ \mathbf{y}_j \mid u_k) = \frac{\exp(r_k^i)}{\exp(r_k^i) + \exp(r_k^j)}$ with $r_1 = (1, 2, 3)$ and $r_2 = (1, 2, 4)$. More generally, the BT assumption implies that, for a fixed reward r^* , the likelihood of a set of pairwise comparisons $\{(\mathbf{y}_{p,1} > \mathbf{y}_{p,2})\}_{p \in [P]}$ is proportional to $\prod_i \exp(r^*(\mathbf{y}_i))^{\lfloor \sum_{p \in [P]} \mathbf{1}_{\mathbf{y}_{p,1}=i} \rfloor}$ and depends only on the number of times each option is preferred in the comparisons. However, as demonstrated above, this property does not hold for a mixture of BT models. \square

Proposition B.3. *Defining l in Eq. (17) as follows results in a consistent estimation of the optimal policy when preferences follow the BT model:*

$$l(\mathbf{y}_1, \mathbf{y}_2, \mathbf{o}; \pi) = \begin{cases} -\sigma(h(\mathbf{y}_1, \mathbf{y}_2; \pi)) - I(\sigma(h(\mathbf{y}_1, \mathbf{y}_2; \pi))), & \mathbf{o} = \mathbf{1}, \\ -\sigma(h(\mathbf{y}_2, \mathbf{y}_1; \pi)) - I(\sigma(h(\mathbf{y}_2, \mathbf{y}_1; \pi))), & \mathbf{o} = \mathbf{0}, \\ 0 & o.w. \end{cases}$$

Here, we define $I(\theta) := \int_1^\theta (\frac{1}{\theta'} - 1)^{|\mathcal{U}|} d\theta'$, and h is the difference of π 's induced rewards (Eq. (15)).

Proof of Proposition B.3. Recall $\Pr(o_u = 1 \mid \mathbf{x}, \mathbf{y}_1, \mathbf{y}_2) = \sigma(\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u))$. We use $z_u(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$ as a shorthand for this quantity and will drop the dependence on $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$ whenever it is clear from the context. We also use s as a shorthand for $\sigma(h)$. In the limit of a very large dataset, the proposed loss approaches

$$\mathcal{L}(s) = -\mathbb{E}_{\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2} \left[\left(\prod_{u \in \mathcal{U}} z_u \right) (s + I(s)) + \left(\prod_{u \in \mathcal{U}} (1 - z_u) \right) (1 - s + I(1 - s)) \right].$$

Note that we wrote \mathcal{L} as a function of s instead of π since s is the only place that π appears. We first show that $\mathcal{L}(s)$ has a unique global minimizer. To show an s is a global minimizer of \mathcal{L} , it suffices to show that s minimizes the term inside expectation for every $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$. Such a minimizer meets the first-order condition:

$$\left(\prod_{u \in \mathcal{U}} z_u \right) \left(1 + \left(\frac{1}{s} - 1 \right)^{|\mathcal{U}|} \right) + \left(\prod_{u \in \mathcal{U}} (1 - z_u) \right) \left(-1 - \left(\frac{1}{1-s} - 1 \right)^{|\mathcal{U}|} \right) = 0.$$

Here, we used $\frac{dI}{ds} = \left(\frac{1}{s} - 1 \right)^{|\mathcal{U}|}$. Define $w := \left(\frac{1-s}{s} \right)^{|\mathcal{U}|}$. Then, the above condition reduces to a quadratic equation in terms of w :

$$1 + w - \left(\prod_{u \in \mathcal{U}} \left(\frac{1}{z_u} - 1 \right) \right) (1 + w^{-1}) = (1 + w^{-1}) \left[w - \prod_{u \in \mathcal{U}} \left(\frac{1}{z_u} - 1 \right) \right] = 0.$$

Solving for w , we obtain

$$s^* = \frac{1}{1 + \left(\prod_{u \in \mathcal{U}} \left(\frac{1}{z_u} - 1 \right) \right)^{\frac{1}{|\mathcal{U}|}}}.$$

For the BT model, a direct calculation then shows

$$s^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2) = \sigma\left(\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \Delta r(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u)\right). \quad (21)$$

In fact, s^* is the only global minimizer of $\mathcal{L}(s)$. This is because $\mathcal{L}(s)$ is convex in s :

$$\frac{d^2 \mathcal{L}}{ds^2} = \mathbb{E}_{\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2} \left[\left(\prod_{u \in \mathcal{U}} z_u \right) \cdot \frac{|\mathcal{U}|}{s^2} \left(\frac{1}{s} - 1 \right)^{|\mathcal{U}|-1} + \left(\prod_{u \in \mathcal{U}} (1 - z_u) \right) \cdot \frac{|\mathcal{U}|}{(1-s)^2} \left(\frac{1}{1-s} - 1 \right)^{|\mathcal{U}|-1} \right] \geq 0.$$

Finally, one can verify that the policy that results in s^* (Eq. (21)) is the optimal policy π^* . This completes the proof that the proposed loss is a consistent loss for π^* . \square

Lemma C.1. *Suppose l in Eq. (17) only depends on $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$ through π and π_{ref} . Then, for any l that gives a consistent estimation of the optimal policy in Eq. (10), there exists an equivalent loss \tilde{l} such that*

$$l(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{o}; \pi) = \tilde{l}(\mathbf{o}; h(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; \pi)),$$

where h is defined in Eq. (15).

Proof of Lemma C.1. Since l only depends on $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$ through π and π_{ref} , we overload the notation and use $l(\mathbf{o}; \pi(\mathbf{y}_1 | \mathbf{x}), \pi(\mathbf{y}_2 | \mathbf{x}))$ to denote the loss from $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{o})$. In the limit of many data points, $\mathcal{L}(\mathcal{D}; \pi)$ converges to

$$\mathcal{L}(\pi) = \mathbb{E}_{\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{o}} [l(\mathbf{o}; \pi(\mathbf{y}_1 | \mathbf{x}), \pi(\mathbf{y}_2 | \mathbf{x}))].$$

Using $\Pr(o_u = 1 | \mathbf{x}, \mathbf{y}_1, \mathbf{y}_2) = \sigma(\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u))$, the tower rule implies

$$\begin{aligned} \mathcal{L}(\pi) &= \mathbb{E}_{\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2} \left[\mathbb{E}_{\mathbf{o}} [l(\mathbf{o}; \pi(\mathbf{y}_1 | \mathbf{x}), \pi(\mathbf{y}_2 | \mathbf{x})) | \mathbf{x}, \mathbf{y}_1, \mathbf{y}_2] \right] \\ &= \mathbb{E}_{\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2} \left[\sum_{\mathbf{o} \in \{0,1\}^{\mathcal{U}}} l(\mathbf{o}; \pi(\mathbf{y}_1 | \mathbf{x}), \pi(\mathbf{y}_2 | \mathbf{x})) \cdot \prod_{u \in \mathcal{U}} \sigma(\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u))^{o_u} (1 - \sigma(\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u)))^{1-o_u} \right]. \end{aligned}$$

For notational simplicity, let's define

$$\begin{aligned} z_u(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2) &:= \sigma(\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u)), \\ \chi_{\mathbf{o}}(\mathbf{z}) &:= \prod_{u \in \mathcal{U}} z_u^{o_u} (1 - z_u)^{1-o_u}. \end{aligned}$$

Note that $\chi_{\mathbf{o}}$ implicitly depends on $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$ through \mathbf{z} , which we drop from the notation when it is clear from the context. Using this notation, we can rewrite the $\mathcal{L}(\pi)$'s expansion as follows:

$$\mathcal{L}(\pi) = \mathbb{E}_{\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2} \left[\sum_{\mathbf{o} \in \{0,1\}^{\mathcal{U}}} l(\mathbf{o}; \pi(\mathbf{y}_1 | \mathbf{x}), \pi(\mathbf{y}_2 | \mathbf{x})) \cdot \chi_{\mathbf{o}}(\mathbf{z}(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)) \right]. \quad (22)$$

Overloading notation, we can always equivalently represent $(\pi(\mathbf{y}_1 | \mathbf{x}), \pi(\mathbf{y}_2 | \mathbf{x}))$ as $(\pi(\mathbf{y}_1, \mathbf{y}_2 | \mathbf{x}), \pi(\mathbf{y}_2 | \mathbf{y}_1, \mathbf{y}_2, \mathbf{x}))$, that is, with the probability that either of the two responses is chosen and the probability that the second one is preferred. Therefore, there exists a loss l' such that

$$l(\mathbf{o}; \pi(\mathbf{y}_1 | \mathbf{x}), \pi(\mathbf{y}_2 | \mathbf{x})) = l'(\mathbf{o}; \pi(\{\mathbf{y}_1, \mathbf{y}_2\} | \mathbf{x}), \pi(\mathbf{y}_2 | \{\mathbf{y}_1, \mathbf{y}_2\}, \mathbf{x})).$$

If π is optimal, $\pi(\mathbf{y}_2 | \mathbf{y}_1, \mathbf{y}_2, \mathbf{x})$ should also be optimal. Since $\pi(\mathbf{y}_2 | \mathbf{y}_1, \mathbf{y}_2, \mathbf{x})$ appears in only one term of the expectation in Eq. (22), we can conclude that

$$\arg \min_{\theta'} \sum_{\mathbf{o} \in \{0,1\}^{\mathcal{U}}} l'(\mathbf{o}; \pi^*(\{\mathbf{y}_1, \mathbf{y}_2\} | \mathbf{x}), \theta') \cdot \chi_{\mathbf{o}}(\mathbf{z}(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2))$$

is the optimal $\pi(\mathbf{y}_2 \mid \{\mathbf{y}_1, \mathbf{y}_2\}, \mathbf{x})$ for every optimal $\pi(\{\mathbf{y}_1, \mathbf{y}_2\} \mid \mathbf{x})$. On the other hand, a property of the optimal policy π^* is that

$$\pi^*(\mathbf{y}_2 \mid \{\mathbf{y}_1, \mathbf{y}_2\}, \mathbf{x}) = \frac{\pi^*(\mathbf{y}_2 \mid \mathbf{x})}{\pi^*(\mathbf{y}_1 \mid \mathbf{x})} = \frac{\pi_{\text{ref}}(\mathbf{y}_2 \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_1 \mid \mathbf{x})} \cdot \exp\left(\frac{1}{\beta} \mathbb{E}_u[\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u)]\right).$$

Therefore, for every optimal policy $\pi^*(\{\mathbf{y}_1, \mathbf{y}_2\} \mid \mathbf{x})$, we have

$$\begin{aligned} & \frac{\pi_{\text{ref}}(\mathbf{y}_2 \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_1 \mid \mathbf{x})} \cdot \exp\left(\frac{1}{\beta} \mathbb{E}_u[\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u)]\right) = \\ & \arg \min_{\theta'} \sum_{\mathbf{o} \in \{0,1\}^{\mathcal{U}}} l'(\mathbf{o}; \pi^*(\{\mathbf{y}_1, \mathbf{y}_2\} \mid \mathbf{x}), \theta') \cdot \chi_{\mathbf{o}}(\mathbf{z}(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)). \end{aligned} \quad (23)$$

Recall from the optimal policy (Eq. (10)) that we can modify the reward function for responses other than $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$ while keeping $\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u)$ constant. This allows arbitrary changes to $\pi^*(\{\mathbf{y}_1, \mathbf{y}_2\} \mid \mathbf{x})$ without altering the rest of Eq. (23). So, we can argue that l' does not depend on $\pi(\{\mathbf{y}_1, \mathbf{y}_2\} \mid \mathbf{x})$ and we drop it from l' notation. Define a new loss based on l' :

$$\tilde{l}(\mathbf{o}; \theta) := l'\left(\mathbf{o}; \frac{\pi_{\text{ref}}(\mathbf{y}_2 \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_1 \mid \mathbf{x})} \cdot \exp\left(\frac{1}{\beta} \theta\right)\right).$$

Note that \tilde{l} implicitly depends on $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$ through π_{ref} which we dropped from notation. Using \tilde{l} , we can write the original loss l as

$$\begin{aligned} l(\mathbf{o}; \pi(\mathbf{y}_1 \mid \mathbf{x}), \pi(\mathbf{y}_2 \mid \mathbf{x})) &= l'(\mathbf{o}; \pi(\mathbf{y}_2 \mid \{\mathbf{y}_1, \mathbf{y}_2\}, \mathbf{x})) \\ &= l'\left(\mathbf{o}; \frac{\pi(\mathbf{y}_2 \mid \mathbf{x})}{\pi(\mathbf{y}_1 \mid \mathbf{x})}\right) \\ &= \tilde{l}\left(\mathbf{o}; \beta \log \frac{\pi(\mathbf{y}_2 \mid \mathbf{x})}{\pi(\mathbf{y}_1 \mid \mathbf{x})} - \beta \log \frac{\pi_{\text{ref}}(\mathbf{y}_2 \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_1 \mid \mathbf{x})}\right) \\ &= \tilde{l}(\mathbf{o}; h(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; \pi)). \end{aligned}$$

This completes the proof. \square

Lemma C.2. *Any loss \tilde{l} as in Lemma C.1 that leads to a consistent estimation of the optimal policy meets*

$$\begin{aligned} & \sum_{\mathbf{o} \in \{0,1\}^{\mathcal{U}}} \frac{\partial \tilde{l}}{\partial \theta}(\mathbf{o}; \theta^*(\mathbf{z})) \cdot \chi_{\mathbf{o}}(\mathbf{z}) = 0, \\ & \sum_{\mathbf{o} \in \{0,1\}^{\mathcal{U}}} \frac{\partial^2 \tilde{l}}{\partial \theta^2}(\mathbf{o}; \theta^*(\mathbf{z})) \cdot \chi_{\mathbf{o}}(\mathbf{z}) \geq 0, \end{aligned}$$

for every $\mathbf{z} \in [0, 1]^{\mathcal{U}}$. Here, we define

$$\chi_{\mathbf{o}}(\mathbf{z}) := \prod_{u \in \mathcal{U}} z_u^{o_u} (1 - z_u)^{1 - o_u}.$$

and

$$\theta^*(\mathbf{z}) := \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sigma^{-1}(z_u).$$

Proof of Lemma C.2. We will refer to the proof of Lemma C.1 in this proof. Using \tilde{l} in place of l' in Eq. (23), since $\exp(\cdot)$ is monotone increasing, we have

$$\mathbb{E}_u[\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u)] = \arg \min_{\theta} \sum_{\mathbf{o} \in \{0,1\}^{\mathcal{U}}} \tilde{l}(\mathbf{o}; \theta) \cdot \chi_{\mathbf{o}}(\mathbf{z}).$$

On the other hand, using the fact that user types in \mathcal{U} are equiprobable, we can write

$$\mathbb{E}_u[\Delta r^*(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2; u)] = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sigma^{-1}(z_u).$$

Putting these together, it is necessary to have

$$\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sigma^{-1}(z_u) = \arg \min_{\theta} \sum_{\mathbf{o} \in \{0,1\}^{\mathcal{U}}} \tilde{l}(\mathbf{o}; \theta) \cdot \chi_{\mathbf{o}}(\mathbf{z})$$

for every $\mathbf{z} \in [0, 1]^{\mathcal{U}}$. The rest of the proof is straightforward. \square

Lemma C.3. *If $|\mathcal{U}| > 3$, for any loss \tilde{l} that meets the first-order condition of Lemma C.2, we have $\frac{\partial \tilde{l}}{\partial \theta}(\mathbf{o}; \theta) = 0$ for every $\mathbf{o} \notin \{\mathbf{0}, \mathbf{1}\}$.*

Proof of Lemma C.3. First of all, for the BT model, a direct calculation shows

$$\theta^*(\mathbf{z}) := \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sigma^{-1}(z_u) = \frac{1}{|\mathcal{U}|} \log \prod_{u \in \mathcal{U}} \left(\frac{z_u}{1 - z_u} \right) = \frac{1}{|\mathcal{U}|} \log \left(\frac{\chi_{\mathbf{1}}}{\chi_{\mathbf{0}}} \right).$$

Since $\theta^*(\mathbf{z})$ depends on \mathbf{z} only through $\frac{\chi_{\mathbf{0}}}{\chi_{\mathbf{1}}}$, we denote $\frac{\partial \tilde{l}}{\partial \theta}(\mathbf{o}; \theta^*)$ by $g(\mathbf{o}; \frac{\chi_{\mathbf{0}}}{\chi_{\mathbf{1}}})$. The proof has two steps: In the first step, we relate $g(\mathbf{o}; \frac{\chi_{\mathbf{0}}}{\chi_{\mathbf{1}}})$ to $g(\mathbf{o}^{\oplus u'}; \frac{\chi_{\mathbf{0}}}{\chi_{\mathbf{1}}})$, where we define

$$\mathbf{o}^{\oplus u'} := \begin{cases} o_u, & u \neq u', \\ 1 - o_u, & u = u'. \end{cases}$$

Using this connection, in the second step, we will show that $g(\mathbf{o}; \frac{\chi_{\mathbf{0}}}{\chi_{\mathbf{1}}}) = 0$ for any $\mathbf{o} \in \{\mathbf{0}, \mathbf{1}\}$ when $|\mathcal{U}| \geq 4$.

Step 1. Consider any $\mathbf{o} \notin \{\mathbf{0}, \mathbf{1}\}$. When $|\mathcal{U}| \geq 3$, there exists $u' \in \mathcal{U}$ such that $\mathbf{o}^{\oplus u'} \notin \{\mathbf{0}, \mathbf{1}\}$. For such \mathbf{o} and u' , we define two non-empty sets

$$\begin{aligned} \mathcal{S}^1 &:= \{u \mid u \in \mathcal{U}, u \neq u', o_u = 1\}, \\ \mathcal{S}^0 &:= \{u \mid u \in \mathcal{U}, u \neq u', o_u = 0\}. \end{aligned}$$

We set \mathbf{z} such that

$$\begin{aligned} \prod_{u \in \mathcal{S}^1} z_u &= \prod_{u \in \mathcal{S}^0} (1 - z_u), \\ z_u &\rightarrow 1^-, \forall u \in \mathcal{S}^1, \\ z_u &\rightarrow 0^+, \forall u \in \mathcal{S}^0. \end{aligned} \tag{24}$$

For this choice of \mathbf{z} , asymptotically, we have

$$\begin{aligned} \frac{\chi_{\mathbf{0}}}{\chi_{\mathbf{1}}} &= \frac{1 - z_{u'}}{z_{u'}}, \\ \chi_{\mathbf{o}} &= z_{u'}^{o_{u'}} (1 - z_{u'})^{1 - o_{u'}}, \\ \chi_{\mathbf{o}^{\oplus u'}} &= z_{u'}^{1 - o_{u'}} (1 - z_{u'})^{o_{u'}}, \\ \chi_{\mathbf{o}'} &= 0, \forall \mathbf{o}' \notin \{\mathbf{o}, \mathbf{o}^{\oplus u'}\}. \end{aligned}$$

Using the above, we can simplify the first-order condition in Lemma C.2 as

$$g\left(\boldsymbol{o}; \frac{1-z_{u'}}{z_{u'}}\right) \cdot z_{u'}^{o_{u'}} (1-z_{u'})^{1-o_{u'}} + g\left(\boldsymbol{o}^{\oplus u'}; \frac{1-z_{u'}}{z_{u'}}\right) \cdot z_{u'}^{1-o_{u'}} (1-z_{u'})^{o_{u'}} = 0.$$

This condition should be held for every $z_{u'} \in [0, 1]$. Therefore, we can conclude

$$g(\boldsymbol{o}^{\oplus u'}; \alpha) = -g(\boldsymbol{o}; \alpha) \cdot \alpha^{1-2o_{u'}}, \quad (25)$$

for every $\alpha \in \mathbb{R}$. This completes the first part of the proof.

Step 2. When $\boldsymbol{o} \notin \{\mathbf{0}, \mathbf{1}\}$ and $|\mathcal{U}| \geq 4$, there exist distinct user types u' and u'' such that none of $\boldsymbol{o}^{\oplus u'}$, $\boldsymbol{o}^{\oplus u''}$, and $\boldsymbol{o}^{\oplus(u', u'')}$ are in $\{\mathbf{0}, \mathbf{1}\}$. Here, we used $\boldsymbol{o}^{\oplus(u', u'')}$ as a shorthand for $(\boldsymbol{o}^{\oplus u'})^{\oplus u''}$. For such \boldsymbol{o} , u' , and u'' , we define two non-empty sets

$$\begin{aligned} \mathcal{S}^1 &:= \{u \mid u \in \mathcal{U} \setminus \{u', u''\}, o_u = 1\}, \\ \mathcal{S}^0 &:= \{u \mid u \in \mathcal{U} \setminus \{u', u''\}, o_u = 0\}. \end{aligned}$$

We set \boldsymbol{z} according to Eq. (24). Then, asymptotically,

$$\chi_{\boldsymbol{o}'} = 0, \quad \forall \boldsymbol{o}' \notin \{\boldsymbol{o}, \boldsymbol{o}^{\oplus u'}, \boldsymbol{o}^{\oplus u''}, \boldsymbol{o}^{\oplus(u', u'')}\}.$$

Using the above, we can simplify the first-order condition in Lemma C.2 as

$$g\left(\boldsymbol{o}; \frac{\chi_0}{\chi_1}\right) \cdot \chi_{\boldsymbol{o}} + g\left(\boldsymbol{o}^{\oplus u'}; \frac{\chi_0}{\chi_1}\right) \cdot \chi_{\boldsymbol{o}^{\oplus u'}} + g\left(\boldsymbol{o}^{\oplus u''}; \frac{\chi_0}{\chi_1}\right) \cdot \chi_{\boldsymbol{o}^{\oplus u''}} + g\left(\boldsymbol{o}^{\oplus(u', u'')}; \frac{\chi_0}{\chi_1}\right) \cdot \chi_{\boldsymbol{o}^{\oplus(u', u'')}} = 0. \quad (26)$$

Because of the symmetry of this equation, we can assume without loss of generality that $o_{u'} = 0$ and $o_{u''} = 1$. Therefore, asymptotically, we have

$$\begin{aligned} \frac{\chi_0}{\chi_1} &= \left(\frac{1-z_{u'}}{z_{u'}}\right) \left(\frac{1-z_{u''}}{z_{u''}}\right), \\ \chi_{\boldsymbol{o}} &= (1-z_{u'}) \cdot z_{u''}, \\ \chi_{\boldsymbol{o}^{\oplus u'}} &= z_{u'} \cdot z_{u''}, \\ \chi_{\boldsymbol{o}^{\oplus u''}} &= (1-z_{u'}) \cdot (1-z_{u''}), \\ \chi_{\boldsymbol{o}^{\oplus(u', u'')}} &= z_{u'} \cdot (1-z_{u''}). \end{aligned}$$

Eq. (25) also implies

$$\begin{aligned} g(\boldsymbol{o}^{\oplus u'}; \alpha) &= -g(\boldsymbol{o}; \alpha) \cdot \alpha, \\ g(\boldsymbol{o}^{\oplus u''}; \alpha) &= -g(\boldsymbol{o}; \alpha) \cdot \alpha^{-1}, \\ g(\boldsymbol{o}^{\oplus(u', u'')}; \alpha) &= g(\boldsymbol{o}; \alpha). \end{aligned}$$

Plugging these into Eq. (26) and simplifying equations, we obtain

$$g\left(\boldsymbol{o}; \left(\frac{1-z_{u'}}{z_{u'}}\right) \left(\frac{1-z_{u''}}{z_{u''}}\right)\right) \cdot (2z_{u'} - 1)(2z_{u''} - 1) = 0.$$

This equation should be held for every $z_{u'}$ and $z_{u''}$. By appropriately setting $z_{u'}$ and $z_{u''}$, we can conclude that $g(\boldsymbol{o}; \alpha)$ should be zero for every α . This completes this proof. \square

D PEW Surveys Experiments: Details and Additional Examples

In this section, we expand on our PEW surveys experiment where we used polling data on key political and social issues to show: (i) how NBC rankings can differ from those maximizing the average reward; (ii) How sensitive NBC is to the sampling distribution of the pairwise preference data.

Data. We use several Pew Research Center surveys, specifically the American Trends Panel surveys number 35, 52, 79, 83, 99, 109, 111, 112, 114, 119, 120, 121, 126, 127, 128, 129, 130, 131, and 132. The choices are a mix of recent surveys and those relevant to science, technology, data and AI. Each survey include questions asked to thousands of participants. We categorize participants by political party leanings to define types. When processing the questions, we discard responses that are empty, as well as discarding the option "Refused". We note that discarding the option "Refused" had no effect on the results as it is not frequently chosen.

Reward Estimation. Although we observe how often each group selects a particular option, we don't directly observe respondents' internal rewards. To estimate this, we apply the Luce-Shepherd model [24, 20]:

$$\Pr(\text{option } i \text{ is chosen from } \mathcal{S}) = \frac{\exp(r(i; u))}{\sum_{j \in \mathcal{S}} \exp(r(j; u))}, \quad (27)$$

where \mathcal{S} is the set of options, and $r(\cdot; u)$ is the reward for type u . This allows us to estimate each option's reward (up to a constant additive term) for each type. From these estimates and observed probabilities, we compute both the expected reward and the NBC metric, where in the latter we assume the uniform probability for alternatives unless specified otherwise.

Sensitivity Experiments. In Section 7.1, we estimate the sensitivity of NBC rankings to the sampling distribution of pairwise preference data by determining the minimum Total Variation (TV) distance required, if possible, to alter NBC rankings from that attained under the uniform distribution.

Recall that NBC is defined as:

$$\text{NBC}(\mathbf{y}; \mathcal{D}) := \mathbb{E}_{\mathbf{y}' \sim \mathcal{D}(\cdot)} [\Pr(\mathbf{y} \succ \mathbf{y}' \mid \mathbf{x}; r)].$$

To compute this, we first estimate the reward function r , then evaluate $\Pr(\mathbf{y} \succ \mathbf{y}'; r)$ for all $(\mathbf{y}, \mathbf{y}') \in \mathcal{Y} \times \mathcal{Y}$, where \mathcal{Y} is the set of alternatives. Next, consider the feasibility of swapping the ranking induced by the uniform distribution \mathcal{D}_U for alternatives \mathbf{y}_i and \mathbf{y}_j with a new distribution \mathcal{D}_a , assuming $\text{NBC}(\mathbf{y}_i; \mathcal{D}_U) > \text{NBC}(\mathbf{y}_j; \mathcal{D}_U)$. This is equivalent to solving the following linear program:

$$\begin{aligned} & \text{minimize} && \frac{1}{2} \mathbf{1}^\top \mathbf{s}, \\ & \text{subject to:} && \mathbf{q} > \epsilon \mathbf{1}, \\ & && \mathbf{s} \geq \frac{1}{N} \mathbf{1} - \mathbf{q}, \\ & && \mathbf{s} \geq \mathbf{q} - \frac{1}{N} \mathbf{1}, \\ & && \mathbf{P}_i \mathbf{q} < \mathbf{P}_j \mathbf{q} + \delta, \\ & && \mathbf{1}^\top \mathbf{q} = 1, \\ & && \mathbf{q} > \mathbf{0}. \end{aligned}$$

Here, $N = |\mathcal{Y}|$, and labeling the alternatives as $\mathbf{y}_1, \dots, \mathbf{y}_N$, we define $\mathbf{P}_{ij} = \Pr(\mathbf{y}_i \succ \mathbf{y}_j; r)$, $q_i = \mathcal{D}_a(\mathbf{y}_i)$, and \mathbf{P}_k as the k -th row of \mathbf{P} . The parameter $\epsilon > 0$ ensures support for all alternatives, and $\delta > 0$ controls the required magnitude of change in NBC beyond what is required for the swap. We set $\epsilon = \delta = 10^{-5}$.

To compute the minimum TV distance, we solve the program for all pairs (i, j) where $\text{NBC}(\mathbf{y}_i; \mathcal{D}_U) > \text{NBC}(\mathbf{y}_j; \mathcal{D}_U)$ and record the smallest objective value. In this analysis, we group respondents by political leaning (specifically, the column `F_PARTYSUM_FINAL`). We also note that in this analysis we exclude survey questions with fewer than three options.

Additional Examples. We highlight some of the examples of discrepancies that we find in some of the surveys listed above in Figures 6, 7, 8, 9, and 10. We note though that while we indeed find a few examples of the discrepancy, NBC rankings is actually aligned with average reward rankings in most cases. One possible explanation for this is that in many questions, the distributions of responses across types were very similar, suggesting that a homogeneous reward model would have been appropriate, causing NBC and average reward to align.

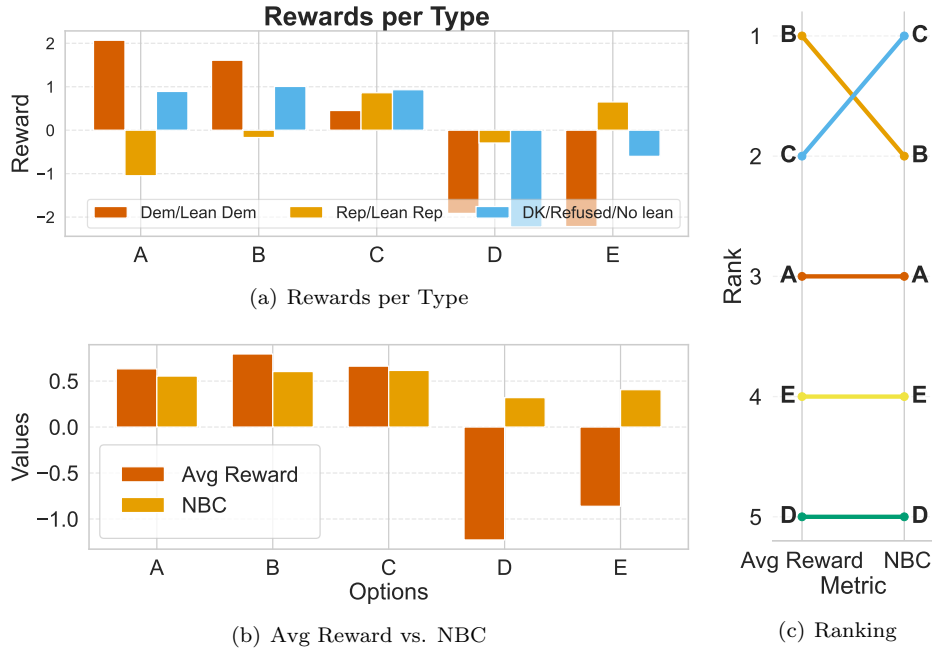


Figure 6: Do you think the plans and policies of the Biden administration will make the country’s response to the coronavirus outbreak: A: A lot better; B: A little better; C: Not much different; D: A little worse; E: A lot worse

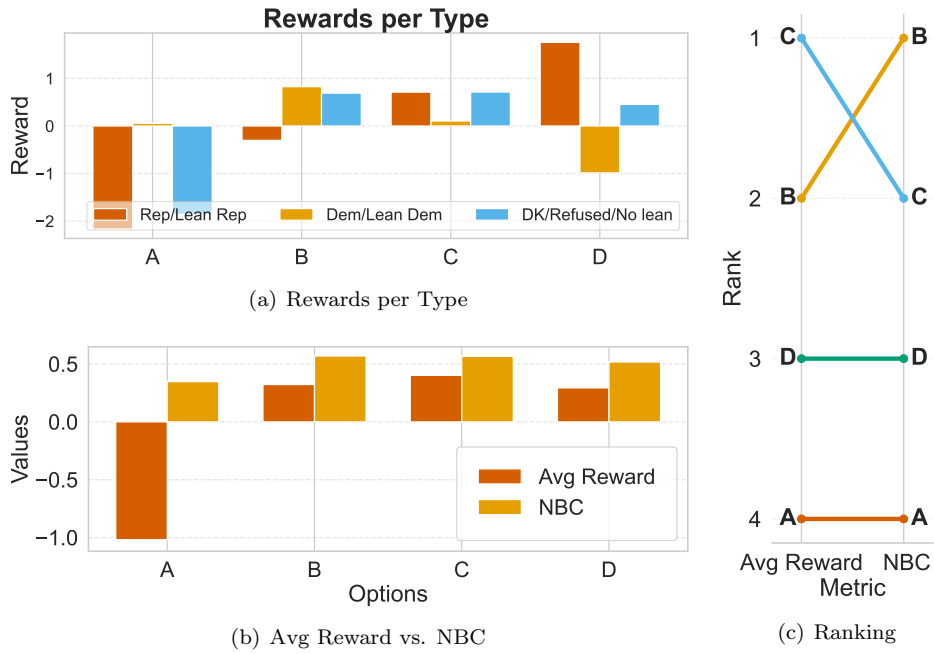


Figure 7: How would you rate the job Joe Biden is doing responding to the coronavirus outbreak? A: Excellent; B: Good; C: Only fair; D: Poor

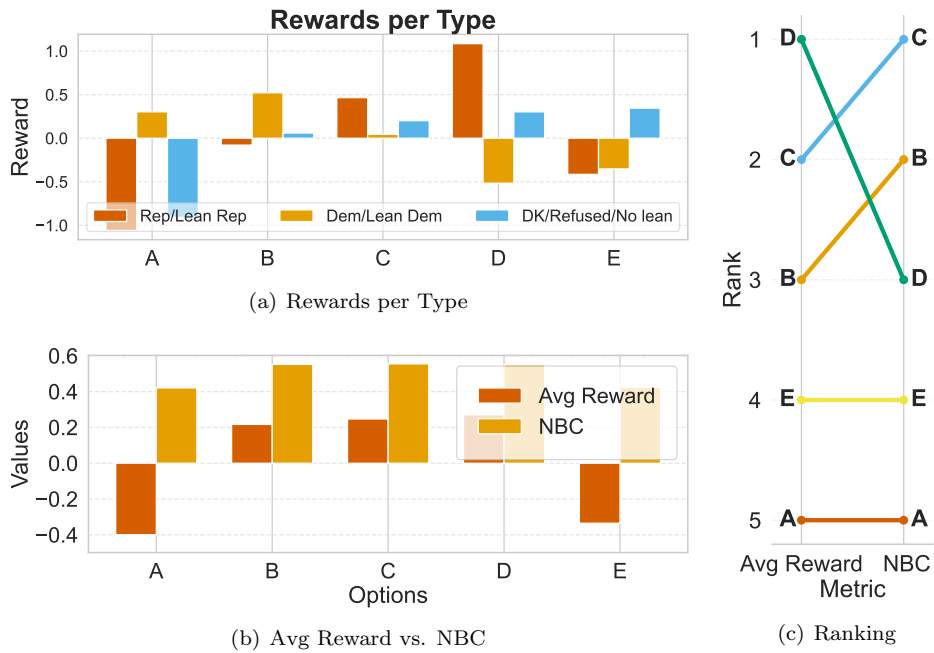


Figure 8: The next time you purchase a vehicle, how likely are you to seriously consider purchasing an electric vehicle? A: Very likely; B: Somewhat likely; C: Not too likely; D: Not at all likely; E: I do not expect to purchase a vehicle

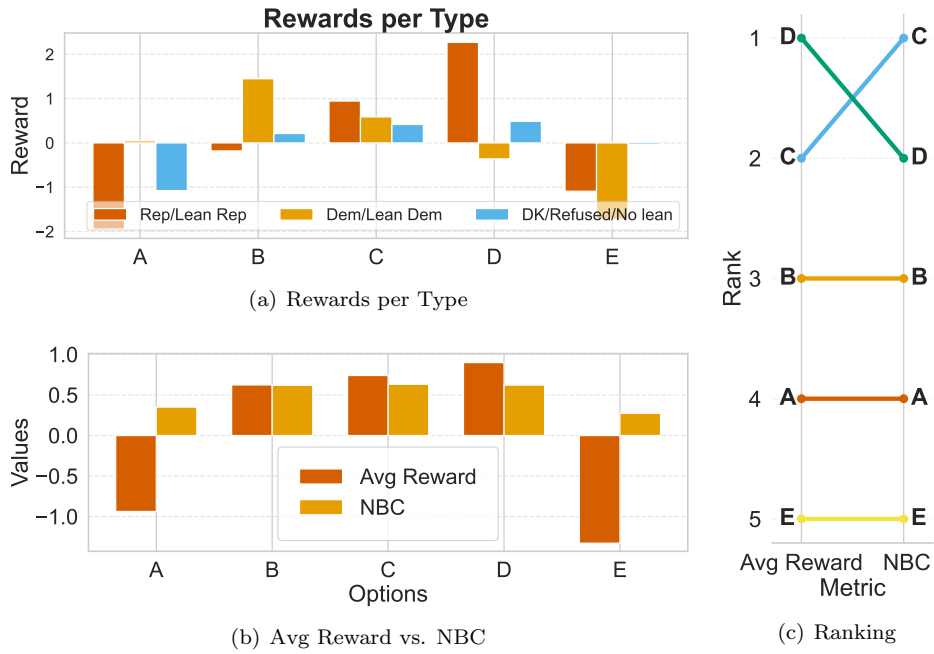


Figure 9: What is your overall opinion of Kamala Harris? A: Very favorable; B: Mostly favorable; C: Mostly unfavorable; D: Very unfavorable; E: Never heard of this person.

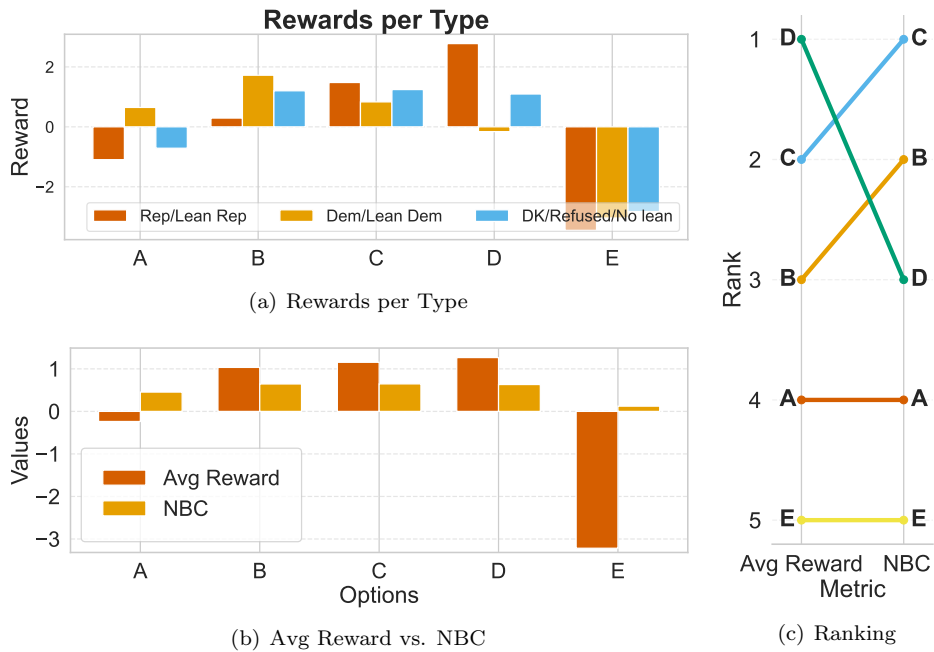


Figure 10: What is your overall opinion of Joe Biden? A: Very favorable; B: Mostly favorable; C: Mostly unfavorable; D: Very unfavorable; E: Never heard of this person.

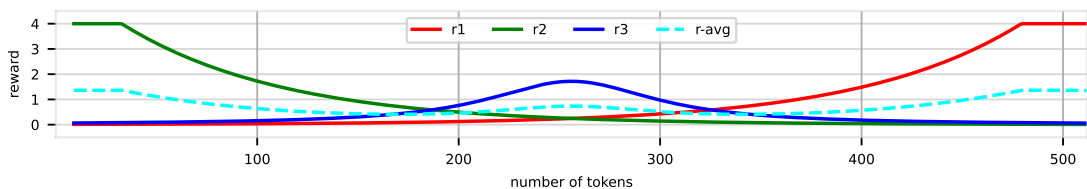


Figure 11: Reward definition for three user types in semi-synthetic experiments (Section 7.3) based on the length of prompt response combination. The first user type prefers long prompt response combinations, the second user type prefers short prompt response combinations, and the third user type prefers mid-length prompt response combinations. The dashed cyan line shows the average reward across the three user types.

E Semi-Synthetic Experiment: Fine-Tuning Llama-3-8B on HH-RLHF

Reward Models. Fig. 11 shows the three distinct rewards we use for the three user types along with their average. In order to have a reliable ground-truth reward which we can rely on in evaluation, we define these rewards as functions of the number of tokens in prompt-response combinations.

Anonymous Dataset. We use prompts and response pairs from both helpfulness and harmless subsets of Anthropic’s HH-RLHF dataset [38] and relabel the *chosen* and *rejected* responses manually. We filter for data points in which the sum of the number of tokens in the prompt and the number of tokens in the longer response do not exceed 512. This leaves us with 160,800 training and 17,104 test data points. For every data point (a prompt with a pair of responses), we sample one of the three user types uniformly at random. Given the type of user, we sample a preference based on BT [22] to label the two alternatives.

Dataset with Maximum Annotator Information. We use prompts and response pairs from both helpfulness and harmless subsets of Anthropic’s HH-RLHF dataset [38] and relabel the *chosen* and *rejected* responses manually. We filter for data points in which the sum of the number of tokens in the prompt and the number of tokens in the longer response do not exceed 512. This leaves us with 160,800 training and 17,104 test data points. For every data point (a prompt with a pair of responses), we keep sampling BT [22] preferences from all user types until they agree with each other. Once the consensus is achieved, we stop sampling and use the agreed-upon preference as the label for this data point.

Fine-Tuning Details. We fine-tune Llama-3-8B [37] base model with LoRA [36]. We fine-tune for one epoch with a batch size of 2, and use a linear learning rate schedule that starts with 3×10^{-5} and decreases to zero. We use the Adam optimizer with a weight decay of 0.001 [74]. Regarding LoRA’s hyper-parameters, we use the matrix rank of $r = 8$, $\alpha = 32$, and the dropout probability of 0.1.

For direct alignment experiments, we use a uniform reference policy. When ignoring heterogeneity, we do vanilla DPO over the anonymous dataset. When modeling heterogeneity, we use the loss function we propose in Proposition 6.1 over the dataset with maximum annotator information. We use the ordinal agreement between the ground-truth average reward and the reward induced by the aligned policy as the measure of accuracy.

For the reward learning experiments, we fine-tune the Llama-3-8B as a reward model. When ignoring heterogeneity, we assume BT and maximize the probability of the anonymous preference dataset under the learned reward model. When modeling heterogeneity, we use the loss function in

Table 1: Raw Accuracy (%) in Alignment Experiments

SEED	IGNORING HOMOGENEITY	MODELING HETEROGENEITY
0	65.61	66.63
1	67.55	69.55
2	68.77	75.21
3	68.70	72.04
4	66.28	74.95

Table 2: Raw Accuracy (%) in Reward Learning Experiments

SEED	IGNORING HOMOGENEITY	MODELING HETEROGENEITY
0	92.33	95.26
1	85.38	92.01
2	88.46	94.6
3	89.69	93.57
4	91.94	93.87

Proposition 6.1 over the dataset with maximum annotator information, but replace $h(\mathbf{y}_1, \mathbf{y}_2; \pi)$ with the difference in rewards, i.e., $r(\mathbf{y}_2) - r(\mathbf{y}_1)$. We use the ordinal agreement between the ground-truth average reward and the learned reward as the measure of accuracy.

Detailed Results. We conduct every experiment with five different random seeds. Fig. 5 shows the average, 25th percentile, and 75th percentile of accuracy across the five random seeds. Tables 1 and 2 show the raw accuracy numbers across the five random seeds.