Algorithmic Impact Reveals the Hidden Social Choice Structure of Alignment

Zachary Wojtowicz
MIT
Cambridge, MA 02139
zachwoj@mit.edu

Michelle Si
Harvard University
Cambridge, MA 02138
msi@g.harvard.edu

Finale Doshi-Velez
Harvard University
Cambridge, MA 02138
finale@seas.harvard.edu

Ariel Procaccia
Harvard University
Cambridge, MA 02138
arielpro@seas.harvard.edu

Abstract

When an AI algorithm makes decisions that affect more than one person, aligning it becomes a problem of social choice: how should people's divergent preferences about system behavior be reconciled and aggregated into a single coherent model? The standard approach to aligning frontier AI models—reinforcement learning from human feedback—largely sidesteps this question and has poor social choice guarantees. However, it remains unclear what alternative should replace it. We show that, by focusing directly on an algorithm's welfare consequences, the alignment problem can be reformulated as linear optimization over a convex *impact space*, which makes it amenable to the standard toolkit of welfare economics and mechanism design. This reformulation clarifies how alignment protocols translate into welfare consequences and, conversely, how a social planner's desired constraints on welfare consequences can be translated back into alignment protocols. We apply this transformation to show that voting-by-issues and random-dictatorship mechanisms are strategyproof and unanimous. Demonstrating the reverse direction, we also apply the impact representation to derive a family of alignment protocols that maximize utilitarian social welfare subject to various social desiderata, such as bounds on individual or group harm. We illustrate the welfare implications of these alignment protocols empirically using real human preferences over kidney allocation, charitable food distribution, LLM responses, and trolley problems. [1]

1 Introduction

Alignment is typically framed as the problem of ensuring that an AI system respects human preferences. However, AI systems increasingly take actions that affect more than one person, and some disagreement about how these systems should behave is inevitable. In such cases, alignment becomes a problem of *social choice*: how should an AI model combine the potentially conflicting preferences of many individuals to select among alternatives?

[1] All code can be found at https://github.com/michelleeesi/hiddenstructure.

Recently, *reinforcement learning from human feedback (RLHF)* has become the standard approach to aligning AI systems [Bai et al., 2022]. In this method, human annotators provide structured feedback on model responses, typically in the form of pairwise comparisons, which are then pooled together and used to train a model, either directly or through a proxy reward model. Although widely used, this approach does not directly engage with the social choice problem intrinsic to alignment, and recent work has shown that RLHF produces models that have a variety of potentially undesirable properties, such as the down-weighting of minority groups [Ge et al., 2024, Gölz et al., 2025, Shirali et al., 2025, Chakraborty et al., 2024, Chidambaram et al., 2024, Siththaranjan et al., 2023, Halpern et al., 2025].

These limitations have prompted recent calls to elicit preferences at a more granular level, such as the group or individual, rather than pooling all feedback anonymously [Park et al., 2024, Shirali et al., 2025, Li et al., 2024, Poddar et al., 2024, Dai and Fleisig, 2024]. However, this leaves open the question: how should individual preferences, once collected, be combined? Social choice theory was developed precisely to guide such decisions [Conitzer et al., 2024]. Yet applying classical results from the field to alignment has been complicated by the structure of modern AI systems, in which alternatives are high-dimensional model parameterizations rather than discrete options. The key insight of our paper is that this complexity is an artifact of the parameterization, not of the underlying alignment problem itself:

> *A decision-making algorithm can be summarized by a single impact vector that captures how its choices affect individual welfare during deployment. Once reformulated in impact space, aligning a model to maximize utilitarian social welfare reduces to optimizing a linear function over a convex polytope.*

Our main theorem establishes this correspondence. Specifically, we give a constructive procedure for both reformulating an alignment problem as an instance of *linear social choice* [Ge et al., 2024] and mapping the solution back to a deployable model parameter.

Contributions: We show that directly considering the distributional welfare impact of an AI model provides clean insights into the social choice properties of alignment protocols. We introduce our main result in Section 4, which shows that maximizing social welfare in the linear social choice setting is a linear optimization problem over a convex polytope that we call the impact space (Theorem 1). In Section 5, we show that in impact space strategyproof alignment mechanisms inherit the classical menu structure of *option-set* mechanisms (Lemma 2, after Barbera and Peleg, 1990), and we show that aggregating each deployment query by a monotone vote yields the strategyproof *voting-by-issues* rules (Theorem 2), with *random dictatorships* as a key example. We also establish that random dictatorships can be implemented, in expected impact, by a single model parameter (Theorem 3). In Section 6, we use our framework to characterize a family of mechanisms that maximize welfare subject to constraints on individual and group outcomes (Theorem 4), such as bounds on harm (Corollary 2) and public-spirited trade-offs between personal loss and public gain (Corollary 3). These applications highlight how working in impact space makes it easy to build social objectives beyond utilitarian welfare maximization into AI systems. Finally, in Section 7 we illustrate the welfare consequences of these alignment protocols using real human preferences from four domains: kidney allocation, charitable food distribution, LLM responses, and trolley problems.

2 Related Work

Our work builds on a recent literature that studies the distributional welfare implications and other social properties of RLHF, particularly for populations with diverse preferences.

Characterizing RLHF Distortion. Understanding the shortcomings of RLHF has become an area of heightened interest [Siththaranjan et al., 2023, Shirali et al., 2025, Gölz et al., 2025]. Gölz et al. [2025] measure how suboptimal (relative to a utilitarian optimum) an alignment method can be in aggregate, while we focus on what generates that suboptimality on a participant-to-participant (or group-to-group) basis. Shirali et al. [2025] show that pooled RLHF differs from using the average of individual choice probabilities by a variance term informed by preference dispersion.

Strategyproof Alignment. A separate line of work studies strategic behavior in preference collection [Sun et al., 2024, Kleine Buening et al., 2026]. Kleine Buening et al. [2026] show that standard RLHF procedures can be manipulated by strategic feedback providers and introduce the Pessimistic Median of MLEs algorithm, which is approximately strategyproof. We study a much more restricted domain (linear rewards) that gives simple results: the strategyproof alignment mechanisms are option-set mechanisms, with the random dictatorship as a focal member, and a random dictatorship need not be implemented by literally sampling one person's personalized model at deployment time—a single preference parameter $\theta_\lambda^{\mathrm{sp}}$ reproduces its expected impact.

Alternative Alignment Protocols. Other papers propose algorithmic alternatives for heterogeneous preferences [Chakraborty et al., 2024, Chidambaram et al., 2024, Park et al., 2024], from MaxMin-RLHF to personalization. These papers study ways to model or optimize over latent preference types given some desiderata, but our approach is more general and geometric: we show how different social objectives correspond to different optimization problems over that same set—the impact space. This lets us compare utilitarian alignment, strategyproof alignment, harm-bounded alignment, and public-spirit constraints within one common representation. Our work is most closely related to the linear social choice framework of Ge et al. [2024]. Theorem 1 formally establishes that, when utilities are linearly representable, alignment is an instance of linear social choice.

3 Preliminaries

3.1 Choice in the Linear Utility Setting

We begin with the standard model of stochastic choice assumed by *Direct Preference Optimization* [Rafailov et al., 2023] and related methods. Let X be a finite set of contexts, Y a finite set of actions, and N the number of agents. Agent n has utility $u_n : X \times Y \to \mathbb{R}$. For a binary choice problem (x, y, y'), we assume stochastic choice follows a Bradley–Terry–Luce model

$$p_n(y \succ y' \mid x, y, y') = \sigma\big(u_n(y \mid x) - u_n(y' \mid x)\big), \qquad \sigma(a) = \frac{\exp(a)}{1 + \exp(a)}.$$

Each context–action pair is embedded as $\phi(x, y) \in \mathbb{R}^k$. Our central assumption is that each agent's utility can be expressed linearly in this feature space: $u_n(y \mid x) = \phi(x, y)^\top \theta_n$ for $\theta_n \in \mathbb{R}^k$. For social welfare weights $w \in \Delta_N$, define the utilitarian preference parameter $\bar{\theta}_w = \sum_n w_n \theta_n$. By linearity,

the w-weighted utilitarian aggregate is represented by the same feature map:

$$\bar{u}_w(y \mid x) = \sum_n w_n u_n(y \mid x) = \phi(x, y)^\top \bar{\theta}_w.$$

Restricting attention to weighted sums of individual utilities is itself justified by the aggregation theorem of Harsanyi [1955], which shows that a social preference satisfying the expected-utility axioms over a convex set of prospects, together with Pareto indifference, must take exactly this form.

Under linear utilities, each pairwise comparison depends only on the feature difference $\alpha(x, y, y') = \phi(x, y) - \phi(x, y')$. We write $\alpha_z \in \mathbb{R}^k$ for the vector associated with a choice problem $z \in Z = X \times Y \times Y$. Let $q_{\text{train}} \in \Delta([N] \times Z)$ be the distribution of training query-agent pairs, and $q_{\text{dep}} \in \Delta(Z)$ be the distribution of deployment problems.

3.2 Deployment Welfare

A central objective for alignment is ensuring that model behaviors improve aggregate social welfare. We measure the welfare agent n receives from a model deployed at parameter θ as the amount of utility it generates relative to the baseline of choosing randomly between the two options,

$$U_n(\theta) = \mathbb{E}_{z \sim q_{\text{dep}}} \left[\alpha_z^\top \theta_n \left(\sigma(\alpha_z^\top \theta) - \frac{1}{2} \right) \right].$$

The utilitarian social welfare of a deployed model is then the sum of the individual welfares, $U(\theta) = \sum_{n=1}^N w_n U_n(\theta)$.

4 Framework: Alignment as Linear Optimization in Impact Space

As reviewed in Section 2, recent literature has catalogued many ways that anonymous RLHF can fail to represent heterogeneous populations, violating basic welfare criteria such as Pareto efficiency [Gölz et al., 2025, Shirali et al., 2025, Ge et al., 2024, Chakraborty et al., 2024, Park et al., 2024]. A natural remedy is to estimate preferences at the individual or group level and then combine these estimates into a single deployed model. But this immediately raises the social choice questions that anonymous RLHF sidesteps: which aggregation rule should be used? What welfare guarantees does it provide? Can agents manipulate it? Answering these questions requires a framework in which the welfare consequences of different alignment protocols can be directly compared. However, even if the goal of alignment is to combine the preferences of many individuals, working directly in model-parameter space obscures the comparison, because welfare is a nonlinear function of the model's parameters.

The key insight of our framework is that this nonlinearity can be avoided if we consider the set of potential welfare implications directly, instead of the parameterization.

Under our linear utility assumptions, a model's welfare effects can be summarized by a single vector—its *impact*—that lives in the same space as preferences and deployment queries (see Figure 1). Welfare is linear in this vector, so the social alignment problem reduces to optimizing a linear function over a convex set: a problem for which we have the tools of welfare economics and mechanism design at our disposal.

Definition 1. The **impact** of a model with parameter θ is $\psi(\theta) = \mathbb{E}_{z \sim q_{\text{dep}}} \left[\alpha_z \left(\sigma(\alpha_z^\top \theta) - \frac{1}{2} \right) \right] \in \mathbb{R}^k$.

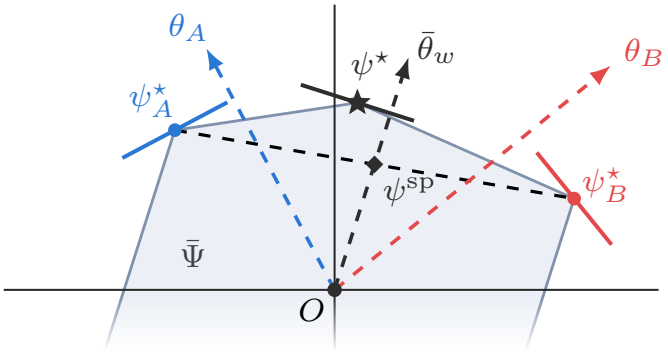


Figure 1: **The impact-space view of social alignment.** Preferences (θ_n, dashed) and the feasible impact set $\bar{\Psi}$ live in the same Euclidean space; for finitely many deployment queries, $\bar{\Psi}$ is a convex, origin-symmetric polytope. Preferences differ in magnitude and need not themselves be feasible impacts: both θ_A and θ_B lie outside $\bar{\Psi}$. Because welfare $U_n(\psi) = \theta_n^\top \psi$ is linear, each agent's ideal impact $\psi_n^\star = \arg\max_{\psi \in \bar{\Psi}} \theta_n^\top \psi$ is the vertex of $\bar{\Psi}$ farthest in the direction of θ_n, which in general is *not* collinear with θ_n. The utilitarian optimum $\psi^\star$ maximizes welfare in the weighted-average direction $\bar{\theta}_w = \sum_n w_n \theta_n$, where the iso-welfare line $\{\psi : \bar{\theta}_w^\top \psi = \text{const}\}$ is tangent to $\bar{\Psi}$. The dashed segment $\mathrm{conv}(\psi_A^\star, \psi_B^\star)$ is the set of expected impacts attainable by strategyproof mechanisms (Section 5), with ψ^{sp} the equal-weight random dictatorship.

The impact vector is a sufficient statistic for welfare. Since $U_n(\theta) = \mathbb{E}_{z \sim q_{\mathrm{dep}}}[\alpha_z^\top \theta_n (\sigma(\alpha_z^\top \theta) - \frac{1}{2})] = \theta_n^\top \psi(\theta)$, agent n's welfare is the inner product of their preference with the impact vector.[2] Utilitarian welfare is therefore also linear: $U(\psi) = \bar{\theta}_w^\top \psi$.

This lets us rewrite social alignment as a linear optimization over the set of feasible impacts:

$$\psi^\star \in \arg\max_{\psi \in \bar{\Psi}} \sum_{n=1}^{N} w_n \theta_n^\top \psi, \qquad \bar{\Psi} = \mathrm{closure}\{\psi(\theta) : \theta \in \mathbb{R}^k\} \subseteq S_{\mathrm{dep}}. \tag{1}$$

Here, S_{dep} is the subspace spanned by deployment queries, and $\bar{\Psi}$ is the feasible set of impacts achievable by some model parameter.[3] The following theorem characterizes the geometry of the impact set.

Theorem 1. *Let $\Psi = \{\psi(\theta) : \theta \in \mathbb{R}^k\}$ and $\bar{\Psi} = \mathrm{cl}(\Psi)$, and let S_{dep} be the subspace of $\mathbb{R}^k$ spanned by the deployment queries. Then:*

1. *$\bar{\Psi} \subseteq S_{\mathrm{dep}}$ is the origin-symmetric* zonotope *$\bar{\Psi} = \{\mathbb{E}_{z \sim q_{\mathrm{dep}}}[t_z\, \alpha_z] : t_z \in [-\frac{1}{2}, \frac{1}{2}]\}$ generated by the deployment queries, and $\Psi = \mathrm{relint}\, \bar{\Psi}$, the interior of $\bar{\Psi}$ relative to S_{dep}. In particular, $\bar{\Psi}$ is convex and centrally symmetric ($\bar{\Psi} = -\bar{\Psi}$), and ψ is antisymmetric ($\psi(-\theta) = -\psi(\theta)$) with $\psi(0) = 0$.*

2. *On S_{dep}, the map ψ is a C^∞ diffeomorphism from S_{dep} onto Ψ. Every feasible impact $\psi \in \Psi$ is implemented by a unique parameter in S_{dep}.*

3. *ψ is bounded, with $\sup_{\theta \in \mathbb{R}^k} \|\psi(\theta)\| \le \frac{1}{2}\, \mathbb{E}_{z \sim q_{\mathrm{dep}}} \|\alpha_z\|$.*

4. *Scaling is monotone in the direction of θ: $\theta^\top \frac{\partial \psi(\nu\theta)}{\partial \nu} = \mathbb{E}_{z \sim q_{\mathrm{dep}}} \left[(\alpha_z^\top \theta)^2 v(\nu \alpha_z^\top \theta)\right] \ge 0$, where $v(x) = \sigma'(x)$.*

[2] Economically, ψ is analogous to a vector of quantities delivered and θ_n encodes agent n's marginal valuations.
[3] We optimize over the closure because deterministic limits can be approached to arbitrary precision.

5. *The deterministic boundary in direction θ is* $\lim_{\nu\to\infty}\psi(\nu\theta) = \frac{1}{2}\mathbb{E}_{z\sim q_{\text{dep}}}\left[\alpha_z \operatorname{sign}(\alpha_z^\top\theta)\right]$. *For generic θ (i.e., $\alpha_z^\top\theta \neq 0$ for all $z \in \operatorname{supp}(q_{\text{dep}})$), this limit is the unique maximizer of $\theta^\top\psi$ over $\bar{\Psi}$, and it is a vertex.*

This structure allows us to restate the utilitarian optimal model as the boundary point of $\bar{\Psi}$ in the direction of the weighted average preference $\bar{\theta}_w$.

Proposition 1. *The limit* $\lim_{\nu\to\infty}\psi(\nu\,\bar{\theta}_w)$ *exists and is a utilitarian-optimal impact:* $\lim_{\nu\to\infty}\psi(\nu\,\bar{\theta}_w) \in \arg\max_{\psi\in\bar{\Psi}}\bar{\theta}_w^\top\psi$. *In particular, we may take* $\psi^\star = \lim_{\nu\to\infty}\psi(\nu\,\bar{\theta}_w)$, *and this maximizer is unique whenever* $\alpha_z^\top\bar{\theta}_w \neq 0$ *for all* $z \in \operatorname{supp}(q_{\text{dep}})$.

We can now explicitly describe how scaling affects individual welfare. For agent n,

$$\frac{\partial U_n(\nu\theta)}{\partial\nu} = \theta_n^\top\frac{\partial\psi(\nu\theta)}{\partial\nu} = \mathbb{E}_{z\sim q_{\text{dep}}}\left[(\alpha_z^\top\theta_n)(\alpha_z^\top\theta)v(\nu\alpha_z^\top\theta)\right].$$

Thus scaling *helps* agents whose preferences align with the model direction across deployment queries, and *harms* agents whose preferences covary negatively.

More generally, translating the alignment problem into impact space and considering the entire problem from the perspective of welfare invokes *externalities* as the objects we would like to regulate as we design alignment protocols. Our framework gives us the machinery to do so. In the remaining sections of this paper, we discuss how our framework leads to straightforward ways of addressing various types of externalities.

5 Strategyproof Alignment in Impact Space

When individuals have a say in how an AI system is aligned, they may be tempted to misreport their preferences to steer the outcome in their favor. Even though this problem is not a major issue for chatbots and other present AI platforms, it will likely become more salient as algorithms are deployed in increasingly consequential social domains, and their behavioral policies therefore become the subject of more direct and explicit contention among impacted stakeholders. Misreporting one's true preferences to distort alignment in a self-serving way imposes welfare externalities on others. In economics, the field of mechanism design regulates such externalities by identifying procedures for combining preferences that are *strategyproof*—i.e., such that no agent can benefit from misreporting, so truthful participation is always a weakly dominant strategy.

The impact-space formulation places social alignment within the classical theory of strategyproof mechanism design, enabling us to apply its toolkit directly. Because welfare is linear in the impact vector and the feasible impact set $\bar{\Psi}$ is convex, we identify strategyproof alignment mechanisms in three nested steps. The general family consists of the *option-set* mechanisms of Barbera and Peleg [1990]. Restricting to per-query aggregation yields *voting-by-issues* rules. A leading example is the *random dictatorship*, which we show is implementable by a single model parameter (exactly or in the limit) and is the deterministic limit of anonymous RLHF when the labeling pool is common across queries.

Our treatment of strategyproofness complements recent work by Kleine Buening et al. [2026], who study strategic behavior in offline RLHF. Their setting is a learning problem: labelers strategically manipulate comparison labels, and they propose an approximately strategyproof learning algorithm. In contrast, we abstract away from statistical estimation and focus directly on the impact

that preference reports have on an AI model's behavior and, through it, the distribution of welfare among those impacted. By taking advantage of our linearity assumption, we can characterize a family of simple and intuitive strategyproof mechanisms.

Definition 2. A **mechanism** is a function $\mu : \mathbb{R}^{N\times k} \to \Delta(\bar{\Psi})$ that maps a report profile $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_N)$ to a probability distribution over the space of impact vectors $\bar{\Psi}$.

Throughout this section we restrict attention to *generic* profiles and reports—those with $\alpha_z^\top \theta_n \neq 0$ for every $z \in \mathrm{supp}(q_{\mathrm{dep}})$ and every n—so that vote signs and ideal impacts $\psi_n^\star = \arg\max_{\psi\in\bar{\Psi}} \theta_n^\top \psi$ are single-valued (Theorem 1). Equivalently, one may fix any tie-breaking convention on the measure-zero set of non-generic reports.

Intuitively, we want a mechanism to satisfy two basic requirements. The first, as we have discussed, is *strategyproofness*: an agent should not be able to obtain a better outcome, according to their true preferences, by misreporting those preferences.

Definition 3. A mechanism is **strategyproof** if, for all preference profiles $\theta \in \mathbb{R}^{N\times k}$ and $n \in [N]$, all θ_n',

$$\mathbb{E}\big[\theta_n^\top \mu(\theta)\big] \geq \mathbb{E}\big[\theta_n^\top \mu(\theta_n', \theta_{-n})\big],$$

where (θ_n', θ_{-n}) denotes the profile obtained by replacing agent n's true preferences with θ_n'.

The second is *unanimity*: If all agents have exactly the same preferences, then there is no social conflict to resolve: the mechanism should simply choose the impact vector that is best according to that shared preference.

Definition 4. A mechanism μ is **unanimous** if, for any profile θ such that $\theta_n = \bar{\theta}$ for all $n \in [N]$, $\mu(\theta)$ is the point mass on $\arg\max_{\psi\in\bar{\Psi}} \bar{\theta}^\top \psi$ (a singleton, since $\bar{\theta}$ is generic).

We point out a powerful consequence of the linearity of $\theta_n^\top \psi$: that a stochastic mechanism can be fully characterized by its *expected* impact vector at each profile.

Lemma 1. *Let $\bar{\mu}(\theta) = \mathbb{E}_{\psi\sim\mu(\theta)}[\psi] \in \bar{\Psi}$ be the expected impact. Then $\mathbb{E}[\theta_n^\top \mu(\theta)] = \theta_n^\top \bar{\mu}(\theta)$, so strategyproofness and unanimity depend on μ only through the expected-impact map $\bar{\mu} : \mathbb{R}^{N\times k} \to \bar{\Psi}$.*

This enables us to characterize classes of alignment protocols satisfying strategyproofness and unanimity while restricting attention to deterministic expected-impact maps.

5.1 The General Family: Option-Set Mechanisms

The impact space reformulation enables us to recognize social alignment as an instance of strategyproof social choice, which can be characterized generally using the option-set principle [Barbera and Peleg, 1990]. For a report profile, let

$$O_n(\theta_{-n}) = \{\bar{\mu}(\theta_n', \theta_{-n}) : \theta_n' \in \mathbb{R}^k\} \subseteq \bar{\Psi}$$

be agent n's *option set*: the impacts it can induce by varying its own report while the others' reports are held fixed.

Lemma 2. *A mechanism μ is strategyproof if and only if there exist menus $M_n(\theta_{-n}) \subseteq \bar{\Psi}$, each depending only on the other agents' reports, such that, at every profile, the mechanism deploys each agent's most-preferred impact from its menu:*

$$\bar{\mu}(\theta) \in \operatorname*{arg\,max}_{\psi \in M_n(\theta_{-n})} \theta_n^\top \psi \qquad \forall n. \tag{2}$$

In particular, we can set $M_n(\theta_{-n}) = O_n(\theta_{-n})$ to be the option set itself.

The lemma holds for any convex feasible set. It says that a strategyproof alignment protocol is equivalent to one that posts a menu of achievable model behaviors to each participant, which they cannot themselves enlarge, and deploys their favorite. The menu form is exact but implicit: an explicit characterization of all strategyproof mechanisms for expected-utility agents is a long-standing open problem even in the classical case of lotteries over finitely many alternatives [Barberà, 2011]. We therefore proceed by explicitly identifying structured subfamilies, each a member of the menu family above.

5.2 Reduction 1: Voting by Deployment Query

As established by Theorem 1, the impact set is a zonotope generated by the deployment queries. By decomposing the expectation over deployment queries into its individual terms, we can construct a strategyproof mechanism that aggregates the population's preferences query by query. To ensure strategyproofness, the mechanism must ensure that agents cannot benefit by exaggerating the magnitude of their reports, which can be accomplished by reducing preference information to directional votes on each query. Write $s_{z,n} = \operatorname{sign}(\alpha_z^\top \theta_n) \in \{\pm 1\}$ for how agent n would decide query z.[4]

Definition 5. A **voting-by-issues** mechanism is specified by, for each query z, an aggregator $f_z : \{\pm 1\}^N \to [-1, 1]$ that is nondecreasing in each argument with $f_z(1, \ldots, 1) = 1$ and $f_z(-1, \ldots, -1) = -1$. Its expected impact is

$$\bar{\mu}(\theta) = \tfrac{1}{2}\, \mathbb{E}_{z \sim q_{\text{dep}}} \big[f_z(s_{z,1}, \ldots, s_{z,N})\, \alpha_z \big] \tag{3}$$

Theorem 2. *Every voting-by-issues mechanism is strategyproof and unanimous.*

Voting-by-issues rules are the impact-space analogue of *generalized median voter schemes* [Barberà et al., 1993]: each deployment query is settled by a monotone aggregation of the participants' preferred sides. Our setting differs from those schemes in that mechanisms are stochastic, outcomes range over the impact space, and preferences are linear in the same space. The fact that the impact set is a zonotope generated by the queries, and therefore exhibits central symmetry, makes this per-query aggregation feasible.

5.3 Reduction 2: Random Dictatorships

Definition 6. A mechanism is a **random dictatorship** with weights $\lambda \in \Delta_N$ if $\mu(\theta) = \sum_{n=1}^N \lambda_n \delta(\psi_n^\star)$. Under Lemma 1, this is impact-equivalent to $\bar{\mu}(\theta) = \sum_{n=1}^N \lambda_n \psi_n^\star$.

[4] For (z, n) such that $\alpha_z^\top \theta_n = 0$, an arbitrary tie-breaking rule can be used, which does not affect welfare and therefore does not affect strategyproofness.

Proposition 2. *A random dictatorship is impact-equivalent to the voting-by-issues mechanism whose aggregators are the vote averages* $f_z(s_{z,\cdot}) = \sum_n \lambda_n s_{z,n}$. *It is therefore strategyproof and unanimous.*

Classical results force strategyproofness to coincide with random dictatorship alone, but only on richer preference domains: a universal domain over the outcomes [Gibbard, 1977], or strictly convex single-peaked preferences over a convex set [Dutta et al., 2002]. The assumption that preferences can be represented linearly in the model's feature space is not without loss, as it restricts the power of more general preference profiles to constrain the space of mechanisms. This yields a richer strategyproof class, with random dictatorship as a focal example.

As defined, a random dictatorship draws a single individual, with probability λ_n, and deploys their ideal impact. By Lemma 1 it is impact-equivalent to the more natural protocol of choosing an individual at random on a query-by-query basis and deciding each query the way that individual would. A naive implementation of either would fit a separate optimal model for each individual (e.g., a collection of person-specific LoRA adapters) and randomly select among them at deployment.

We now show something more powerful: this apparently impractical lottery can be compiled into a *single* static parameter. It can be implemented exactly when its target impact lies in the relative interior of $\bar{\Psi}$, and as a limit of single parameters otherwise.

Theorem 3. *The random dictatorship with weights* $\lambda \in \Delta_N$ *has target impact* $\bar{\psi}_\lambda = \sum_{n=1}^N \lambda_n \psi_n^\star \in \bar{\Psi}$. *This impact is attained by a single model parameter, which then reproduces the random dictatorship's welfare for every agent, if and only if* $\bar{\psi}_\lambda \in \Psi$, *in which case the parameter is* $\theta_\lambda^{\mathrm{sp}} = \psi^{-1}(\bar{\psi}_\lambda)$. *When* $\bar{\psi}_\lambda$ *lies on the boundary of* $\bar{\Psi}$, *it is instead the limit of the impacts of a sequence of single parameters.*

Even though $\theta_\lambda^{\mathrm{sp}}$ is one fixed parameterization, its stochastic responses reproduce, in expectation, the same impact as randomizing over agents' ideal models. Mathematically, it follows from the invertibility of ψ; conceptually, it reflects that model stochasticity is itself a form of compromise between conflicting individuals (see Figure 3). Recall $\psi_n^\star = \arg\max_{\psi \in \bar{\Psi}} \theta_n^\top \psi$ is agent n's ideal impact.

Having established that a random dictatorship is strategyproof and can be compiled into a single model parameter, we now ask how much doing so constrains the achievable welfare. First, note that we can select a set of weights λ to achieve any $\psi \in \Psi^{\mathrm{sp}} = \mathrm{Hull}(\{\psi_n^\star\}_{n=1}^N)$. Intuitively, each $\psi_n^\star$ is the support point of $\bar{\Psi}$ in the direction of θ_n. If agent preferences are spread over all directions, then their convex hull Ψ^{sp} will approximate $\bar{\Psi}$. Welfare will be low when social welfare maximization selects a "compromise" impact vector that is dissimilar to all individual vectors.

Proposition 3. *Let* $\bar{\Psi} \subset \mathbb{R}^k$ *be any compact, convex, origin-symmetric set with* $\dim(\bar{\Psi}) \geq 2$, *suppose the welfare weights* $w \in \Delta_N$ *are not a point mass, and write* $U^\star(\theta) = \max_{\psi \in \bar{\Psi}} \theta_w^\top \psi$ *for the optimal utilitarian welfare. For any weights* $\lambda \in \Delta_N$ *and any* $\delta > 0$, *there exists a preference profile* θ *with* $U^\star(\theta) > 0$ *at which the random dictatorship* μ_λ^{sp} *attains at most a* δ *fraction of the optimum,* $\mathbb{E}[U(\mu_\lambda^{sp}(\theta))] \leq \delta\, U^\star(\theta)$. *Hence, for non-degenerate welfare weights, the worst-case distortion* $U^\star(\theta)/\mathbb{E}[U(\mu_\lambda^{sp}(\theta))]$ *of every random dictatorship is unbounded.*[5]

[5]Some restriction on w is necessary: if $w = \delta_i$, then $U^\star(\theta) = \max_{\psi \in \bar{\Psi}} \theta_i^\top \psi = \theta_i^\top \psi_i^\star$, while origin symmetry gives $\theta_i^\top \psi \geq -U^\star(\theta)$ for all $\psi \in \bar{\Psi}$; hence every profile satisfies $\mathbb{E}[U(\mu_\lambda^{\mathrm{sp}}(\theta))] = \sum_n \lambda_n \theta_i^\top \psi_n^\star \geq (2\lambda_i - 1)\, U^\star(\theta)$, so for $\lambda_i > \frac{1}{2}$ the distortion is bounded.

Intuitively, the proof shows that this worst case arises when an individual or small group has a very strong preference along some dimension that the overwhelming majority has a very weak preference in the opposite direction. In such cases, a random dictatorship will implement the weak preferences of the majority, which causes outsized harm to the minority. In the applied domains that we study, this structure does not generally arise and the strategyproof mechanism's welfare sub-optimality is bounded (see Figure 8 in Appendix E).

These results also shed new light on the standard practice of RLHF. Anonymous RLHF fits a single parameter $\hat{\theta}^{\mathrm{RLHF}}$ to the pooled labels, ignoring who produced them. Our unifying result is that its deployed impact is exactly that of a simple stochastic protocol—answer each query by sampling a labeler and following their choice—and therefore depends on the training data only through the population's per-query label frequencies.

Proposition 4. *Suppose q_{train} and q_{dep} share the same query marginal, and let $\hat{\theta}^{\mathrm{RLHF}}$ be any stationary point of the anonymous RLHF objective. Writing $\mu_z = \mathbb{E}_{n\sim q_{\mathrm{train}}(\cdot|z)}[\mathbb{E}[c \mid n, z]]$ for the population's label frequency on query z,*

$$\psi(\hat{\theta}^{\mathrm{RLHF}}) = \mathbb{E}_{z\sim q_{\mathrm{dep}}}\big[(\mu_z - \tfrac{1}{2})\,\alpha_z\big].$$

That is, the deployed model is impact-equivalent to the random-labeler protocol that answers each deployment query z by sampling an agent $n \sim q_{\mathrm{train}}(\cdot \mid z)$ and a response $c \sim p(\cdot \mid n, z)$.

The identity holds because the RLHF first-order condition matches the first moment of $\sigma(\alpha_z^\top \hat{\theta}^{\mathrm{RLHF}})$ against α_z to that of μ_z, and the impact ψ is exactly that moment. In the deterministic limit of labeler accuracy, this turns into a per-query vote count, and the impact into that of a voting-by-issues rule.

Corollary 1. *Suppose, as in Proposition 4, that q_{train} and q_{dep} share the same query marginal. In the deterministic labeling limit $\mathbb{E}[c \mid n, z] \to \mathbf{1}\{\alpha_z^\top \theta_n > 0\}$, anonymous RLHF converges to the voting-by-issues mechanism with per-query aggregators $f_z(s_{z,\cdot}) = \sum_n q_{\mathrm{train}}(n \mid z)\, s_{z,n}$, and is therefore strategyproof and unanimous. When the labeling weights are query-independent, $q_{\mathrm{train}}(n \mid z) = \lambda_n$, it is the random dictatorship with weights λ.*

This highlights a connection between RLHF and the strategyproof voting-by-issues family, reaching the random-dictatorship point when the labeling pool is the same across queries.

6 Linear Constraints Can Enforce Externality Bounds and Trade-offs

In practice, alignment designers have objectives beyond strictly maximizing utilitarian welfare or ensuring strategyproofness. A designer may want to pursue high total welfare while guaranteeing that no individual or group is harmed beyond some threshold or, more generally, to guarantee normative commitments about how the benefits and costs of alignment are distributed.

The impact-space formulation makes this task straightforward. Because welfare is linear in the impact vector, any welfare guarantee expressible as a linear inequality becomes a halfspace constraint on ψ. As we show below, a floor on individual welfare, a bound on the ratio of personal harm to social benefit, or a constraint that no individual imposes externalities above some level can all be expressed in this way.

When augmented with such a linear constraint, the designer's problem becomes:

$$\psi^{\text{wel}} \in \arg\max_{\psi \in \bar{\Psi}} \bar{\theta}_w^\top \psi \quad \text{s.t.} \quad \gamma_n^\top \psi \geq c_n \ \forall\, n, \tag{4}$$

for constraint directions $\gamma_n \in \mathbb{R}^k$ and thresholds $c_n \in \mathbb{R}$. Since each constraint is a halfspace and $\bar{\Psi}$ is convex, closed, and bounded, this is a linear optimization with linear constraints whenever the feasible set is nonempty. Different choices of γ_n and c_n encode different commitments. We develop three natural families below—absolute welfare floors, public-spirit constraints, and participation-externality bounds—and show that all three admit closed-form characterizations.

Theorem 4. *If $\bar{\Psi} \cap \bigcap_n \{\psi : \gamma_n^\top \psi \geq c_n\}$ is nonempty, then an optimum ψ^{wel} exists. Moreover, there are multipliers $\eta_n \geq 0$ such that*

$$\psi^{\text{wel}} \in \arg\max_{\psi \in \bar{\Psi}} \left(\bar{\theta}_w + \sum_{n=1}^{N} \eta_n \gamma_n \right)^\top \psi,$$

with complementary slackness $\eta_n(\gamma_n^\top \psi^{\text{wel}} - c_n) = 0$ for all n. Equivalently, only the binding constraints $\mathcal{B} = \{n : \eta_n > 0\}$ tilt the objective, giving effective direction $\bar{\theta}_w + \sum_{n \in \mathcal{B}} \eta_n \gamma_n$. The optimal utilitarian value is weakly decreasing and concave in the thresholds c.

6.1 Example: Bounding Harm

A natural welfare guarantee is to require every agent's welfare to exceed a floor b_n. This is the special case $\gamma_n = \theta_n$ and $c_n = b_n$.

Definition 7 (Harm-bounded alignment)**.** For welfare floors $b \in \mathbb{R}^N$,

$$\psi_b^{\text{h}} \in \arg\max_{\psi \in \bar{\Psi}} \bar{\theta}_w^\top \psi \quad \text{s.t.} \quad \theta_n^\top \psi \geq b_n \ \forall n.$$

Corollary 2 (of Theorem 4)**.** *The harm-bounded optimum satisfies*

$$\psi_b^{\text{h}} \in \arg\max_{\psi \in \bar{\Psi}} \left(\sum_{n=1}^{N} (w_n + \eta_n) \theta_n \right)^\top \psi,$$

where $\eta_n \geq 0$ and $\eta_n(\theta_n^\top \psi_b^{\text{h}} - b_n) = 0$ for all n.

The corollary has a simple interpretation. Instead of maximizing with fixed weights w_n, the planner maximizes with effective weights $w_n + \eta_n$. The extra term η_n is positive exactly for agents who would fall below their required welfare level without additional protection. Thus the constrained solution can be implemented as an ordinary weighted-welfare maximization problem, but with extra weight placed on the agents that need protection.

Useful choices of welfare floor include: (i) $b_n = 0$, so no agent is worse off than under a random model; (ii) $b_n = \theta_n^\top \psi_w^{\text{sp}}$, so every agent weakly prefers the mechanism to the welfare-weighted strategyproof solution; and (iii) $b_n = -\epsilon h_{\bar{\Psi}}(\theta_n)$, so agent n can lose at most an ϵ fraction of their maximum achievable welfare.[6]

[6]The same construction applies to groups by replacing θ_n with a group preference $\bar{\theta}_{\mathcal{G}} = \sum_{n \in \mathcal{G}} w_n \theta_n$ and b_n with a group floor.

6.2 Example: Public Spirit

Another example of a welfare constraint is *public spirit* [Flanigan et al., 2023]. This constraint allows people to tolerate personal harm in proportion to the social benefit created, where $\gamma \in [0,1]$ is an individual's degree of public spirit.[7]

Definition 8 (Public-Spirit Alignment)**.** For a public-spirit parameter $\gamma \in [0,1]$ and reference impact ψ_0:

$$\psi_\gamma^{\mathrm{ps}} \in \arg\max_{\psi \in \bar{\Psi}} \quad \bar{\theta}_w^\top \psi \qquad \text{s.t.} \quad (\gamma\,\bar{\theta}_w + (1-\gamma)\theta_n)^\top \psi \geq (\gamma\,\bar{\theta}_w + (1-\gamma)\theta_n)^\top \psi_0 \quad \forall\, n.$$

The constraint has a direct interpretation, saying that agent n will accept some personal loss only when it is justified by enough social gain: $(1-\gamma)\,\theta_n^\top(\psi_0 - \psi) \leq \gamma\,\bar{\theta}_w^\top(\psi - \psi_0)$. When $\gamma = 0$, this reduces to an individual rationality constraint: no agent can be made worse off relative to ψ_0.

Corollary 3 (of Theorem 4)**.** *The public-spirit optimum satisfies*

$$\psi_\gamma^{\mathrm{ps}} \in \arg\max_{\psi \in \bar{\Psi}} \quad \left(\left(1 + \gamma \sum_{n \in \mathcal{B}} \eta_n\right)\bar{\theta}_w + (1-\gamma)\sum_{n \in \mathcal{B}} \eta_n\,\theta_n\right)^\top \psi, \tag{5}$$

where $\eta_n \geq 0$ *and* $\eta_n \left[\left(\gamma\bar{\theta}_w + (1-\gamma)\theta_n\right)^\top(\psi_\gamma^{\mathrm{ps}} - \psi_0)\right] = 0 \qquad \forall n.$

Figure 2 shows how each mechanism distributes welfare across individuals at high choice precision. Under the utilitarian, RLHF, and strategyproof mechanisms the distribution is wide and carries mass below zero—agents who are actively harmed. The harm-bounded mechanism removes this left tail: as the welfare floor b tightens toward 0, more agents' constraints bind and the distribution compresses upward against the floor, at the cost of some reduction in mean welfare. Appendix E.2 traces the same effect as a function of the choice-precision parameter β.

6.3 Example: Bounding Participation Externalities

The welfare floor (Corollary 2) protects agents who are *harmed*—it increases their influence over the deployed model. Another potential concern is limiting the harm that any single agent's participation *imposes on everyone else*. We formally define the participation externality in Appendix D.1. Here we need only the *leave-one-out utilitarian preference*, the welfare-weighted average preference of everyone other than n,

$$\bar{\theta}_{-n} = \frac{1}{1 - w_n} \sum_{m \neq n} w_m \theta_m, \tag{6}$$

which is well defined whenever $w_n < 1$.

Definition 9 (Externality-bounded alignment)**.** For a reference impact ψ_0 and tolerances $b \in \mathbb{R}^N_{\geq 0}$: $\psi_b^{\mathrm{ext}} \in \arg\max_{\psi \in \bar{\Psi}} \bar{\theta}_w^\top \psi \quad \text{s.t.} \quad \bar{\theta}_{-n}^\top \psi \geq \bar{\theta}_{-n}^\top \psi_0 - b_n \quad \forall\, n.$

[7] This is a special case of Theorem 4 with $\gamma_n = \gamma\bar{\theta}_w + (1-\gamma)\theta_n$ and $c_n = (\gamma\bar{\theta}_w + (1-\gamma)\theta_n)^\top \psi_0$, where ψ_0 is a reference impact (e.g., the status quo or the random model).

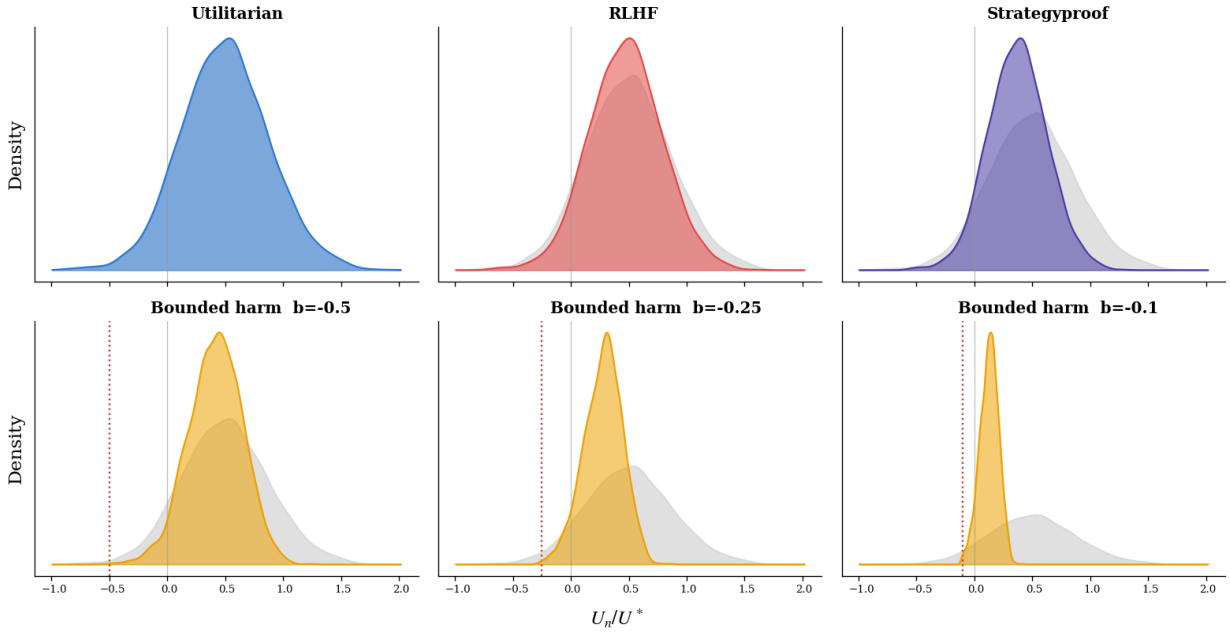


Figure 2: Community Alignment dataset: distribution of individual welfare $U_n/U^\star$ across the 2387 annotators under each mechanism, at high choice precision ($\beta = 100$). Top: utilitarian, RLHF, and strategyproof. Bottom: harm-bounded alignment at three welfare floors b; the faint gray curve repeats the utilitarian distribution for reference, and the dotted red line marks the floor. Tightening the floor (toward 0) binds more agents and compresses the welfare distribution, removing the harmed left tail.

The constraint says: the aggregate welfare of everyone except n, evaluated at the deployed impact, must not fall more than $(1 - w_n)\, b_n$ below its value at the reference (recall that $\bar{\theta}_{-n}$ normalizes the leave-one-out weights by $1/(1 - w_n)$, which requires $w_n < 1$ for all n). This bounds the cost that accommodating n's preferences imposes on the rest of the population. When $b_n = 0$, the mechanism cannot reduce the rest of the population's aggregate welfare by including n; as $b_n \to \infty$, the constraint becomes vacuous.

Corollary 4 (of Theorem 4)**.** *The externality-bounded optimum satisfies*

$$\psi_b^{\text{ext}} \in \arg\max_{\psi \in \bar{\Psi}} \left(\left(1 + \sum_{n \in \mathcal{B}} \frac{\eta_n}{1 - w_n} \right) \bar{\theta}_w \; - \; \sum_{n \in \mathcal{B}} \frac{w_n\, \eta_n}{1 - w_n}\, \theta_n \right)^\top \psi, \tag{7}$$

where $\eta_n \geq 0$ *and* $\eta_n(\bar{\theta}_{-n}^\top \psi_b^{\text{ext}} - \bar{\theta}_{-n}^\top \psi_0 + b_n) = 0$ *for all* n.

The effective direction has a striking structure that is the *opposite* of the welfare floor. In harm-bounded alignment (Corollary 2), binding agents receive *additional* weight: $w_n \to w_n + \eta_n$. Here, binding agents have their weight *reduced*: their preferences are subtracted from the effective direction, and the social welfare direction $\bar{\theta}_w$ is amplified. Intuitively, the welfare floor asks "who is being harmed?" and gives them more voice; the externality bound asks "who is causing harm?" and attenuates their influence.

With equal weights $w_n = 1/N$ and large N, the effective direction simplifies to approximately $(1 + \sum_{n \in \mathcal{B}} \eta_n)\, \bar{\theta}_w - \frac{1}{N} \sum_{n \in \mathcal{B}} \eta_n\, \theta_n$: the dominant effect is an amplification of the utilitarian direction, with a small per-agent correction of order $1/N$.

A bounded harm mechanism allows us to regulate individuals' participation externalities (Figure 4 in Appendix E.3). The bright rows in the bounded harm panel are the agents for whom the constraint was binding—agents whose welfare would otherwise fall below the floor. When such an agent is removed from the dataset, their binding constraint disappears and everyone else benefits from returning closer to the utilitarian regime; hence, their inclusion imposes a drastic negative externality on every other member of the population.

7 Empirical Illustrations

Setup. We illustrate the framework on real human pairwise-choice data from four domains: kidney allocation [Keswani et al., 2024], charitable food distribution [Lee et al., 2019], trolley problems from Moral Machine [Awad et al., 2018], and preferences over LLM responses from the Community Alignment dataset [Zhang et al., 2025]. In each domain, options are described by interpretable features (k ranging from 5 to 20) and we fit each participant's linear preference θ_n with a Bradley–Terry–Luce model on their own choices; a precision parameter β scales all preferences, with choices becoming deterministic as $\beta \to \infty$. For Community Alignment, responses are first embedded by a frozen sentence encoder and projected onto a $K = 15$-dimensional interpretable basis using the method of Wojtowicz et al. [2026], yielding $N = 2387$ annotators. Taking the observed queries as the deployment distribution, the impact zonotope $\bar{\Psi}$—and with it every mechanism in this paper—can be computed exactly. Dataset details and preprocessing are given in Appendix E.

Mechanisms and measures. We compare the deployed impacts of the utilitarian optimum (Proposition 1), anonymous RLHF (Proposition 4), the uniform-weight random dictatorship compiled into a single model parameter (Theorem 3), and harm-bounded alignment at several welfare floors (Corollary 2). For each mechanism we compute every agent's welfare U_n relative to the utilitarian optimum $U^\star$, as well as the person-on-person participation externalities $\Gamma_{n,m}$ (Appendix D.1).

Findings. On Community Alignment, welfare varies widely across agents under the utilitarian, RLHF, and strategyproof mechanisms, and under all three some agents are actively harmed (Figure 2). Under harm-bounded alignment no agent's welfare falls below the chosen floor, at a cost in mean welfare that grows as the floor tightens. The externality measures show that this cost comes from a small number of agents whose harm constraint binds, and that including these agents is costly to everyone else. Across the four datasets, agents disagree about the *direction* of their preferences in Moral Machine and Community Alignment, but mostly about the *magnitude* of their preferences in the kidney and food-rescue domains (Figure 7 in Appendix E). As a result, the strategyproof mechanism loses a fraction of the optimal welfare in these datasets (Figure 8), because the adversarial structure behind Proposition 3 does not arise in them. At high choice precision, welfare also varies least across agents under this mechanism (Figures 9 and 10). Under the utilitarian, RLHF, and strategyproof mechanisms, welfare varies more across agents as the choice precision β grows and individual choices become more deterministic (Figure 3).

8 Limitations and Discussion

Limitations. Our framework builds on, and therefore relies on, the assumption that preferences are linear in a model's feature space. For LLM alignment, making such a linear representation

interpretable is nontrival, which we demonstrate on the Community Alignment dataset by importing the feature extraction pipeline of [Wojtowicz et al., 2026]. Additionally, some assumptions in our results about query distributions may not hold in practice, and the strategyproofness results rely on the central planner broadcasting individual probabilities of being selected ahead of time. The "optimal" strategyproof solution may require solving a fixed-point problem where the optimal probabilities are learned over time, a technical problem we propose for future work.

Discussion. We have shown that linear social choice provides the right setting in which to reformulate alignment as a linear optimization problem. By analyzing welfare not as a second-order consequence of model parameterization but as a first-order object, we are able to develop an impact-space framework that directly informs how we design strategyproof mechanisms as well as mechanisms that implement a variety of welfare-specific desiderata. The information granularity resulting from the impact space analysis leads to a rich set of future questions: can we use this framework to constrain the welfare impacts of one demographic group on another? Can every alignment mechanism (traditional RLHF, DPO, Max-Min, Nash) be characterized by the patterns of externalities that it creates over the population of individuals, and do we have the power to regulate these in a unified way? Prioritizing welfare as the primary object of analysis may also lead to other elegant structures in the analysis of alignment protocols—we believe this viewpoint will only become more important as we begin to align agents and other systems capable of making decisions on behalf of humans, such as negotiation or high-stakes resource allocation.

References

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

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

Salvador Barberà. Strategyproof social choice. In Kenneth J. Arrow, Amartya Sen, and Kotaro Suzumura, editors, *Handbook of Social Choice and Welfare*, volume 2, chapter 25, pages 731–831. Elsevier, 2011.

Salvador Barbera and Bezalel Peleg. Strategy-proof voting schemes with continuous preferences. *Social choice and welfare*, 7(1):31–38, 1990.

Salvador Barberà, Faruk Gul, and Ennio Stacchetti. Generalized median voter schemes and committees. *Journal of Economic Theory*, 61(2):262–289, 1993.

Kyle Boerstler, Vijay Keswani, Lok Chan, Jana Schaich Borg, Vincent Conitzer, Hoda Heidari, and Walter Sinnott-Armstrong. On the stability of moral preferences: A problem with computational elicitation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 156–167, 2024.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-RLHF: Alignment with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

Keertana Chidambaram, Karthik Vinay Seetharaman, and Vasilis Syrgkanis. Direct preference optimization with unobserved preference heterogeneity. *arXiv preprint arXiv:2405.15065*, 2024.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. Social choice should guide AI alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*, 2024.

Jessica Dai and Eve Fleisig. Mapping social choice theory to RLHF. *arXiv preprint arXiv:2404.13038*, 2024. doi: 10.48550/arXiv.2404.13038. URL `https://arxiv.org/abs/2404.13038`.

Bhaskar Dutta, Hans Peters, and Arunava Sen. Strategy-proof probabilistic mechanisms in economies with pure public goods. *Journal of Economic Theory*, 106(2):392–416, 2002.

Bailey Flanigan, Ariel D Procaccia, and Sven Wang. Distortion under public-spirited voting. *arXiv preprint arXiv:2305.11736*, 2023.

Luise Ge, Daniel Halpern, Evi Micha, Ariel D Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. Axioms for AI alignment from human feedback. *Advances in Neural Information Processing Systems*, 37:80439–80465, 2024.

Allan Gibbard. Manipulation of schemes that mix voting with chance. *Econometrica: Journal of the Econometric Society*, pages 665–681, 1977.

Paul Gölz, Nika Haghtalab, and Kunhe Yang. Distortion of AI alignment: Does preference optimization optimize for preferences? *arXiv preprint arXiv:2505.23749*, 2025.

Daniel Halpern, Evi Micha, Ariel D Procaccia, and Itai Shapira. Pairwise calibrated rewards for pluralistic alignment. *arXiv preprint arXiv:2506.06298*, 2025.

Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: A survey. *Machine Learning*, 113(5):2351–2403, 2024.

John C Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy*, 63(4):309–321, 1955.

Vijay Keswani, Vincent Conitzer, Hoda Heidari, Jana Schaich Borg, and Walter Sinnott-Armstrong. On the pros and cons of active learning for moral preference elicitation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 711–723, 2024.

Thomas Kleine Buening, Jiarui Gan, Debmalya Mandal, and Marta Kwiatkowska. Strategyproof reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, 38:101431–101464, 2026.

Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on human-computer interaction*, 3(CSCW): 1–35, 2019.

Xinyu Li, Ruiyang Zhou, Zachary C. Lipton, and Liu Leqi. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*, 2024. doi: 10.48550/arXiv.2402.05133. URL `https://arxiv.org/abs/2402.05133`.

Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman Ozdaglar. RLHF from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1664. URL `https://proceedings.neurips.cc/paper_files/paper/2024/hash/5e1c255653eb98cef13f45b2d337c882-Abstract-Conference.html`.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Ali Shirali, Arash Nasr-Esfahany, Abdullah Alomar, Parsa Mirtaheri, Rediet Abebe, and Ariel Procaccia. Direct alignment with heterogeneous preferences. *arXiv preprint arXiv:2502.16320*, 2025.

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

Haoran Sun, Yurong Chen, Siwei Wang, Xu Chu, Wei Chen, and Xiaotie Deng. Mechanism design for LLM fine-tuning with multiple reward models. *arXiv preprint arXiv:2405.16276*, 2024.

Zachary Wojtowicz, Ayush Nayak, and Jacob Andreas. From weights to words: Expressing and editing preference model inferences in natural language. *arXiv preprint arXiv:2607.16232*, 2026.

Lily Hong Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, et al. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. *arXiv preprint arXiv:2507.09650*, 2025.

A Proofs for Sections 3 and 4

Throughout the proofs, let $\Psi = \{\psi(\theta) : \theta \in \mathbb{R}^k\}$ and $\bar{\Psi} = \mathrm{cl}(\Psi)$. We optimize over $\bar{\Psi}$, but ψ^{-1} is used only on Ψ, where the impact map is invertible on the deployment subspace.

For any non-empty closed convex set A, the support function of A is

$$h_A(x) = \sup_{a \in A} x^\top a. \tag{8}$$

Proof of Theorem 1.

1. Write the impact map in terms of its per-query coordinates

$$\psi(\theta) = \mathbb{E}_{z \sim q_{\text{dep}}}\big[t_z(\theta)\,\alpha_z\big], \qquad t_z(\theta) = \sigma(\alpha_z^\top \theta) - \tfrac{1}{2} \in (-\tfrac{1}{2}, \tfrac{1}{2}), \tag{9}$$

and let $\mathcal{Z} = \{\mathbb{E}_{z \sim q_{\text{dep}}}[t_z \alpha_z] : t_z \in [-\frac{1}{2}, \frac{1}{2}]\}$ be the origin-symmetric zonotope generated by $\{\frac{1}{2} q_{\text{dep}}(z)\alpha_z\}$, i.e. the linear image $M([-\frac{1}{2}, \frac{1}{2}]^Z)$ of the cube (one coordinate per query $z \in Z$) under $M(t) = \mathbb{E}_{z \sim q_{\text{dep}}}[t_z \alpha_z]$. Since each $\alpha_z \in S_{\text{dep}}$, both Ψ and $\mathcal{Z}$ lie in S_{dep}. Because a linear map carries the relative interior of a convex set onto the relative interior of its image, M maps the open cube onto $\operatorname{relint} \mathcal{Z}$, the interior of $\mathcal{Z}$ relative to S_{dep}. Note that this is nonempty, since the generators of $\mathcal{Z}$ span S_{dep}. Moreover, because $t(\theta)$ lies in the open cube, $\Psi \subseteq \operatorname{relint} \mathcal{Z}$. A zonotope is convex and origin-symmetric. We have $\psi(0) = 0$ since $\sigma(0) = \frac{1}{2}$, and $\psi(-\theta) = -\psi(\theta)$ since $\sigma(-x) - \frac{1}{2} = -(\sigma(x) - \frac{1}{2})$.

2. The map ψ is the gradient of the convex log-partition potential

$$\Phi(\theta) = \mathbb{E}_{z \sim q_{\text{dep}}}\big[\log(1 + e^{\alpha_z^\top \theta}) - \tfrac{1}{2}\alpha_z^\top \theta\big], \qquad \nabla\Phi = \psi, \quad \nabla^2\Phi(\theta) = \mathbb{E}_{z \sim q_{\text{dep}}}\big[v(\alpha_z^\top \theta)\,\alpha_z \alpha_z^\top\big], \tag{10}$$

with $v = \sigma' > 0$. The Hessian is positive definite on S_{dep}, i.e., if $u \in S_{\text{dep}}$ satisfies $u^\top \nabla^2\Phi(\theta)\, u = \mathbb{E}_{z \sim q_{\text{dep}}}[v(\alpha_z^\top \theta)(\alpha_z^\top u)^2] = 0$, then $\alpha_z^\top u = 0$ for every $z \in \operatorname{supp}(q_{\text{dep}})$, so $u \perp S_{\text{dep}}$, which together with $u \in S_{\text{dep}}$ forces $u = 0$. Hence Φ restricted to S_{dep} is finite, C^∞, and strictly convex. Given that it is finite and differentiable everywhere, it is essentially smooth, and is therefore a convex function of Legendre type [Rockafellar, 1970, §26]. We restrict attention to S_{dep} throughout, as on the ambient space Φ has flat directions and $\mathcal{Z}$ has empty interior whenever $S_{\text{dep}} \neq \mathbb{R}^k$, so all interiors below are relative to S_{dep}. The recession function of Φ is

$$\lim_{\nu \to \infty} \frac{\Phi(\nu u)}{\nu} = \mathbb{E}_{z \sim q_{\text{dep}}}\big[(\alpha_z^\top u)_+ - \tfrac{1}{2}\,\alpha_z^\top u\big] = \tfrac{1}{2}\,\mathbb{E}_{z \sim q_{\text{dep}}}\big[\,|\alpha_z^\top u|\,\big] = h_{\mathcal{Z}}(u), \tag{11}$$

the support function of $\mathcal{Z}$. By Rockafellar [1970, Theorem 13.3], the support function of $\operatorname{dom}\Phi^*$ is the recession function of Φ; hence $\operatorname{cl}(\operatorname{dom}\Phi^*) = \mathcal{Z}$, and so $\operatorname{relint}(\operatorname{dom}\Phi^*) = \operatorname{relint}\mathcal{Z}$. By Rockafellar [1970, Theorem 26.5], the gradient map of a Legendre-type function is a bijection from the interior of its domain onto the interior of the domain of its convex conjugate, with continuous inverse $\nabla\Phi^*$; here these interiors are S_{dep} and $\operatorname{relint}\mathcal{Z}$, so $\psi = \nabla\Phi$ is a bijection from S_{dep} onto $\operatorname{relint}\mathcal{Z}$, and, by the inverse function theorem, a C^∞ diffeomorphism. Therefore $\Psi = \operatorname{relint}\mathcal{Z}$, so $\bar{\Psi} = \operatorname{cl}(\operatorname{relint}\mathcal{Z}) = \mathcal{Z}$ is the stated zonotope.

3. From $|t_z(\theta)| \le \frac{1}{2}$, $\|\psi(\theta)\| \le \mathbb{E}[\|\alpha_z\|\,|t_z(\theta)|] \le \frac{1}{2}\mathbb{E}[\|\alpha_z\|]$.

4. Differentiating under the expectation, $\frac{\partial}{\partial\nu}\psi(\nu\theta) = \mathbb{E}[(\alpha_z^\top\theta)v(\nu\alpha_z^\top\theta)\alpha_z]$, so

$$\theta^\top \frac{\partial}{\partial\nu}\psi(\nu\theta) = \mathbb{E}_{z \sim q_{\text{dep}}}\big[(\alpha_z^\top\theta)^2 v(\nu\alpha_z^\top\theta)\big] \ge 0. \tag{12}$$

5. By dominated convergence, $\lim_{\nu\to\infty} t_z(\nu\theta) \to \frac{1}{2}\operatorname{sign}(\alpha_z^\top\theta)$, so

$$\lim_{\nu\to\infty} \psi(\nu\theta) = \tfrac{1}{2}\,\mathbb{E}_{z\sim q_{\mathrm{dep}}}\big[\alpha_z \operatorname{sign}(\alpha_z^\top\theta)\big], \tag{13}$$

which attains $\max_{\psi\in\bar\Psi}\theta^\top\psi = \frac{1}{2}\,\mathbb{E}_{z\sim q_{\mathrm{dep}}}|\alpha_z^\top\theta|$. For generic θ (i.e., $\alpha_z^\top\theta \neq 0$ for all $z \in \operatorname{supp}(q_{\mathrm{dep}})$), it is the unique maximizer: any maximizer may be written $\mathbb{E}_{z\sim q_{\mathrm{dep}}}[t_z\,\alpha_z]$ with $t_z \in [-\frac{1}{2},\frac{1}{2}]$, and $\theta^\top\psi = \mathbb{E}_{z\sim q_{\mathrm{dep}}}[t_z\,(\alpha_z^\top\theta)]$ attains this value only if $t_z = \frac{1}{2}\operatorname{sign}(\alpha_z^\top\theta)$ for every $z \in \operatorname{supp}(q_{\mathrm{dep}})$, which determines the point. As the singleton exposed face of the polytope $\bar\Psi$ in direction θ, it is a vertex; when $\theta = \theta_n$, it is agent n's ideal impact $\psi_n^\star$.

□

Proof of Proposition 1. By Theorem 1.5, the limit exists and equals $p = \frac{1}{2}\,\mathbb{E}_{z\sim q_{\mathrm{dep}}}[\alpha_z \operatorname{sign}(\alpha_z^\top\bar\theta_w)]$, and $p \in \bar\Psi$ because $\bar\Psi$ is closed and p is a limit of points of Ψ. Moreover, $\bar\theta_w^\top p = \frac{1}{2}\,\mathbb{E}_{z\sim q_{\mathrm{dep}}}[\,|\alpha_z^\top\bar\theta_w|\,] = h_{\bar\Psi}(\bar\theta_w)$, the support function of the zonotope computed in the proof of Theorem 1, so p attains the maximum of $\bar\theta_w^\top\psi$ over $\bar\Psi$, i.e. $p \in \arg\max_{\psi\in\bar\Psi}\bar\theta_w^\top\psi$. For uniqueness under genericity: any maximizer can be written $\psi = \mathbb{E}_{z\sim q_{\mathrm{dep}}}[t_z\alpha_z]$ with $t_z \in [-\frac{1}{2},\frac{1}{2}]$, and $\bar\theta_w^\top\psi = \mathbb{E}[t_z(\alpha_z^\top\bar\theta_w)]$ attains $\frac{1}{2}\mathbb{E}|\alpha_z^\top\bar\theta_w|$ only if $t_z = \frac{1}{2}\operatorname{sign}(\alpha_z^\top\bar\theta_w)$ for every z with $\alpha_z^\top\bar\theta_w \neq 0$; when $\alpha_z^\top\bar\theta_w \neq 0$ for all $z \in \operatorname{supp}(q_{\mathrm{dep}})$, this pins down every representation, so the maximizer equals p. □

B Proofs for Section 5

Proof of Lemma 1. By linearity of $\theta_n^\top(\cdot)$, $\mathbb{E}[\theta_n^\top\mu(\theta)] = \theta_n^\top\mathbb{E}_{\psi\sim\mu(\theta)}[\psi] = \theta_n^\top\bar\mu(\theta)$. Strategyproofness is an inequality between such expectations and so depends only on $\bar\mu$. For unanimity, on the generic domain where it is defined, the point $\psi^\star(\bar\theta) = \arg\max_{\psi\in\bar\Psi}\bar\theta^\top\psi$ is an exposed vertex—and hence an extreme point—of $\bar\Psi$ (Theorem 1). A distribution whose mean is an extreme point is the point mass, so $\mu(\theta,\dots,\theta) = \delta(\psi^\star(\bar\theta))$ iff $\bar\mu(\theta,\dots,\theta) = \psi^\star(\bar\theta)$. □

Proof of Lemma 2. By Lemma 1 we restrict attention to $\bar\mu$. The argument follows that of Barbera and Peleg [1990].

- ⇐ Suppose such menus M_n exist. Fix n, θ_{-n}, and true type θ_n. For any misreport θ_n', the mechanism applied at the profile (θ_n',θ_{-n}) gives $\bar\mu(\theta_n',\theta_{-n}) \in M_n(\theta_{-n})$, and applied at θ it says $\bar\mu(\theta)$ maximizes $\theta_n^\top\psi$ over $M_n(\theta_{-n})$. Hence $\theta_n^\top\bar\mu(\theta) \geq \theta_n^\top\bar\mu(\theta_n',\theta_{-n})$. Thus, it is strategyproof.
- ⇒ Suppose μ is strategyproof and take $M_n(\theta_{-n}) = O_n(\theta_{-n}) = \{\bar\mu(\theta_n',\theta_{-n}) : \theta_n' \in \mathbb{R}^k\}$, which depends only on θ_{-n}. Then $\bar\mu(\theta) \in O_n(\theta_{-n})$ trivially, and strategyproofness says $\theta_n^\top\bar\mu(\theta) \geq \theta_n^\top\psi$ for every $\psi \in O_n(\theta_{-n})$, i.e. $\bar\mu(\theta) \in \arg\max_{\psi\in O_n(\theta_{-n})}\theta_n^\top\psi$.

□

Proof of Theorem 2. Fix agent n and the others' reports. By Definition 5, agent n's expected utility is

$$\theta_n^\top\bar\mu(\theta) = \tfrac{1}{2}\,\mathbb{E}_{z\sim q_{\mathrm{dep}}}\big[f_z(s_z)\,(\alpha_z^\top\theta_n)\big], \tag{14}$$

a sum over queries in which n controls only its own votes $\{s_{z,n}\}$. For each z, this is increasing in f_z when $\alpha_z^\top\theta_n > 0$ and decreasing when $\alpha_z^\top\theta_n < 0$. Since f_z is nondecreasing in $s_{z,n}$, it is maximized

by the truthful vote $s_{z,n} = \operatorname{sign}(\alpha_z^\top \theta_n)$. The truthful report realizes all these votes simultaneously, so it maximizes every term and hence the sum.

For unanimity, note that identical reports give $s_{z,n} = s_z = \operatorname{sign}(\alpha_z^\top \bar{\theta})$, so $f_z = s_z$ and $\bar{\mu} = \frac{1}{2}\mathbb{E}_z[s_z \alpha_z] = \psi^\star(\bar{\theta})$, which is unanimous by Lemma 1. □

Proof of Proposition 2. The aggregator $f_z(s_z) = \sum_n \lambda_n s_{z,n}$ is non-decreasing in each vote (given $\lambda_n \geq 0$) with $f_z(\pm\mathbf{1}) = \pm 1$, so it defines a voting-by-issues mechanism, which is strategyproof and unanimous by Theorem 2. Its expected impact is $\frac{1}{2}\mathbb{E}_z\big[\big(\sum_n \lambda_n s_{z,n}\big)\alpha_z\big] = \sum_n \lambda_n \cdot \frac{1}{2}\mathbb{E}_z[s_{z,n}\alpha_z] = \sum_n \lambda_n \psi_n^\star$, the random dictatorship with weights λ. □

Proof of Theorem 3. Each ideal impact $\psi_n^\star$ lies in $\bar{\Psi}$, which is convex (Theorem 1), so the target impact $\bar{\psi}_\lambda = \sum_n \lambda_n \psi_n^\star \in \bar{\Psi}$. By Theorem 1(1), the impacts attainable by a single parameter are exactly $\Psi = \{\psi(\theta) : \theta \in \mathbb{R}^k\}$, the relative interior of $\bar{\Psi}$, and by Theorem 1(2), ψ is a bijection from S_{dep} onto Ψ. Hence $\bar{\psi}_\lambda$ is attained by a single parameter if and only if $\bar{\psi}_\lambda \in \Psi$, in which case the unique such parameter in S_{dep} is $\theta_\lambda^{\text{sp}} = \psi^{-1}(\bar{\psi}_\lambda)$, with impact exactly $\bar{\psi}_\lambda$. This is the mean impact of the random dictatorship (Proposition 2). Since welfare depends on a mechanism only through its expected impact (Lemma 1), this single parameter reproduces the random dictatorship's welfare for every agent.

When $\bar{\psi}_\lambda$ instead lies on the boundary of $\bar{\Psi}$, it remains a limit of single-parameter impacts: since $0 = \psi(0) \in \Psi = \operatorname{relint} \bar{\Psi}$ and $\bar{\psi}_\lambda \in \bar{\Psi}$, the half-open segment $\big\{(1-\varepsilon)\bar{\psi}_\lambda : \varepsilon \in (0,1]\big\}$ lies in Ψ. Thus $\theta_{\lambda,\varepsilon}^{\text{sp}} := \psi^{-1}\big((1-\varepsilon)\bar{\psi}_\lambda\big)$ is well-defined for every $\varepsilon \in (0,1)$, and $\psi\big(\theta_{\lambda,\varepsilon}^{\text{sp}}\big) = (1-\varepsilon)\bar{\psi}_\lambda \to \bar{\psi}_\lambda$ as $\varepsilon \to 0$. □

Proof of Proposition 3. Choice of directions. Since $\bar{\Psi}$ is convex and origin-symmetric, its affine hull is the linear span $S = \operatorname{span}(\bar{\Psi})$ and $0 \in \operatorname{relint} \bar{\Psi}$; hence $h_{\bar{\Psi}}(d) > 0$ for every nonzero $d \in S$. The support function $h_{\bar{\Psi}}$ is finite and convex on S, hence differentiable at Lebesgue-almost every direction in S, and since $\bar{\Psi} \subseteq S$ it is differentiable at $d \in S$ exactly when $\arg\max_{\psi \in \bar{\Psi}} d^\top \psi$ is a singleton. Fix such a $d_1 \in S$ with $d_1 \neq 0$ and let v denote the unique maximizer in direction d_1; then $d_1^\top v = h_{\bar{\Psi}}(d_1) > 0$ and $v \neq 0$, and by symmetry $-v$ is the unique maximizer in direction $-d_1$. Because $\dim(\bar{\Psi}) \geq 2$ and $v \in S$, we may choose $d_2 \in S$ with $d_2 \perp v$ and $d_2 \neq 0$, and then $h_{\bar{\Psi}}(d_2) > 0$.

Profile. Since w is not a point mass, fix i with $0 < w_i < 1$. For $\varepsilon > 0$, set

$$\theta_n = \begin{cases} d_1 + \varepsilon\, d_2 & n = i \\ -\dfrac{w_i}{1 - w_i}\, d_1 + \varepsilon\, d_2 & n \neq i. \end{cases} \tag{15}$$

(The coefficient for $n \neq i$ is common to all agents, so zero welfare weights are permitted.) The d_1 components cancel in the welfare-weighted average:

$$\bar{\theta}_w = \sum_n w_n\, \theta_n = w_i\, d_1 - \frac{w_i}{1 - w_i}\,(1 - w_i)\, d_1 + \varepsilon\, d_2 = \varepsilon\, d_2, \tag{16}$$

so the first-best welfare is $U^\star = \varepsilon\, h_{\bar{\Psi}}(d_2) > 0$.

Welfare of the random dictatorship. As $\varepsilon \downarrow 0$, the normalized direction of θ_i converges to that of d_1 and the normalized direction of every θ_n, $n \neq i$, converges to that of $-d_1$. The correspondence $d \mapsto \arg\max_{\psi \in \bar{\Psi}} d^\top \psi$ is upper hemicontinuous (Berge's maximum theorem, using compactness of $\bar{\Psi}$) and singleton-valued at $\pm d_1$, so *every* selection of maximizers satisfies $\psi_i^\star \to v$ and $\psi_n^\star \to -v$

for $n \neq i$; no continuity or strict-convexity assumption on $\bar{\Psi}$ is needed, and the argument covers the zonotopes of Theorem 1. Therefore

$$\psi^{\mathrm{sp}} = \sum_n \lambda_n \psi_n^\star \;\longrightarrow\; \lambda_i\, v - (1-\lambda_i)\, v = (2\lambda_i - 1)\, v \qquad (\varepsilon \downarrow 0), \tag{17}$$

and, since $d_2^\top v = 0$ by construction,

$$U^{\mathrm{sp}} = \bar{\theta}_w^\top \psi^{\mathrm{sp}} = \varepsilon\, d_2^\top \psi^{\mathrm{sp}} = \varepsilon \cdot o(1). \tag{18}$$

Combining with $U^\star = \varepsilon\, h_{\bar{\Psi}}(d_2) > 0$, the welfare ratio is $U^{\mathrm{sp}}/U^\star = o(1)$ as $\varepsilon \downarrow 0$, which is below δ once ε is small enough. The construction places no restriction on λ. □

C Anonymous RLHF Through the Impact Identity

This section collects our results on anonymous RLHF. We first prove the identity and derive its mechanism consequences: the deterministic labeling limit (Corollary 1) and manipulability at finite labeling precision (Corollary 5). We then develop the marginal welfare effect of reweighting training sources and the resulting contrast between RLHF stationarity and utilitarian efficiency (Theorem 5). Finally, we characterize the identification limits imposed by the identity's information bottleneck: which impacts reweighting can and cannot recover (Section C.5).

C.1 The Identity and Its Mechanism Consequences

Proof of Proposition 4. The anonymous RLHF objective is the cross-entropy loss

$$-\,\mathbb{E}_{(n,z)\sim q_{\mathrm{train}}}\, \mathbb{E}_{c|n,z}\Big[c \log \sigma(\alpha_z^\top \theta) + (1-c)\log\big(1-\sigma(\alpha_z^\top \theta)\big)\Big], \tag{19}$$

whose stationarity condition is

$$\mathbb{E}_{z\sim q_{\mathrm{train}}}\Big[\big(\mu_z - \sigma(\alpha_z^\top \hat{\theta}^{\mathrm{RLHF}})\big)\,\alpha_z\Big] = 0, \qquad \mu_z = \mathbb{E}_{n\sim q_{\mathrm{train}}(\cdot|z)}\big[\mathbb{E}[c \mid n, z]\big]. \tag{20}$$

Since q_{train} and q_{dep} share the same query marginal, the same moment identity $\mathbb{E}[\sigma(\alpha_z^\top \hat{\theta}^{\mathrm{RLHF}})\alpha_z] = \mathbb{E}[\mu_z \alpha_z]$ holds under q_{dep}. Subtracting $\frac{1}{2}\mathbb{E}_{z\sim q_{\mathrm{dep}}}[\alpha_z]$ from both moments,

$$\psi(\hat{\theta}^{\mathrm{RLHF}}) = \mathbb{E}_{z\sim q_{\mathrm{dep}}}\Big[\big(\sigma(\alpha_z^\top \hat{\theta}^{\mathrm{RLHF}}) - \tfrac{1}{2}\big)\alpha_z\Big] = \mathbb{E}_{z\sim q_{\mathrm{dep}}}\big[\big(\mu_z - \tfrac{1}{2}\big)\alpha_z\big]. \tag{21}$$

The argument uses only first-order stationarity and the shared query marginal. For the equivalence claim, the random-labeler protocol answers query z with choice probability μ_z, so it delivers welfare $\theta_n^\top \mathbb{E}_{z\sim q_{\mathrm{dep}}}[(\mu_z - \frac{1}{2})\alpha_z]$ to each agent n—by the display above, the same as the deployed model. Note, finally, that the right-hand side of the identity is a function of $\{\mu_z\}$ alone, so the deployed impact depends on the training data only through the per-query label frequencies. □

Proof of Corollary 1. In the deterministic labeling limit $\mathbb{E}[c \mid n, z] \to \mathbf{1}\{\alpha_z^\top \theta_n > 0\}$, so $\mu_z \to \sum_n q_{\mathrm{train}}(n \mid z)\, \mathbf{1}\{\alpha_z^\top \theta_n > 0\}$. Using $\mathbf{1}\{x > 0\} - \frac{1}{2} = \frac{1}{2}\operatorname{sign}(x)$ and $\sum_n q_{\mathrm{train}}(n \mid z) = 1$,

$$\mu_z - \tfrac{1}{2} = \tfrac{1}{2}\sum_n q_{\mathrm{train}}(n \mid z)\, \operatorname{sign}(\alpha_z^\top \theta_n) = \tfrac{1}{2}\, f_z(s_{z,\cdot}), \qquad f_z(s_{z,\cdot}) = \sum_n q_{\mathrm{train}}(n \mid z)\, s_{z,n}. \tag{22}$$

By Proposition 4 and dominated convergence ($|(\mu_z - \frac{1}{2})\alpha_z| \le \frac{1}{2}\|\alpha_z\|$), the deployed impact converges to $\frac{1}{2}\mathbb{E}_{z\sim q_{\mathrm{dep}}}[f_z(s_{z,\cdot})\alpha_z]$, the impact of the voting-by-issues mechanism of Definition 5 with aggregator f_z. Each f_z is a convex combination of the votes, hence nondecreasing in every argument with $f_z(\pm\mathbf{1}) = \pm 1$, so the mechanism is strategyproof and unanimous by Theorem 2. When $q_{\mathrm{train}}(n \mid z) = \lambda_n$ for all z, $f_z(s_{z,\cdot}) = \sum_n \lambda_n s_{z,n}$ and the impact is $\sum_n \lambda_n \psi_n^\star$, the random dictatorship with weights λ (Proposition 2). □

At finite labeling precision, the identity instead exposes a manipulation channel: the frequencies $\{\mu_z\}$, and hence the deployed impact, respond continuously to the *intensity* of an agent's reported preferences.

Corollary 5 (Finite-temperature RLHF is manipulable)**.** *Suppose, as in Corollary 1, that* q_{train} *and* q_{dep} *share the same query marginal, that the labeling weights are query-independent,* $q_{\mathrm{train}}(n \mid z) = \lambda_n$*, and that labels follow the BTL model at the* reported *preferences,* $\mathbb{E}[c \mid n, z] = \sigma(\alpha_z^\top \hat{\theta}_n)$*. Then the anonymous RLHF impact is the* λ*-average of the reported individual impacts,*

$$\psi\big(\hat{\theta}^{\mathrm{RLHF}}\big) = \sum_{n=1}^{N} \lambda_n \, \psi(\hat{\theta}_n),$$

and the induced mechanism is not strategyproof: for every agent n *with* $\lambda_n > 0$ *and generic true preference* θ_n*, the utility of reporting* $\nu\theta_n$ *is strictly increasing in* ν*, so any* $\nu > 1$ *strictly improves on truthful reporting and no optimal report exists.*

Proof. Since every $\mu_z = \sum_n \lambda_n \sigma(\alpha_z^\top \hat{\theta}_n)$ lies in $(0, 1)$, the cross-entropy objective is strictly convex and coercive on $S_{\mathrm{train}} = S_{\mathrm{dep}}$ (the supports of the query marginals coincide), so a unique stationary point exists there. By Proposition 4 and $\sum_n \lambda_n = 1$,

$$\psi\big(\hat{\theta}^{\mathrm{RLHF}}\big) = \mathbb{E}_{z\sim q_{\mathrm{dep}}}\big[(\mu_z - \tfrac{1}{2})\alpha_z\big] = \sum_{n=1}^{N} \lambda_n \, \mathbb{E}_{z\sim q_{\mathrm{dep}}}\big[\big(\sigma(\alpha_z^\top \hat{\theta}_n) - \tfrac{1}{2}\big)\alpha_z\big] = \sum_{n=1}^{N} \lambda_n \, \psi(\hat{\theta}_n).$$

Fix agent n and the others' reports. By the identity above, agent n's deployment welfare from reporting θ' is $\lambda_n \, \theta_n^\top \psi(\theta')$ plus a term that does not depend on its report, and by Theorem 1(4),

$$\frac{d}{d\nu} \, \theta_n^\top \psi(\nu\theta_n) = \mathbb{E}_{z\sim q_{\mathrm{dep}}}\big[(\alpha_z^\top \theta_n)^2 \, v(\nu \, \alpha_z^\top \theta_n)\big] > 0$$

for generic θ_n. Hence reporting $\nu\theta_n$ with $\nu > 1$ strictly improves on the truthful report. The supremum $h_{\bar{\Psi}}(\theta_n)$ of $\theta_n^\top \psi(\theta')$ over reports is approached along the ray $\nu\theta_n$ as $\nu \to \infty$ but is not attained by any report: for generic θ_n the unique maximizer $\psi_n^\star$ is an exposed vertex of $\bar{\Psi}$ and hence lies outside $\Psi = \operatorname{relint} \bar{\Psi}$. □

Kleine Buening et al. [2026] show that RLHF with stochastic labelers is not strategyproof: a labeler can manipulate its choice probabilities to steer the learned policy. Corollary 1 is complementary: in the deterministic limit, where labelers can only manipulate their choice directions (not intensities), RLHF converges to a strategyproof voting-by-issues rule. The additional manipulation channel available at finite temperature is exactly the nonlinearity that makes the finite-temperature mechanism deviate from this limit.

C.2 The Differential Companion: Marginal Alignment Externalities

The identity pins down the deployed impact at a fixed training distribution; its differential companion describes how the impact—and hence welfare—moves as the training distribution is reweighted. For an RLHF procedure that trains on a query distribution $q_{\text{train}} \in \Delta([N] \times Z)$, we define the *marginal* alignment externality: the first-order welfare effect of infinitesimally upweighting a specific training source (n, z). This is the alignment analogue of the influence function from robust statistics [see Hammoudeh and Lowd, 2024, for a review]. Here $\hat{\theta}_q$ denotes the anonymous RLHF parameter trained on distribution q; recall that $U_m(\theta) = \theta_m^\top \psi(\theta)$ is agent m's deployment welfare.

Definition 10 (Marginal Alignment Externality)**.** For a training source (n, z), let $q_\varepsilon = (1 - \varepsilon)\, q_{\text{train}} + \varepsilon\, \delta_{(n,z)}$, where $\delta_{(n,z)}$ is a point mass on (n, z). The **marginal alignment externality** of source (n, z) on agent m is

$$\gamma_{n,z}^m = \frac{d}{d\varepsilon} U_m(\hat{\theta}_{q_\varepsilon})\bigg|_{\varepsilon=0} = \theta_m^\top \frac{d\,\psi(\hat{\theta}_{q_\varepsilon})}{d\varepsilon}\bigg|_{\varepsilon=0}, \tag{23}$$

the inner product of m's preference with the **marginal impact shift** from upweighting source (n, z).

The alignment externality Γ (Appendix D.1) and the marginal externality γ are related: the former is the discrete welfare change from participation, the latter the infinitesimal welfare change from reweighting. Both are inner products of preferences with impact shifts—Γ uses the finite shift $\Delta_n = \psi - \psi_{-n}$, while γ uses the infinitesimal shift $d\psi/d\varepsilon$—but are conceptually distinct: Γ asks "what if agent n had never participated?" while γ asks "what if we gave source (n, z) slightly more weight?"

C.3 Anonymous RLHF and the Closed-Form Marginal Externality

Anonymous RLHF fits a single parameter $\hat{\theta}^{\text{RLHF}}$ to pooled data, ignoring agent identities at training time:

$$\hat{\theta}^{\text{RLHF}}_{q_{\text{train}}} \in \arg\max_{\theta \in \mathbb{R}^k} \; -\mathbb{E}_{(n,z)\sim q_{\text{train}}}\, \mathbb{E}_{c|n,z}\left[-c \log \sigma(\alpha_z^\top \theta) - (1-c) \log\left(1 - \sigma(\alpha_z^\top \theta)\right)\right] \tag{24}$$

The marginal externality has a closed form, obtained from the influence function of the estimator. Throughout the remainder of this section we assume the Bradley–Terry–Luce labeler model $\mathbb{E}[c \mid n, z] = \sigma(\alpha_z^\top \theta_n)$ and that the training queries span $\mathbb{R}^k$, so that H below is positive definite and $\hat{\theta}_q$ is the unique stationary point of the training objective, depending smoothly on q by the implicit function theorem. (Otherwise, every statement should be read as restricted to $S_{\text{train}} = \text{span}\{\alpha_z : z \in \text{supp}(q_{\text{train}})\}$, with $S_{\text{dep}} \subseteq S_{\text{train}}$ required.) The first-order condition is $g(\hat{\theta}^{\text{RLHF}}; q_{\text{train}}) = 0$ with $g(\theta; q) = \mathbb{E}_{(n',z')\sim q}[(\sigma(\alpha_{z'}^\top \theta_{n'}) - \sigma(\alpha_{z'}^\top \theta))\, \alpha_{z'}]$. Differentiating $g(\hat{\theta}_{q_\varepsilon}; q_\varepsilon) = 0$ at $\varepsilon = 0$ and using the base condition $g(\hat{\theta}^{\text{RLHF}}; q_{\text{train}}) = 0$ to cancel the distributional term gives the influence function

$$\frac{d\,\hat{\theta}_{q_\varepsilon}}{d\varepsilon}\bigg|_0 = \left(\sigma(\alpha_z^\top \theta_n) - \sigma(\alpha_z^\top \hat{\theta}^{\text{RLHF}})\right) H^{-1} \alpha_z, \qquad H = \mathbb{E}_{(n',z')\sim q_{\text{train}}}\left[v(\alpha_{z'}^\top \hat{\theta}^{\text{RLHF}})\, \alpha_{z'} \alpha_{z'}^\top\right]. \tag{25}$$

Applying the Jacobian $\nabla\psi(\hat{\theta}^{\text{RLHF}}) = \mathbb{E}_{z'\sim q_{\text{dep}}}[v(\alpha_{z'}^\top \hat{\theta}^{\text{RLHF}}) \alpha_{z'} \alpha_{z'}^\top]$ (Theorem 1) and summing over agents with the welfare weights w yields a closed form for the *social welfare effect* $\gamma_{n,z}^W = \sum_m w_m \gamma_{n,z}^m$

of source (n, z):

$$\gamma^W_{n,z} = \underbrace{\left(\sigma(\alpha_z^\top \theta_n) - \sigma(\alpha_z^\top \hat{\theta}^{\mathrm{RLHF}})\right)}_{\text{preference deviation on query } z} \quad \underbrace{\left(\alpha_z^\top \eta\right)}_{\text{welfare sensitivity of query } z} \tag{26}$$

where

$$\eta = \mathbb{E}_{(n',z')\sim q_{\mathrm{train}}}\left[v(\alpha_{z'}^\top \hat{\theta}^{\mathrm{RLHF}})\,\alpha_{z'}\,\alpha_{z'}^\top\right]^{-1} \mathbb{E}_{z\sim q_{\mathrm{dep}}}\left[(\alpha_z^\top \bar{\theta}_w)\,v(\alpha_z^\top \hat{\theta}^{\mathrm{RLHF}})\,\alpha_z\right] \tag{27}$$

is independent of (n, z) and acts as a sufficient statistic for the welfare implications of parameter changes. Anonymous RLHF distorts welfare along directions that are simultaneously (i) systematically misrepresented in the training data and (ii) socially important at deployment.

C.4 RLHF Stationarity vs. Utilitarian Efficiency

The central result of this section is that RLHF's first-order condition and utilitarian efficiency impose qualitatively different requirements on the marginal externalities. RLHF ensures only that the *social welfare effect* of marginal perturbations cancels on average across training sources; utilitarian efficiency requires that it vanishes for *every* training source individually.

Equivalently, $\gamma^W_{n,z} = \bar{\theta}_w^\top \frac{d\,\psi(\hat{\theta}_{q_\varepsilon})}{d\varepsilon}\big|_{\varepsilon=0}$, the inner product of the utilitarian preference with the marginal impact shift from upweighting source (n, z).

Theorem 5 (RLHF Stationarity vs. Utilitarian Efficiency)**.**

1. ***RLHF stationarity (average zero).*** *At the anonymous RLHF solution, the training-weighted average social welfare effect vanishes:*

$$\mathbb{E}_{(n,z)\sim q_{\mathrm{train}}}\left[\gamma^W_{n,z}\right] = 0. \tag{28}$$

   Individual sources may have $\gamma^W_{n,z} > 0$ (upweighting would improve welfare) or $\gamma^W_{n,z} < 0$, but these cancel on average under q_{train}.

2. ***Utilitarian efficiency (pointwise zero).*** *If the training weighting is* welfare-optimal*—if q_{train} maximizes the deployed utilitarian welfare $\bar{\theta}_w^\top \psi(\hat{\theta}_q)$ over all reweightings q of its sources—then the social welfare effect vanishes for every source individually:*

$$\gamma^W_{n,z} = 0 \qquad \forall\,(n, z) \in \mathrm{supp}(q_{\mathrm{train}}). \tag{29}$$

Anonymous RLHF guarantees only the average condition. Whenever it leaves some source with $\gamma^W_{n,z} \neq 0$, upweighting the sources with positive marginal effect (and downweighting those with negative effect) strictly improves deployed welfare to first order.

Proof of Theorem 5. (1). By the closed form (26), $\mathbb{E}_{(n,z)\sim q_{\mathrm{train}}}[\gamma^W_{n,z}] = \big(\mathbb{E}_{(n,z)\sim q_{\mathrm{train}}}[(\sigma(\alpha_z^\top \theta_n) - \sigma(\alpha_z^\top \hat{\theta}^{\mathrm{RLHF}}))\,\alpha_z]\big)^\top \eta = 0$, because the RLHF first-order condition makes the bracketed vector vanish.

(2). Let $W(q) = U(\hat{\theta}_q) = \bar{\theta}_w^\top \psi(\hat{\theta}_q)$ be the deployed welfare as a function of the training weighting. Since $q_\varepsilon = (1-\varepsilon) q_{\mathrm{train}} + \varepsilon \delta_{(n,z)}$, the definition of $\gamma^W_{n,z}$ is the directional derivative toward the vertex (n, z) of the simplex, $\gamma^W_{n,z} = \frac{d}{d\varepsilon} W(q_\varepsilon)\big|_0 = \partial_{n,z} W - \mathbb{E}_{q_{\mathrm{train}}}[\partial W]$. If q_{train} maximizes W over the simplex of reweightings, the first-order (KKT) condition is that the partials $\partial_{n,z} W$ are equal—to a common multiplier λ—across $\mathrm{supp}(q_{\mathrm{train}})$; their q_{train}-average is then also λ, so $\gamma^W_{n,z} = \lambda - \lambda = 0$ for every $(n, z) \in \mathrm{supp}(q_{\mathrm{train}})$. □

The welfare consequences of RLHF distortion are fully characterized by the single vector $\Delta\psi \in \mathbb{R}^k$: the aggregate welfare loss is $\bar{\theta}_w^\top \Delta\psi$, and the per-agent redistribution is $(\theta_n^\top \Delta\psi)_{n=1}^N$.

Remark 1 (Query-level decomposition). The impact gap decomposes across deployment queries as

$$\bar{\theta}_w^\top \Delta\psi = \mathbb{E}_{z\sim q_{\text{dep}}}\left[(\bar{\theta}_w^\top \alpha_z)\Big(\tfrac{1}{2}\operatorname{sign}(\alpha_z^\top \bar{\theta}_w) - \big(\sigma(\alpha_z^\top \hat{\theta}^{\text{RLHF}}) - \tfrac{1}{2}\big)\Big)\right] \tag{30}$$

The distortion on each query z is the difference between the deterministic utilitarian decision and the probabilistic RLHF decision, weighted by the welfare importance of the query.

C.5 Recoverability under Anonymous RLHF

A natural question is whether reweighting the training distribution can close the gap. By the impact identity (Proposition 4), reweighting moves the deployed impact only through the per-query label frequencies $\{\mu_z\}$, and the answer depends on whether the reweighting conditions on individual identities.

Let $C = \text{Conv}(\{\theta_n\}_{n=1}^N)$ be the convex hull of the population preferences, let $S = \text{span}\{\alpha_z : z \in \text{supp}(q_{\text{train}})\}$, and let P_S denote projection onto S.

Definition 11. A parameter $\theta \in \mathbb{R}^k$ is **individually mix-recoverable** if there exists, for each query $z \in Z$, a probability vector $\pi(\cdot \mid z) \in \Delta_N$ such that

$$\sigma(\alpha_z^\top \theta) = \sum_{n=1}^N \pi(n \mid z)\, \sigma(\alpha_z^\top \theta_n) \qquad \text{for every } z \in Z$$

Let Θ^{indiv} denote the set of all such parameters. A parameter is **anonymously mix-recoverable** if $\pi(n \mid z) = \frac{1}{N}$ for every n. Denote the set of such parameters Θ^{anon}.

Proposition 5.

1. *A parameter $\theta \in \Theta^{\text{indiv}}$ if and only if $\alpha_z^\top \theta \in \left[\min_n \alpha_z^\top \theta_n,\ \max_n \alpha_z^\top \theta_n\right]$ for every $z \in Z$.*
2. *$P_S\, C \subseteq P_S\, \Theta^{\text{indiv}}$: the convex hull of individual preferences is contained in the individually recoverable set.*

Proof of Proposition 5.

Part 1: $\Rightarrow$ Per the definition of $\theta \in \Theta^{\text{indiv}}$, there exists a π such that, for all z,

$$\sigma(\alpha_z^\top \theta) = \sum_{n=1}^N \pi(n|z)\, \sigma(\alpha_z^\top \theta_n) \tag{31}$$

As a convex combination of scalars, the right-hand side must lie between its extreme points, so that $\sigma(\alpha_z^\top \theta) \in [\min_n \sigma(\alpha_z^\top \theta_n), \max_n \sigma(\alpha_z^\top \theta_n)]$ for all z. But then, by the strict monotonicity of σ, this implies that $\alpha_z^\top \theta \in [\min_n \alpha_z^\top \theta_n, \max_n \alpha_z^\top \theta_n]$ for all z.

$\Leftarrow$ Let $z \in Z$ be arbitrary, and suppose that $\sigma(\alpha_z^\top \theta) \in [\min_n \sigma(\alpha_z^\top \theta_n), \max_n \sigma(\alpha_z^\top \theta_n)]$. Let $\underline{n} = \arg\min_n \sigma(\alpha_z^\top \theta_n)$ and $\overline{n} = \arg\max_n \sigma(\alpha_z^\top \theta_n)$. Since $\sigma(\alpha_z^\top \theta)$ lies in the closed interval with endpoints $\sigma(\alpha_z^\top \theta_{\underline{n}})$ and $\sigma(\alpha_z^\top \theta_{\overline{n}})$, there exists $t_z \in [0,1]$ such that

$$\sigma(\alpha_z^\top \theta) = (1 - t_z)\, \sigma(\alpha_z^\top \theta_{\underline{n}}) + t_z\, \sigma(\alpha_z^\top \theta_{\overline{n}}). \tag{32}$$

Set $\pi(\underline{n} \mid z) = 1 - t_z$, $\pi(\overline{n} \mid z) = t_z$, and $\pi(n \mid z) = 0$ for every other n. If $\underline{n} = \overline{n}$, the interval is a single point, and we set $\pi(\underline{n} \mid z) = 1$ instead. In either case $\pi(\cdot \mid z) \in \Delta_N$, as the definition of Θ^{indiv} requires. Applying this construction at each $z \in Z$ yields the result.

Part 2: $P_S C \subseteq P_S \Theta^{\text{indiv}}$. Fix any $\theta \in C$. By definition of C, there exists $\lambda \in \Delta_N$ such that $\theta = \sum_{n=1}^N \lambda_n \theta_n$. Then, for every $z \in Z$, $\alpha_z^\top \theta = \sum_{n=1}^N \lambda_n \alpha_z^\top \theta_n$, so $\alpha_z^\top \theta$ lies between $\min_n \alpha_z^\top \theta_n$ and $\max_n \alpha_z^\top \theta_n$. By Part 1, $\theta \in \Theta^{\text{indiv}}$, and therefore $P_S C \subseteq P_S \Theta^{\text{indiv}}$.

□

Thus, with individual-level label weights, RLHF can recover any parameter in the convex hull C of the population preferences—in particular the welfare-weighted parameter $\bar{\theta}_w$ for every $w \in \Delta_N$: if $\theta \in \Theta^{\text{indiv}}$ with mixture weights π, then setting $q_{\text{train}}(n \mid z) = \pi(n \mid z)$ makes θ satisfy the stationarity condition of Proposition 4 exactly, with zero population loss. Anonymous weighting, by contrast, is far more rigid:

Proposition 6. *If $S \neq \{0\}$, then $P_S\, \Theta^{\text{anon}}$ is either empty or a singleton.*

Proof of Proposition 6. When $\theta \in \Theta^{\text{anon}}$,

$$\alpha_z^\top \theta = \sigma^{-1}\left(\frac{1}{N}\sum_{n=1}^N \sigma(\alpha_z^\top \theta_n)\right) \qquad \forall z \in Z \tag{33}$$

where the right-hand side is determined entirely by the population $\{\theta_n\}_{n=1}^N$ for each z. If there exists no θ satisfying this system of linear constraints, then $\Theta^{\text{anon}} = \varnothing$ and hence $P_S\, \Theta^{\text{anon}}$ is empty.

Otherwise, suppose $\theta, \theta' \in \Theta^{\text{anon}}$. They both satisfy the system of linear constraints, so for every $z \in Z$, $\alpha_z^\top \theta = \alpha_z^\top \theta'$ and hence $\alpha_z^\top(\theta - \theta') = 0$ for $z \in Z$. This implies $\theta - \theta'$ is orthogonal to $\text{span}\{\alpha_z : z \in Z\} \supseteq S$, and hence to S. Thus all elements of Θ^{anon} have the same projection onto S. Therefore $P_S\, \Theta^{\text{anon}}$ is at most a singleton. □

Anonymous RLHF is pinned to at most one parameter in the identifiable subspace, regardless of how the query distribution is chosen.

Corollary 6 (Impact-space consequence)**.** *Let $\Psi^{\text{indiv}} = \{\psi(\theta) : \theta \in \Theta^{\text{indiv}}\}$ and $\Psi^{\text{anon}} = \{\psi(\theta) : \theta \in \Theta^{\text{anon}}\}$. Then $\psi(\bar{\theta}_w) \in \Psi^{\text{indiv}}$ for every $w \in \Delta_N$, but Ψ^{anon} is at most a singleton. If $\Psi^{\text{anon}} = \{\psi^{\text{anon}}\}$, the welfare gap from the anonymity constraint is $\bar{\theta}_w^\top(\psi^\star - \psi^{\text{anon}})$.*

Proof of Corollary 6. The proof of Proposition 5(2) shows $C \subseteq \Theta^{\text{indiv}}$, and $\bar{\theta}_w \in C$, so $\psi(\bar{\theta}_w) \in \Psi^{\text{indiv}}$. For the second claim, any $\theta, \theta' \in \Theta^{\text{anon}}$ satisfy (33), so $\alpha_z^\top \theta = \alpha_z^\top \theta'$ for every $z \in Z$; since ψ depends on its argument only through $(\alpha_z^\top \theta)_{z \in \text{supp}(q_{\text{dep}})}$ and $\text{supp}(q_{\text{dep}}) \subseteq Z$, the map ψ is constant on Θ^{anon}, and Ψ^{anon} is at most a singleton. □

Letting $\mu_z = \mathbb{E}_{n \sim q_{\text{train}}(n|z)}[\sigma(\alpha_z^\top \theta_n)]$ denote the population-level choice frequency for z, anonymous RLHF effectively computes the M-projection of the choice data onto the logistic model family:

$$\hat{\theta}^{RLHF}_{q_{train}} \in \arg\min_{\theta \in \mathbb{R}^k} \; \mathbb{E}_{(z,n) \sim q_{train}}\Big[\mathrm{D}_{\mathrm{KL}}\Big(\mathrm{Bern}(\mu_z) \,\Big\|\, \mathrm{Bern}(\sigma(\alpha_z^\top \theta))\Big)\Big]$$

Proposition 7. *Suppose every query is represented in training,* $\mathrm{supp}(q_{\mathrm{train}}) = Z$. *If* $\hat{\theta}^{\mathrm{RLHF}}_{q_{\mathrm{train}}}$ *achieves zero population training loss, then* $P_S\hat{\theta}^{\mathrm{RLHF}}_{q_{\mathrm{train}}} \in P_S\Theta^{\mathrm{indiv}}$.

Proof of Proposition 7. If $\hat{\theta}^{\mathrm{RLHF}}_{q_{\mathrm{train}}}$ achieves zero excess loss, then every KL term vanishes on the support of q_{train}, hence

$$\sigma(\alpha_z^\top \hat{\theta}^{\mathrm{RLHF}}_{q_{\mathrm{train}}}) = \sum_{n=1}^{N} q_{\mathrm{train}}(n \mid z)\, \sigma(\alpha_z^\top \theta_n) \qquad \forall z \in \mathrm{supp}(q_{\mathrm{train}}) \tag{34}$$

Since $\mathrm{supp}(q_{\mathrm{train}}) = Z$, this is exactly the mixture-recovery condition of Proposition 5(1) at every $z \in Z$. As $\hat{\theta}^{\mathrm{RLHF}}_{q_{\mathrm{train}}}$ is only identifiable in the subspace S, applying Proposition 5 gives $P_S\hat{\theta}^{\mathrm{RLHF}}_{q_{\mathrm{train}}} \in P_S\Theta^{\mathrm{indiv}}$. □

D Additional Material for Section 6

D.1 Alignment Externalities

Define $\Delta\psi = \psi^\star - \psi^{\mathrm{RLHF}}$ as the impact gap. Writing $U^\star = h_{\bar{\Psi}}(\bar{\theta}_w) = \bar{\theta}_w^\top \psi^\star$ for the optimal utilitarian welfare, the welfare loss from RLHF is $U^\star - \bar{\theta}_w^\top \psi^{\mathrm{RLHF}} = \bar{\theta}_w^\top \Delta\psi \;\geq\; 0$, with equality if and only if ψ^{RLHF} is itself utilitarian-optimal.

Agent n's welfare change from deploying ψ^{RLHF} instead of $\psi^\star$ is $\theta_n^\top \psi^{\mathrm{RLHF}} - \theta_n^\top \psi^\star = -\theta_n^\top \Delta\psi$. Agents whose preferences are aligned with the impact gap ($\theta_n^\top \Delta\psi > 0$) are harmed by RLHF's distortion; agents anti-aligned with $\Delta\psi$ benefit.

D.1.1 Participation Alignment Externality

The participation alignment externality measures the welfare impact of including an agent's data in the training process relative to the counterfactual in which they are excluded. This is the alignment analogue of the VCG pivot from mechanism design: it quantifies how much an agent's participation affects others.

To see this formally, let $\hat{\theta}$ denote the parameter obtained from training on the full population and $\hat{\theta}_{-n}$ the parameter obtained when agent n is excluded (and remaining agents' weights are renormalized). The corresponding deployed impacts are $\psi = \psi(\hat{\theta})$ and $\psi_{-n} = \psi(\hat{\theta}_{-n})$.

Definition 12 (Participation Alignment Externality)**.** The participation alignment externality of agent n on agent m is

$$\Gamma_{n,m} = U_m(\hat{\theta}) - U_m(\hat{\theta}_{-n}) = \theta_m^\top(\psi - \psi_{-n}) \tag{35}$$

The second equality follows from $U_m(\theta) = \theta_m^\top \psi(\theta)$. In impact space, the alignment externality is the inner product of m's preference with the impact shift $\Delta_n = \psi - \psi_{-n}$ caused by n's participation.

The impact shift $\Delta_n \in \mathbb{R}^k$ is a vector that represents the welfare consequences of n's participation: agent m benefits from n's inclusion whenever $\theta_m^\top \Delta_n > 0$ (their preferences are aligned with the impact shift) and is harmed whenever $\theta_m^\top \Delta_n < 0$. Thus, the aggregate externality of n on all other agents is

$$\Gamma_{n,-n} = \sum_{m \neq n} w_m\, \Gamma_{n,m} = (1 - w_n)\, \bar{\theta}_{-n}^\top \Delta_n \tag{36}$$

where $\bar{\theta}_{-n} = \frac{1}{1-w_n}\sum_{m\neq n} w_m\theta_m$ is the leave-one-out utilitarian preference. This is positive if the addition of n's data moves the estimated preference parameter in a direction the rest of the population likes and negative when n's participation moves the parameter in a direction the rest of the population dislikes.

Proof of Theorem 4. Define the Lagrangian

$$\mathcal{L}(\psi,\eta) = \bar{\theta}_w^\top\psi + \sum_{n=1}^{N}\eta_n(\gamma_n^\top\psi - c_n), \qquad \eta_n \geq 0. \tag{37}$$

Equivalently,

$$\mathcal{L}(\psi,\eta) = \left(\bar{\theta}_w + \sum_{n=1}^{N}\eta_n\gamma_n\right)^\top \psi - \sum_{n=1}^{N}\eta_n c_n. \tag{38}$$

Since the feasible set is non-empty and compact and the objective is continuous, an optimum exists.

By Theorem 1, $\bar{\Psi}$ is a zonotope generated by finitely many segments, hence a polytope, so the feasible set $\bar{\Psi}\cap\bigcap_n\{\gamma_n^\top\psi \geq c_n\}$ is a polyhedron and the problem is a finite linear program. Linear-programming duality therefore yields—with no constraint qualification required—multipliers $\eta_n \geq 0$ such that the constrained optimum also solves

$$\psi^{\text{wel}} \in \arg\max_{\psi\in\bar{\Psi}}\left(\bar{\theta}_w + \sum_{n=1}^{N}\eta_n\gamma_n\right)^\top \psi. \tag{39}$$

Complementary slackness gives

$$\eta_n(\gamma_n^\top\psi^{\text{wel}} - c_n) = 0 \qquad \forall n. \tag{40}$$

Hence only binding constraints enter the effective objective direction, so if $\mathcal{B} = \{n : \eta_n > 0\}$, then

$$\bar{\theta}_w + \sum_{n\in\mathcal{B}}\eta_n\gamma_n \tag{41}$$

is the effective welfare direction. Finally, let $V(c)$ denote the optimal value as a function of the threshold vector c. Raising any c_n shrinks the feasible set, so V is weakly decreasing in each c_n. For concavity, let ψ_1,ψ_2 be optimal at thresholds $c^{(1)},c^{(2)}$ and $t\in[0,1]$; then $t\psi_1+(1-t)\psi_2\in\bar{\Psi}$ by convexity and satisfies $\gamma_n^\top(t\psi_1+(1-t)\psi_2) \geq tc_n^{(1)}+(1-t)c_n^{(2)}$ for all n, so it is feasible at $tc^{(1)}+(1-t)c^{(2)}$ and $V(tc^{(1)}+(1-t)c^{(2)}) \geq tV(c^{(1)})+(1-t)V(c^{(2)})$. Whenever V is differentiable, its derivative satisfies

$$\frac{\partial V}{\partial c_n} = -\eta_n. \tag{42}$$

□

Proof of Corollary 2. Apply Theorem 4 with $\gamma_n = \theta_n$ and $c_n = b_n$. The effective welfare direction is

$$\bar{\theta}_w + \sum_{n=1}^{N}\eta_n\theta_n = \sum_{n=1}^{N}(w_n+\eta_n)\theta_n.$$

Complementary slackness gives

$$\eta_n(\theta_n^\top\psi_b^{\text{h}} - b_n) = 0 \qquad \forall n.$$

□

Proof of Corollary 3. Apply Theorem 4 with

$$\gamma_n^{\text{ps}} = \gamma\bar{\theta}_w + (1-\gamma)\theta_n \qquad \text{and} \qquad c_n = \big(\gamma\bar{\theta}_w + (1-\gamma)\theta_n\big)^\top \psi_0.$$

The effective welfare direction is

$$\bar{\theta}_w + \sum_{n=1}^{N} \eta_n\big(\gamma\bar{\theta}_w + (1-\gamma)\theta_n\big) = \left(1 + \gamma\sum_{n=1}^{N}\eta_n\right)\bar{\theta}_w + (1-\gamma)\sum_{n=1}^{N}\eta_n\theta_n.$$

By complementary slackness, only binding constraints enter this expression, giving

$$\left(1 + \gamma\sum_{n\in\mathcal{B}}\eta_n\right)\bar{\theta}_w + (1-\gamma)\sum_{n\in\mathcal{B}}\eta_n\theta_n.$$

The complementary slackness condition is

$$\eta_n\left[\big(\gamma\bar{\theta}_w + (1-\gamma)\theta_n\big)^\top(\psi_\gamma^{\text{ps}} - \psi_0)\right] = 0 \qquad \forall n.$$

□

Proof of Corollary 4. Apply Theorem 4 with $\gamma_n = \bar{\theta}_{-n}$. Substituting $\bar{\theta}_{-n} = \frac{1}{1-w_n}(\bar{\theta}_w - w_n\theta_n)$ into the effective direction:

$$\begin{aligned}\bar{\theta}_w + \sum_{n\in\mathcal{B}}\eta_n\,\bar{\theta}_{-n} &= \bar{\theta}_w + \sum_{n\in\mathcal{B}}\frac{\eta_n}{1-w_n}(\bar{\theta}_w - w_n\,\theta_n)\\ &= \left(1 + \sum_{n\in\mathcal{B}}\frac{\eta_n}{1-w_n}\right)\bar{\theta}_w \;-\; \sum_{n\in\mathcal{B}}\frac{w_n\,\eta_n}{1-w_n}\,\theta_n.\end{aligned}$$

Complementary slackness follows from the general theorem. □

The constrained welfare mechanisms introduced in Section 6 can be understood as directly regulating the alignment externality system defined in Appendix D.1.

Proposition 8 (Externality Decomposition of Welfare Floors)**.** *Let $\psi = \psi(\hat{\theta})$ be the impact deployed by the full-population training procedure and $\psi_{-n} = \psi(\hat{\theta}_{-n})$ the leave-one-out impact, as in Definition 12. Then the welfare floor constraint $\theta_m^\top\psi \geq b_m$ is equivalent to*

$$\Gamma_{n,m} \;\geq\; b_m - \theta_m^\top\psi_{-n} \qquad \forall\, n. \tag{43}$$

That is, the welfare floor constrains each agent's participation externality on m to be large enough to lift m's welfare above the floor, given m's baseline welfare without n.

Proof. From $\theta_m^\top\psi = \theta_m^\top\psi_{-n} + \Gamma_{n,m}$, the floor $\theta_m^\top\psi \geq b_m$ is equivalent to $\Gamma_{n,m} \geq b_m - \theta_m^\top\psi_{-n}$. □

Public spirit changes how binding agents pull on the solution. When γ is small, their constraints act mostly like personal welfare protections, tilting the objective toward their own preferences. When γ is large, the same constraints put more weight on the social objective itself. Thus public spirit makes the constrained solution look less like individualized compensation and more like utilitarian alignment.

E Additional Experiments

E.1 Dataset Information and Experimental Setup

All experiments were run locally on commodity hardware. An A100 was rented for conducting the linear feature embeddings for the Community Alignment dataset, which took approximately 5 minutes to run. All code can be found at `https://github.com/michelleeesi/hiddenstructure`.

E.1.1 412 Food Rescue

We use the data from Lee et al. [2019], which records pairwise choices over food-rescue recipients along $k = 7$ features (*size, access, income, poverty, last_donation, total_donation, dist*). Each row of the CSV is one binary comparison by one participant: we set $\alpha_z = \phi_A - \phi_B$ and $y_z = \mathbf{1}\{\text{chose } A\}$, with 19 participants (different stakeholders for the 412 Food Rescue program) and 45 pairwise responses each. This data is not public and was accessed with permission from the original authors on the WeBuildAI paper.

E.1.2 Kidney Exchange

We use the kidney-allocation data from Keswani et al. [2024], Boerstler et al. [2024], in which respondents choose between two patients described by $k = 5$ features (*elderlyDep, lifeYearsGained, obesity, weeklyWorkhours, yearsWaiting*). We construct α_z from the released columns as the left-option minus right-option feature difference and y_z from the `chosen` indicator. Each respondent responded to ~ 400 pairwise comparisons. The data is publicly available online at `https://github.com/vijaykeswani/Preference-Instability/tree/main/Study%201%262%20-%20AIES%202024`.

E.1.3 Moral Machine

We use the trolley-problem dataset from Awad et al. [2018], which records, for each of millions of participants, a choice between two groups of characters drawn from 20 character types (*Man, Woman, Pregnant, OldMan, Boy, Girl, Homeless, Criminal, MaleExecutive, FemaleExecutive, MaleAthlete, FemaleAthlete, MaleDoctor, FemaleDoctor, Dog, Cat,* etc.; $k = 20$). We process the public data by restricting to annotators with between 100 and 500 responses, and by screening out bots. The Moral Machine dataset is publicly available online at `https://osf.io/mxa6z/overview`.

E.1.4 Community Alignment

We use the Community Alignment Dataset of Zhang et al. [2025], which contains over 200,000 pairwise preference judgments collected from over 3000 annotators across five countries. We apply the natural language preference learning method of Wojtowicz et al. [2026]. Options are encoded by a frozen sentence encoder and pairwise differences are projected onto a $K = 15$-dimensional human-interpretable basis of preference-relevant axes of variation in the choice domain. Per-participant preferences are then estimated under a Bradley-Terry-Luce model on this subspace. We filtered for annotators with ≥ 20 responses, leaving $N = 2387$ annotators as agents.

E.2 Welfare Dispersion vs. Choice Precision

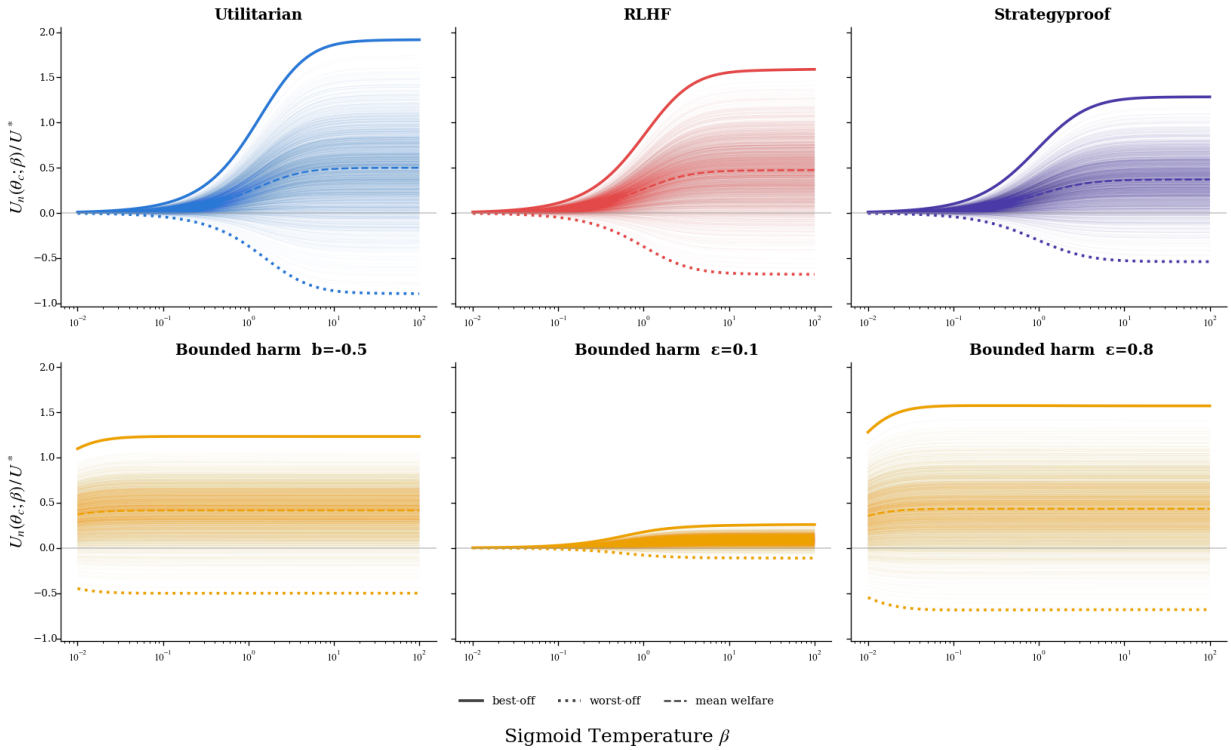


Figure 3: Community Alignment dataset: per-agent welfare $U_n/U^\star$ as a function of the choice-precision parameter β, for each mechanism. The shaded band is the density of all 2387 agents; solid, dashed, and dotted lines mark the best-off agent, mean welfare, and worst-off agent. As β grows, choices become more deterministic, individual preferences become more powerful, and the welfare distribution spreads—widening the gap between best- and worst-off agents. Strategyproof and harm-bounded mechanisms produce lower welfare dispersion, and the bounded-harm constraints are empirically satisfied.

E.3 Per-Agent Harm Bound

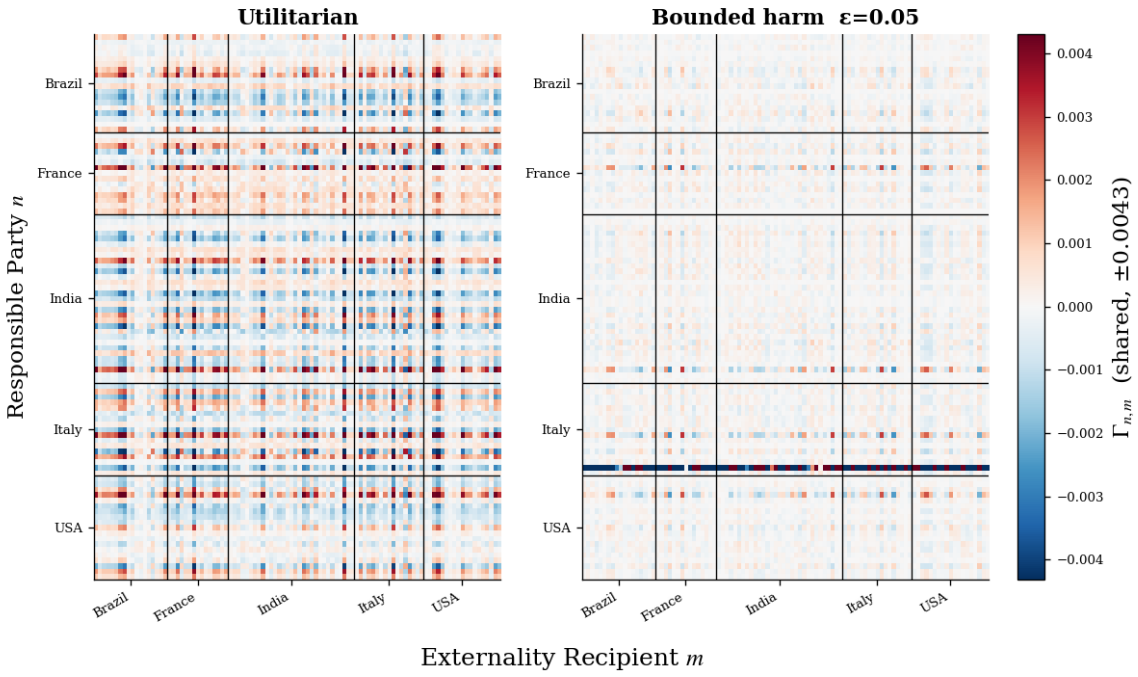


Figure 4: Participation externalities under the *per-agent* harm bound. Here each agent's welfare floor is a fixed fraction of their own maximum achievable welfare, $b_n = -\epsilon\, h_{\bar{\Psi}}(\theta_n)$ with $\epsilon = 0.05$ (lose at most 5% of one's own best case), rather than a single absolute floor. The left panel is the utilitarian baseline; the right is the harm-bounded model. The qualitative story matches that of the single absolute floor discussed in Section 6: a small set of binding agents accounts for almost all of the participation externalities.

E.4 Full Alignment Externalities

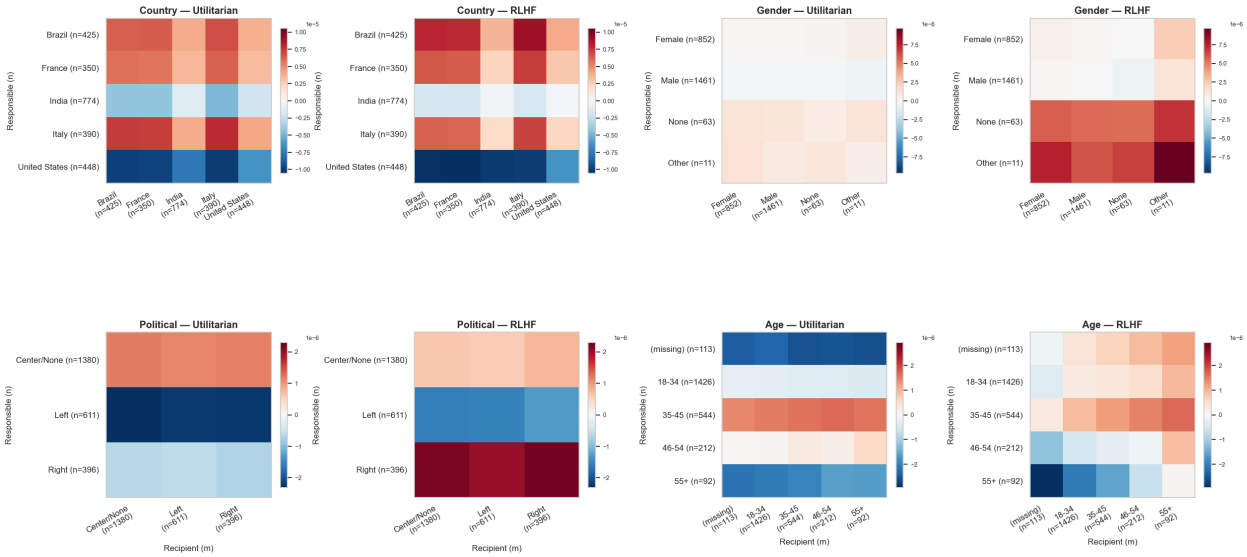
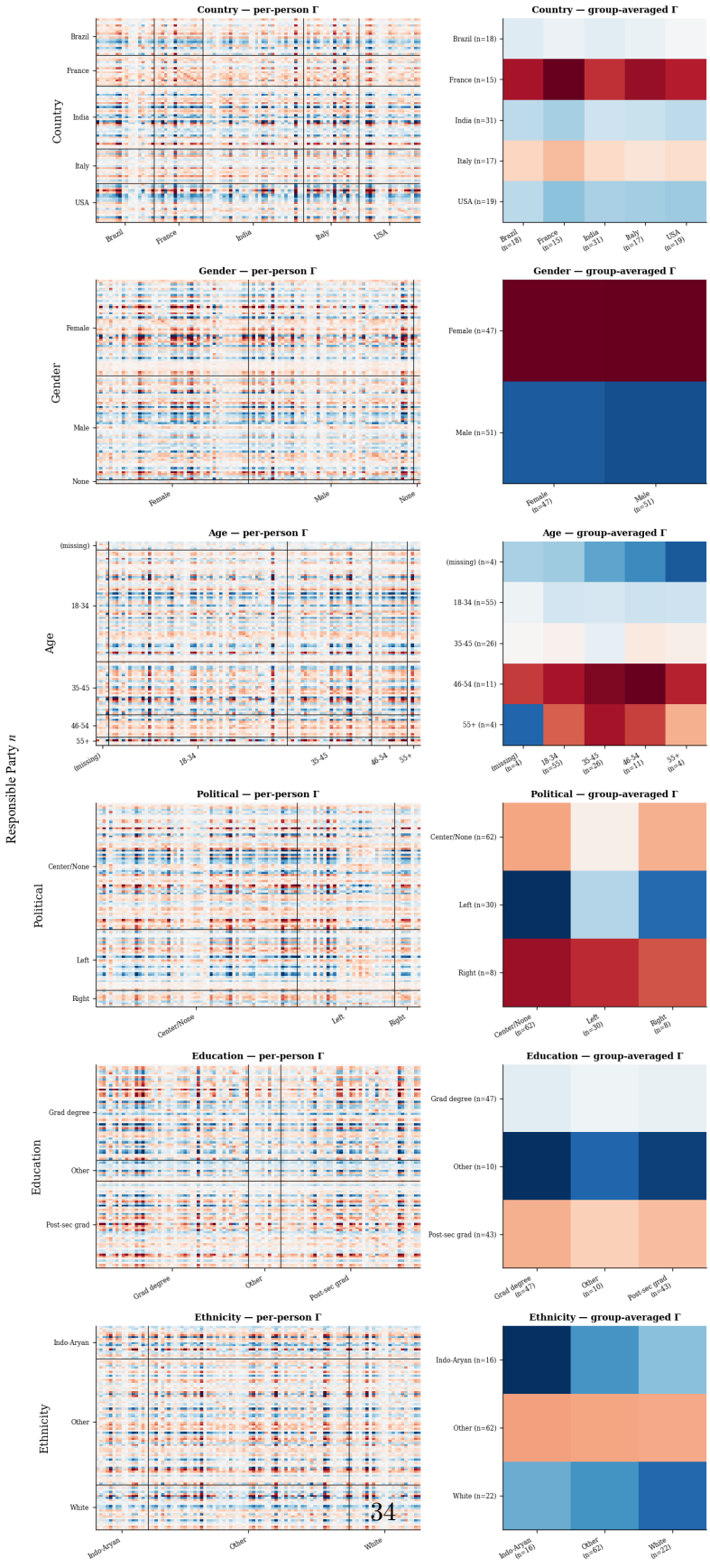


Figure 5: Community Alignment Dataset: Utilitarian vs RLHF participation alignment externalities by demographic. The figure includes 2387 annotators, with full alignment externalities averaged within each group.


Country — per-person Γ
Country — group-averaged Γ
Gender — per-person Γ
Gender — group-averaged Γ
Age — per-person Γ
Age — group-averaged Γ
Political — per-person Γ
Political — group-averaged Γ
Education — per-person Γ
Education — group-averaged Γ
Ethnicity — per-person Γ
Ethnicity — group-averaged Γ
Responsible Party n
Externality Recipient m
Brazil
France
India
Italy
USA
Brazil (n=18)
France (n=15)
India (n=31)
Italy (n=17)
USA (n=19)
Female
Male
None
Female (n=47)
Male (n=51)
(missing)
18-34
35-45
46-54
55+
(missing) (n=4)
18-34 (n=55)
35-45 (n=26)
46-54 (n=11)
55+ (n=4)
Center/None
Left
Right
Center/None (n=62)
Left (n=30)
Right (n=8)
Grad degree
Other
Post-sec grad
Grad degree (n=47)
Other (n=10)
Post-sec grad (n=43)
Indo-Aryan
Other
White
Indo-Aryan (n=16)
Other (n=62)
White (n=22)

E.5 Four Dataset Comparisons

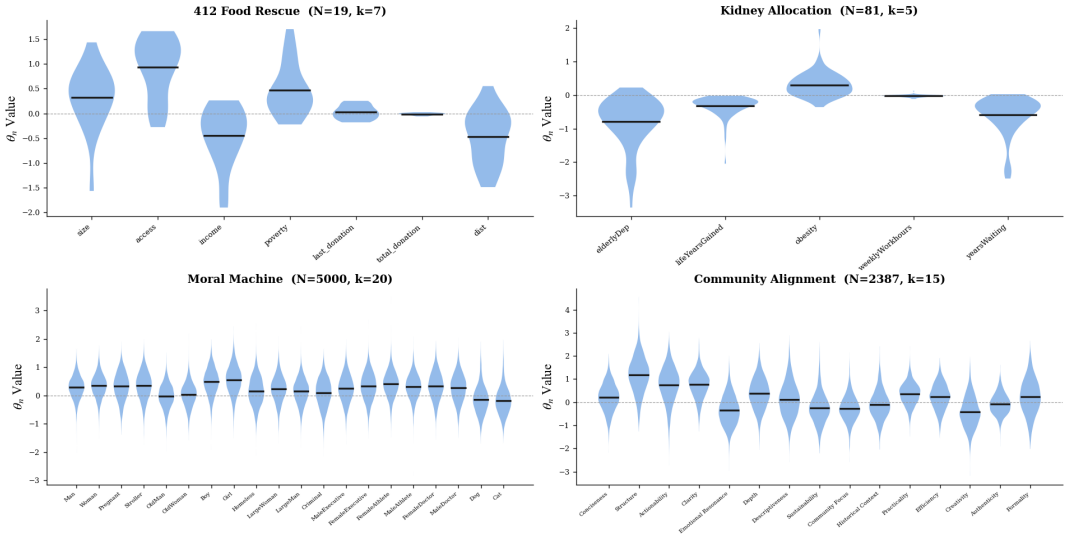


Figure 7: Preference distribution per feature dimension for four datasets. Moral Machine and Community Alignment have high disagreement, with a high number of preferences on either side of 0, while Kidney Allocation and 412 Food Rescue have less disagreement: people disagree on the magnitude of preference, but the direction is more unanimous.

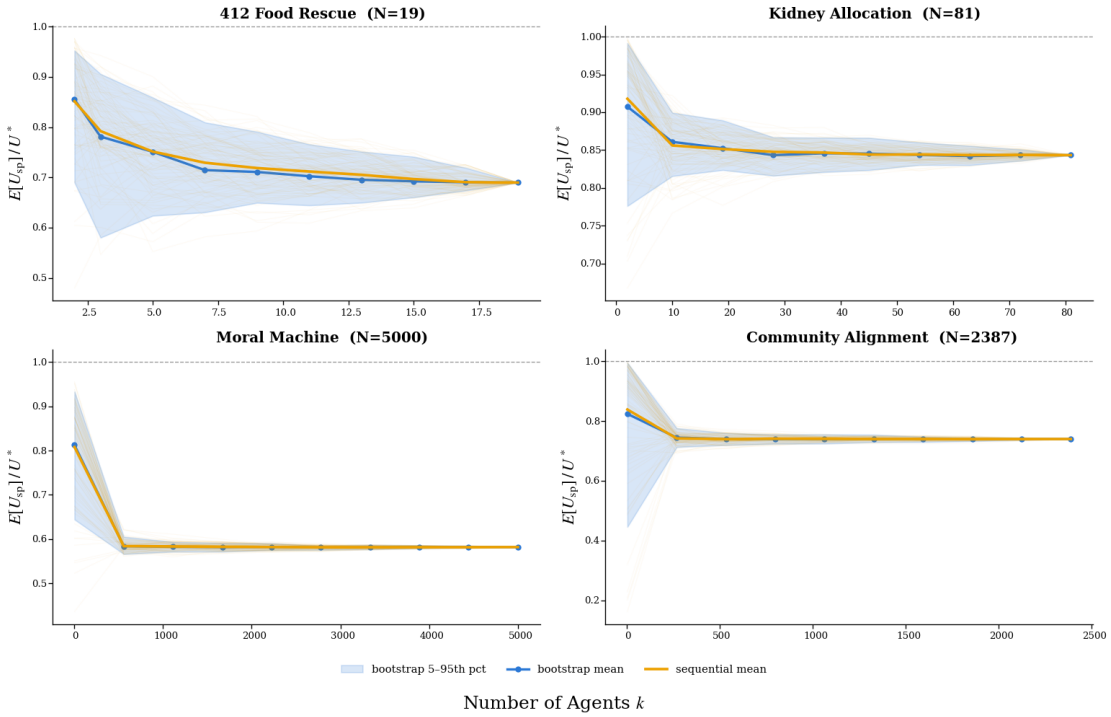


Figure 8: Expected welfare for four datasets. Bootstrap mean and std. deviations were calculated by bootstrapping l agents 100 times and calculating the welfare of the strategyproof mechanism. Sequential addition denotes the procedure where 100 permutations of agents were generated, and welfare was calculated for each coalition created by adding a new agent.

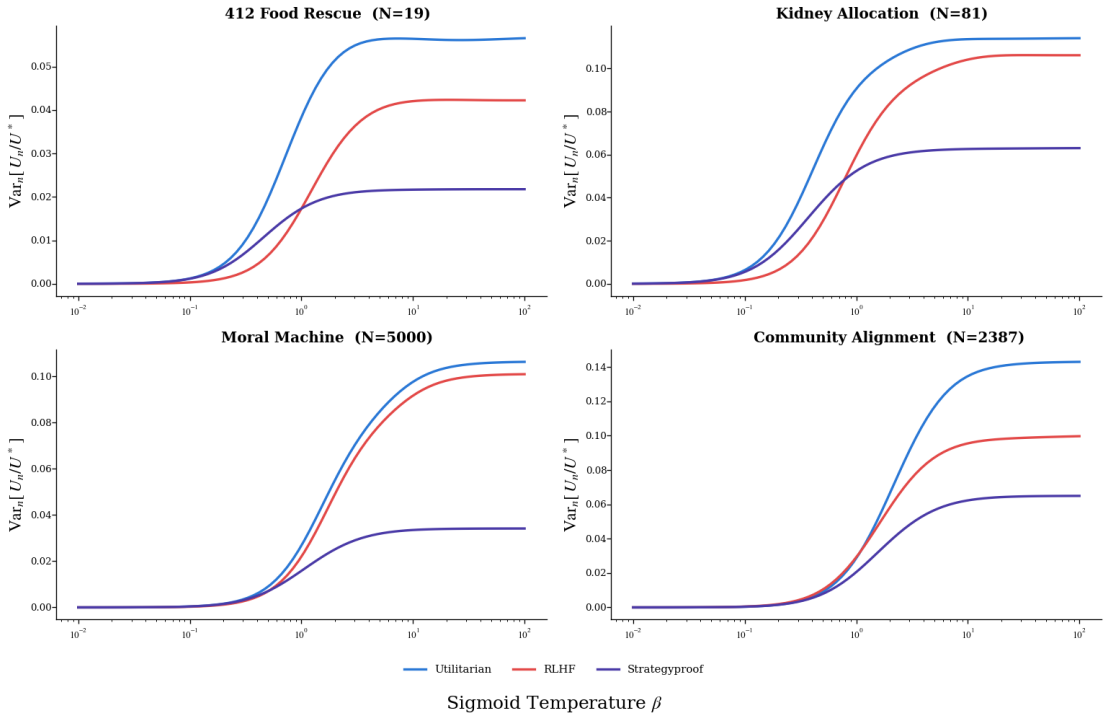


Figure 9: Welfare variance over four datasets. The strategyproof mechanism leads to the least welfare dispersion, while RLHF follows as a far second for larger values of β.

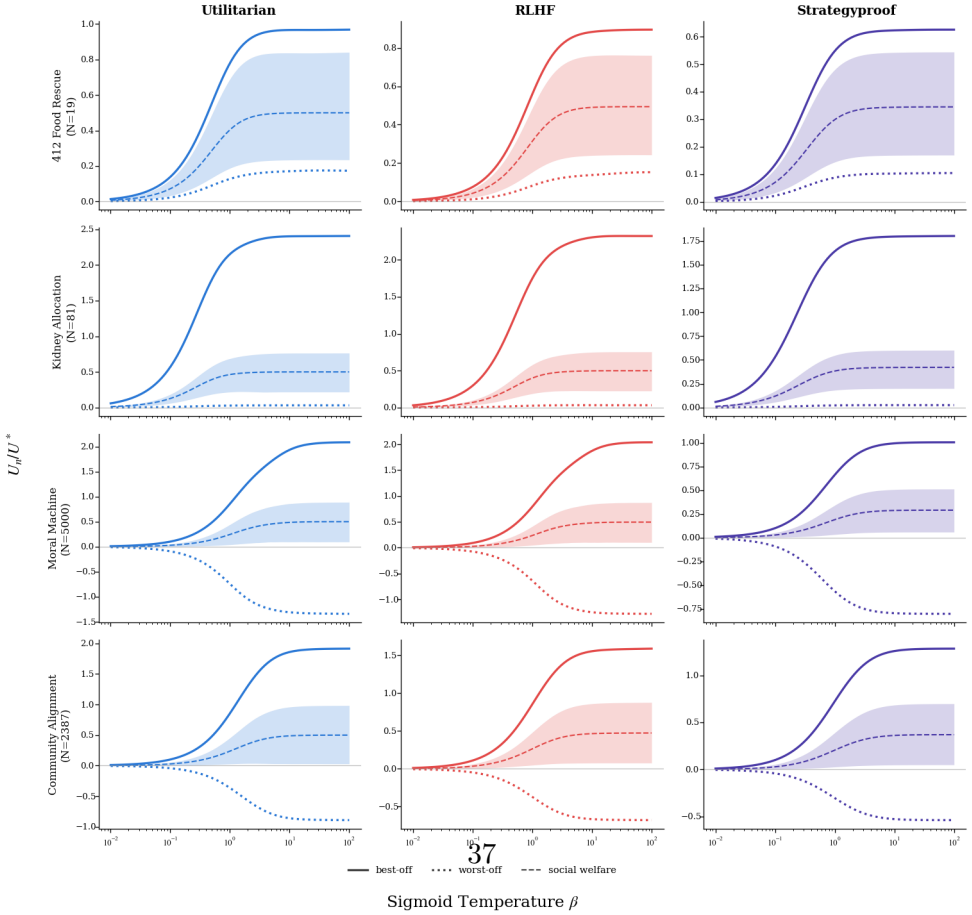


Figure 10: Welfare of the best- and worst-off agents in four datasets for three model parameters.