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, when agent utilities can be expressed linearly in a fixed feature representation, 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?

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

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

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

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, under the assumed linear utility model, working in impact space removes this parameterization complexity:

> *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 central new result, Theorem 2, connects model parameters to welfare impacts and shows how to move between them. This lets us apply existing results from *linear social choice* [Ge et al., 2024] to alignment.

Contributions: Theorem 2 in Section 4 is our main contribution, 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. In Section 5, we show that in impact space strategyproof alignment mechanisms inherit the classical menu structure of *option-set* mechanisms (Lemma 8, 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 10), 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 13). 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 17), such as bounds on harm (of which max-min protection of the worst-off agent is a special case, Corollary 19) and public-spirited trade-offs between personal loss and public gain (Corollary 35). 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]. The main difference between Kleine Buening et al. [2026] and our paper is the strategyproof object. Kleine Buening et al. [2026] study the finite-sample RLHF learning pipeline, where labelers manipulate noisy pairwise labels; they propose an approximately strategyproof learner, the Median-of-MLEs. We abstract away preference estimation and focus on the downstream social-choice problem of aggregating reported linear preferences, which lets us characterize strategyproof mechanisms directly in impact space, with random dictatorship

as the focal example. Thus, the two papers address different layers of the problem—algorithmic robustness vs. mechanism structure.

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 optimize over latent preference types given particular desiderata; we instead show how utilitarian, strategyproof, harm-bounded, and public-spirit alignment are all optimization problems over the same set, the impact space. Our work is most closely related to the linear social choice framework of Ge et al. [2024]. Theorem 2 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

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

Each context–action pair is embedded as $\phi(x, y) \in \mathbb{R}^k$. Our central assumption, standard in theoretical analyses of RLHF [Zhu et al., 2023, Nika et al., 2024, Ge et al., 2024, Kleine Buening et al., 2026, Barreto et al., 2026], 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 $\theta_n \in \mathbb{R}^k$. Linearity is in the chosen feature map ϕ, not in the raw attributes of an outcome; utility that is nonlinear in those attributes is accommodated by enriching ϕ, and everything below applies to the enriched map unchanged (Appendix E.2 tests this on our data). 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 weighted 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

A natural remedy to the challenge of social alignment is to estimate preferences at the individual or group level and then combine these estimates into a single deployed model. But this immediately

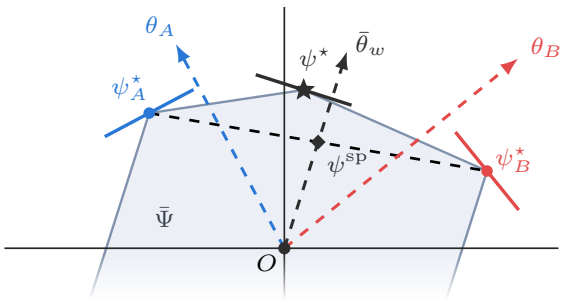


Figure 1: **The impact-space view of social alignment.** Preferences (θ_n, dashed) and the feasible impact set $\bar{\Psi}$ live in the same Euclidean space; for finitely many deployment queries, $\bar{\Psi}$ is a convex, origin-symmetric polytope. Preferences differ in magnitude and need not themselves 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 random dictatorships (Section 5), with ψ^{sp} the equal-weight random dictatorship.

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 them in model-parameter space is hard because welfare is a nonlinear function of the parameters.

The key insight of our framework is that this nonlinearity can be avoided by working directly with the set of potential welfare consequences rather than with 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 $\psi(\theta) = \mathbb{E}_{z \sim q_{\mathrm{dep}}} \left[\alpha_z \left(\sigma(\alpha_z^\top \theta) - \frac{1}{2}\right)\right] \in \mathbb{R}^k$.

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 is our main contribution: it characterizes the geometry of the impact space and connects impacts to model parameters.

Theorem 2. *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}.*

[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.

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* $v(x)=\sigma'(x)$*.*

5. *The deterministic boundary in direction* θ *is* $\lim_{\nu\to\infty}\psi(\nu\theta) = \frac{1}{2}\mathbb{E}_{z\sim q_{\mathrm{dep}}}\left[\alpha_z\,\mathrm{sign}(\alpha_z^\top\theta)\right]$*. For generic* θ *(i.e.,* $\alpha_z^\top\theta\neq 0$ *for all* $z\in\mathrm{supp}(q_{\mathrm{dep}})$*), 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 3. *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* $z\in\mathrm{supp}(q_{\mathrm{dep}})$*.*

Boundary impacts require parameter norms to grow without bound: Proposition 20 in Appendix A.1 shows that the relative welfare loss at scale ν is at most $2e^{-\nu m_w}$, where m_w is the smallest nonzero margin $|\alpha_z^\top\bar\theta_w|$, so the scale needed for a given accuracy diverges near indifference and practical norm limits can prevent exact implementation.

Appendix A.1 gives the scaling formula and its implications for individual welfare. More generally, translating the alignment problem into impact space and considering the entire problem from the perspective of welfare brings *externalities* into focus as the objects we would like to regulate as we design alignment protocols. In the remaining sections of this paper, we discuss how our framework helps us straightforwardly address these externalities.

5 Strategyproof Alignment in Impact Space

When participants can influence alignment, they may misreport their preferences to improve their own welfare. Misreporting to distort alignment in a self-serving way imposes welfare externalities on others. In economics, the field of mechanism design regulates these by identifying procedures for combining preferences that are *strategyproof*—i.e., such that no agent can benefit from misreporting.

The impact-space representation lets us apply classical mechanism design: we study option-set mechanisms [Barbera and Peleg, 1990], voting-by-issues rules, and random dictatorships. The latter can be implemented by one model parameter, exactly or in the limit, and arise from anonymous RLHF in the deterministic limit with query-independent labeling weights.

Definition 4. A **mechanism** is a function $\mu : \mathbb{R}^{N\times k} \to \Delta(\bar\Psi)$ that maps a report profile $\hat\theta = (\hat\theta_1,\dots,\hat\theta_N)$ 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 2).

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

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

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, there is no social conflict to resolve, and the mechanism should simply choose the impact vector that is best according to that shared preference.

Definition 6. 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 7. *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, holding the others' fixed.

Lemma 8. *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 identify structured subfamilies of the menu family.

5.2 Reduction 1: Voting by Deployment Query

As established by Theorem 2, 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 guarantee 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 9. 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 10. *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.

[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.

5.3 Reduction 2: Random Dictatorships

Definition 11. A mechanism is a **random dictatorship** with weights $\lambda \in \Delta_N$ if $\mu(\theta) = \sum_{n=1}^{N} \lambda_n \delta(\psi_n^\star)$. By Lemma 7, this is impact-equivalent to $\bar{\mu}(\theta) = \sum_{n=1}^{N} \lambda_n \psi_n^\star$.

Proposition 12. *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]. Linearity is a smaller domain and so admits 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 7 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. The weights λ are fixed by the designer before reports arrive, not estimated from them; since Proposition 12 holds for every $\lambda \in \Delta_N$, truthful reporting is weakly dominant even for participants who do not know λ, provided λ does not depend on their reports.

This lottery need not be implemented literally: it can be compiled into a *single* static parameter, exactly when its target impact lies in the relative interior of $\bar{\Psi}$ and as a limit otherwise.

Theorem 13. *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).

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 $\psi \in \Psi^{\mathrm{sp}} = \mathrm{conv}(\{\psi_n^\star\}_{n=1}^N)$. 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 14. *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]

Intuitively, the proof shows that this worst case arises when an individual or small group has a very strong preference along some dimension on which the overwhelming majority has a 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.[6] Proposition 21 in Appendix B.1 bounds the welfare loss of any random dictatorship by the dispersion of preferences around the utilitarian direction, and Table 3 shows the loss on our four datasets falling as the population's average majority margin rises.

[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.

[6] 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 15. *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,*

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

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 16. *Suppose, as in Proposition 15, 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 connects RLHF to the strategyproof voting-by-issues family, with random dictatorship when labeling weights are query-independent. Corollary 23 bounds the gap at finite labeling precision.

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:

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

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 welfare floors below and two further families, public spirit and participation-externality bounds, in Appendix D.

Theorem 17. *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*

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

with complementary slackness $\eta_n(\gamma_n^\top \psi^{\mathrm{wel}} - c_n) = 0$ 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.

We now go through one 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 $c_n = b_n$.

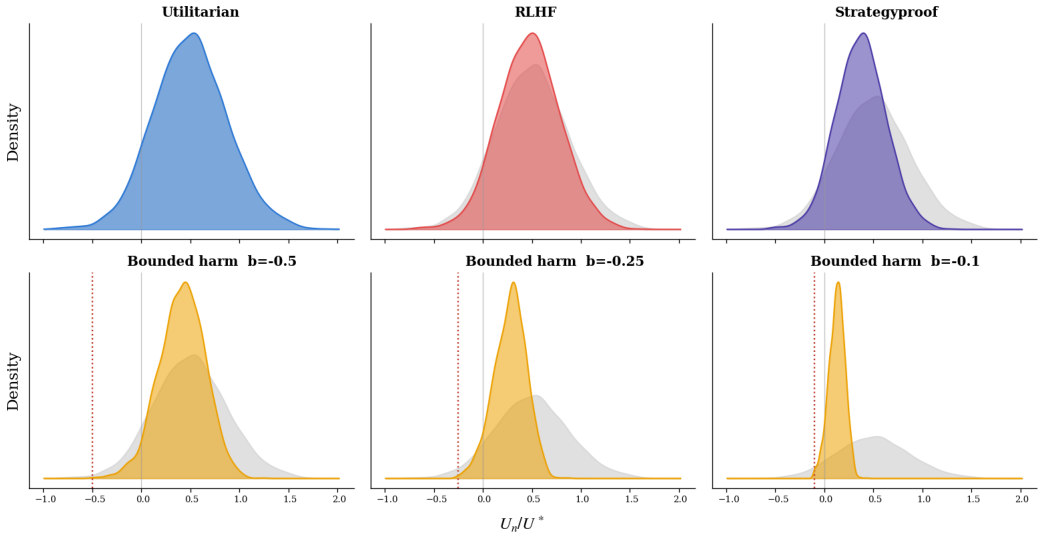


Figure 2: Community Alignment dataset: distribution of individual welfare $U_n/U^\star$ across the 2387 annotators under each mechanism, at high choice precision ($\beta = 100$). Top: utilitarian, RLHF, and strategyproof. Bottom: harm-bounded alignment at three welfare floors b; the faint gray curve 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.

Definition 18 (Harm-bounded alignment)**.** For welfare floors $b \in \mathbb{R}^N$,

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

Corollary 19 (of Theorem 17)**.** *The harm-bounded optimum satisfies*

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

where $\eta_n \geq 0$ *and* $\eta_n(\theta_n^\top \psi_b^{\mathrm{h}} - b_n) = 0$ *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) $b_n = \theta_n^\top \psi_w^{\mathrm{sp}}$, so every agent weakly prefers the mechanism to the welfare-weighted strategyproof solution; (iii) $b_n = -\epsilon h_{\bar{\Psi}}(\theta_n)$, so agent n can lose at most an ϵ fraction of their maximum achievable welfare; and (iv) $b_n = b^\star = \max_{\psi \in \bar{\Psi}} \min_m \theta_m^\top \psi$, the max-min value, which confines the mechanism to the max-min-optimal impacts and selects among them one that is Pareto efficient for strictly positive w, since any Pareto improvement would keep every agent above $b^\star$ and raise the objective.[7]

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

[7]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.

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 (Qwen3-Embedding-8B) 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 3), anonymous RLHF (Proposition 15), the uniform-weight random dictatorship compiled into a single model parameter (Theorem 13), and harm-bounded alignment at several welfare floors (Corollary 19). 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 only a limited fraction of the optimal welfare in these datasets (Figure 8), because the adversarial structure behind Proposition 14 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 fixed feature map ϕ, and the quality of that feature map limits what our guarantees mean. If the representation of the model being aligned is held fixed and only a linear head is trained on top of it, our guarantees apply exactly to the induced model class. If alignment training also changes the representation, or if the chosen features omit aspects of people's preferences, then the guarantees apply only to that approximation. Measuring this representation error is an important empirical step that we leave to future work. The 15-dimensional interpretable basis of our Community Alignment demonstration is a convenience for exposition, not a requirement of the method, and we do not claim that fifteen dimensions suffice in practice. Additionally, our assumption that the query distribution is known, or at least can be approximated, may not hold in practice.

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 mechanisms that satisfy various desiderata, such as strategyproofness. The designer's degrees of freedom this framework exposes—the weights λ, the floors b_n, and the externality tolerances—concentrate power in whoever sets them, but such choices are not new: anonymous RLHF already fixes λ implicitly through who enters the labeling pool, and it already distributes harm across people. Impact space makes the welfare consequences of these choices explicit and contestable, but who should make them is a governance question that we do not address. 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?

Acknowledgments and Disclosure of Funding

This work was partially supported by the National Science Foundation under grant IIS-2229881; by the Office of Naval Research under grants N00014-24-1-2704 and N00014-25-1-2153; and by grants from the Cooperative AI Foundation and the Foresight Institute. Michelle Si was supported by an NSF GRFP.

AI Disclosure. ChatGPT and Gemini were used in the course of brainstorming and exploring related work, and Claude Code was used in implementing the algorithms and in editing text. We reviewed and edited the content produced by these tools and are ultimately responsible for the publication's content.

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.

André Barreto, Vincent Dumoulin, Yiran Mao, Mark Rowland, Nicolas Perez-Nieves, Bobak Shahriari, Yann Dauphin, Doina Precup, and Hugo Larochelle. Capturing individual human preferences with reward features. *Advances in Neural Information Processing Systems*, 38:82119–82154, 2026.

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.

Andi Nika, Debmalya Mandal, Parameswaran Kamalaruban, Georgios Tzannetos, Goran Radanović, and Adish Singla. Reward model learning vs. direct policy optimization: A comparative analysis of learning from human preferences. *arXiv preprint arXiv:2403.01857*, 2024.

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.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR, 2023.

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

$$h_A(x) = \sup_{a \in A} x^\top a. \tag{5}$$

Proof of Theorem 2.

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

$$\psi(\theta) = \mathbb{E}_{z \sim q_{\mathrm{dep}}}\big[t_z(\theta)\,\alpha_z\big], \qquad t_z(\theta) = \sigma(\alpha_z^\top \theta) - \tfrac{1}{2} \in (-\tfrac{1}{2}, \tfrac{1}{2}), \tag{6}$$

and let $\mathcal{Z} = \{\mathbb{E}_{z \sim q_{\mathrm{dep}}}[t_z \alpha_z] : t_z \in [-\frac{1}{2}, \frac{1}{2}]\}$ be the origin-symmetric zonotope generated by $\{\frac{1}{2} q_{\mathrm{dep}}(z)\alpha_z\}$, i.e. the linear image $M([-\frac{1}{2}, \frac{1}{2}]^Z)$ of the cube (one coordinate per query $z \in Z$) under $M(t) = \mathbb{E}_{z \sim q_{\mathrm{dep}}}[t_z \alpha_z]$. Since each $\alpha_z \in S_{\mathrm{dep}}$, 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 $\mathrm{relint}\,\mathcal{Z}$, 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, $\Psi \subseteq \mathrm{relint}\,\mathcal{Z}$. 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

$$\Phi(\theta) = \mathbb{E}_{z \sim q_{\mathrm{dep}}}\big[\log(1+e^{\alpha_z^\top \theta}) - \tfrac{1}{2}\alpha_z^\top \theta\big], \qquad \nabla\Phi = \psi, \quad \nabla^2\Phi(\theta) = \mathbb{E}_{z \sim q_{\mathrm{dep}}}\big[v(\alpha_z^\top \theta)\,\alpha_z \alpha_z^\top\big], \tag{7}$$

with $v = \sigma' > 0$. The Hessian is positive definite on S_{dep}, i.e., if $u \in S_{\mathrm{dep}}$ satisfies $u^\top \nabla^2\Phi(\theta)\, u = \mathbb{E}_{z \sim q_{\mathrm{dep}}}[v(\alpha_z^\top \theta)(\alpha_z^\top u)^2] = 0$, then $\alpha_z^\top u = 0$ for every $z \in \mathrm{supp}(q_{\mathrm{dep}})$, so $u \perp S_{\mathrm{dep}}$, which together with $u \in S_{\mathrm{dep}}$ 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 $S_{\mathrm{dep}} \neq \mathbb{R}^k$, so all interiors below are relative to S_{dep}. The recession function of Φ is

$$\lim_{\nu \to \infty} \frac{\Phi(\nu u)}{\nu} = \mathbb{E}_{z \sim q_{\mathrm{dep}}}\big[(\alpha_z^\top u)_+ - \tfrac{1}{2}\,\alpha_z^\top u\big] = \tfrac{1}{2}\,\mathbb{E}_{z \sim q_{\mathrm{dep}}}\big[\,|\alpha_z^\top u|\,\big] = h_{\mathcal{Z}}(u), \tag{8}$$

the support function of $\mathcal{Z}$. By Rockafellar [1970, Theorem 13.3], the support function of $\mathrm{dom}\,\Phi^*$ is the recession function of Φ; hence $\mathrm{cl}(\mathrm{dom}\,\Phi^*) = \mathcal{Z}$, and so $\mathrm{relint}(\mathrm{dom}\,\Phi^*) = \mathrm{relint}\,\mathcal{Z}$. 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 $\mathrm{relint}\,\mathcal{Z}$, so $\psi = \nabla\Phi$ is a bijection from S_{dep} onto $\mathrm{relint}\,\mathcal{Z}$, and, by the inverse function theorem, a C^∞ diffeomorphism. Therefore $\Psi = \mathrm{relint}\,\mathcal{Z}$, so $\bar{\Psi} = \mathrm{cl}(\mathrm{relint}\,\mathcal{Z}) = \mathcal{Z}$ 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

$$\theta^\top \frac{\partial}{\partial \nu}\psi(\nu\theta) = \mathbb{E}_{z \sim q_{\mathrm{dep}}}\big[(\alpha_z^\top \theta)^2 v(\nu\alpha_z^\top \theta)\big] \ge 0. \tag{9}$$

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

$$\lim_{\nu \to \infty} \psi(\nu\theta) = \tfrac{1}{2}\,\mathbb{E}_{z \sim q_{\mathrm{dep}}}\big[\alpha_z\,\mathrm{sign}(\alpha_z^\top \theta)\big], \tag{10}$$

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 $z \in \mathrm{supp}(q_{\mathrm{dep}})$), 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 $\theta^\top \psi = \mathbb{E}_{z\sim q_{\text{dep}}}[t_z\,(\alpha_z^\top \theta)]$ attains this value only if $t_z = \frac{1}{2}\operatorname{sign}(\alpha_z^\top \theta)$ for every $z \in \operatorname{supp}(q_{\text{dep}})$, 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 3. By Theorem 2(5), the limit exists and equals $p = \frac{1}{2}\,\mathbb{E}_{z\sim q_{\text{dep}}}[\alpha_z \operatorname{sign}(\alpha_z^\top \bar{\theta}_w)]$, and $p \in \bar{\Psi}$ because $\bar{\Psi}$ is closed and p is a limit of points of Ψ. Moreover, $\bar{\theta}_w^\top p = \frac{1}{2}\,\mathbb{E}_{z\sim q_{\text{dep}}}[\,|\alpha_z^\top \bar{\theta}_w|\,] = h_{\bar{\Psi}}(\bar{\theta}_w)$, the support function of the zonotope computed in the proof of Theorem 2, 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 $\psi = \mathbb{E}_{z\sim q_{\text{dep}}}[t_z \alpha_z]$ 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 $z \in \operatorname{supp}(q_{\text{dep}})$, this pins down every representation, so the maximizer equals p. □

A.1 Scaling and 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.

Proposition 20. *Let $u_z = \alpha_z^\top \bar{\theta}_w$, let $U^\star = \sup_\theta U(\theta) = \frac{1}{2}\,\mathbb{E}_{z\sim q_{\text{dep}}}\big[|u_z|\big]$, and let $m_w = \min\{|u_z| : z \in \operatorname{supp}(q_{\text{dep}}),\, u_z \neq 0\}$. For every $\nu > 0$,*

$$U^\star - U(\nu\bar{\theta}_w) = \mathbb{E}_{z\sim q_{\text{dep}}}\big[|u_z|\,\sigma(-\nu|u_z|)\big] \le 2e^{-\nu m_w}\,U^\star. \tag{11}$$

Consequently, under a norm constraint $\|\theta\| \le r$, the parameter $(r/\|\bar{\theta}_w\|)\,\bar{\theta}_w$ has relative welfare loss at most ϵ whenever $r \ge (\|\bar{\theta}_w\|/m_w)\log(2/\epsilon)$.

Proof. Since $|\sigma(t) - \frac{1}{2}| \le \frac{1}{2}$, $U(\theta) = \mathbb{E}_z[u_z(\sigma(\alpha_z^\top\theta) - \frac{1}{2})] \le \frac{1}{2}\mathbb{E}_z\big[|u_z|\big]$ for every θ, and Proposition 3 shows the bound is attained in the limit, so $U^\star = \frac{1}{2}\mathbb{E}_z\big[|u_z|\big]$. For $u_z > 0$, $\frac{1}{2}u_z - u_z(\sigma(\nu u_z) - \frac{1}{2}) = u_z\,\sigma(-\nu u_z)$; for $u_z < 0$ the same algebra gives $|u_z|\,\sigma(-\nu|u_z|)$; for $u_z = 0$ the term vanishes. Taking expectations gives the identity. Since $\sigma(-x) \le e^{-x}$ for $x \ge 0$ and $|u_z| \ge m_w$ whenever $u_z \neq 0$, the gap is at most $e^{-\nu m_w}\mathbb{E}_z\big[|u_z|\big] = 2e^{-\nu m_w}U^\star$. The last claim follows by setting $\nu = r/\|\bar{\theta}_w\|$ and solving $2e^{-\nu m_w} \le \epsilon$. □

B Proofs for Section 5

Proof of Lemma 7. 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 2). 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 8. By Lemma 7 we restrict attention to $\bar{\mu}$. The argument follows that of Barbera and Peleg [1990].

- $\Leftarrow$ 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 $\theta_n^\top \bar{\mu}(\theta) \ge \theta_n^\top \bar{\mu}(\theta_n', \theta_{-n})$. Thus, it is strategyproof.
- $\Rightarrow$ 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 $\theta_n^\top \bar{\mu}(\theta) \ge \theta_n^\top \psi$ 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 10. Fix agent n and the others' reports. By Definition 9, agent n's expected utility is

$$\theta_n^\top \bar{\mu}(\theta) = \tfrac{1}{2}\,\mathbb{E}_{z\sim q_{\mathrm{dep}}}\big[f_z(s_z)\,(\alpha_z^\top \theta_n)\big], \tag{12}$$

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

For unanimity, note that identical reports give $s_{z,n} = s_z = \mathrm{sign}(\alpha_z^\top \bar{\theta})$, 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 7. □

Proof of Proposition 12. The aggregator $f_z(s_z) = \sum_n \lambda_n s_{z,n}$ is nondecreasing 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 10. 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 13. Each ideal impact $\psi_n^\star$ lies in $\bar{\Psi}$, which is convex (Theorem 2), so the target impact $\bar{\psi}_\lambda = \sum_n \lambda_n \psi_n^\star \in \bar{\Psi}$. By Theorem 2(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 2(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 $\theta_\lambda^{\mathrm{sp}} = \psi^{-1}(\bar{\psi}_\lambda)$, with impact exactly $\bar{\psi}_\lambda$. This is the mean impact of the random dictatorship (Proposition 12). Since welfare depends on a mechanism only through its expected impact (Lemma 7), 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 $0 = \psi(0) \in \Psi = \mathrm{relint}\,\bar{\Psi}$ 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 $\theta_{\lambda,\varepsilon}^{\mathrm{sp}} := \psi^{-1}\big((1-\varepsilon)\bar{\psi}_\lambda\big)$ is well-defined for every $\varepsilon \in (0,1)$, and $\psi\big(\theta_{\lambda,\varepsilon}^{\mathrm{sp}}\big) = (1-\varepsilon)\bar{\psi}_\lambda \to \bar{\psi}_\lambda$ as $\varepsilon \to 0$. □

Proof of Proposition 14. Choice of directions. Since $\bar{\Psi}$ is convex and origin-symmetric, its affine hull is the linear span $S = \mathrm{span}(\bar{\Psi})$ and $0 \in \mathrm{relint}\,\bar{\Psi}$; 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

$$\theta_n = \begin{cases} d_1 + \varepsilon\, d_2 & n = i \\ -\dfrac{w_i}{1-w_i}\, d_1 + \varepsilon\, d_2 & n \neq i. \end{cases} \tag{13}$$

(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:

$$\bar{\theta}_w = \sum_n w_n\, \theta_n = w_i\, d_1 - \frac{w_i}{1-w_i}\,(1-w_i)\, d_1 + \varepsilon\, d_2 = \varepsilon\, d_2, \tag{14}$$

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 2. Therefore

$$\psi^{\mathrm{sp}} = \sum_n \lambda_n \psi_n^\star \;\longrightarrow\; \lambda_i v - (1-\lambda_i) v = (2\lambda_i - 1) v \qquad (\varepsilon \downarrow 0), \tag{15}$$

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

$$U^{\mathrm{sp}} = \bar{\theta}_w^\top \psi^{\mathrm{sp}} = \varepsilon\, d_2^\top \psi^{\mathrm{sp}} = \varepsilon \cdot o(1). \tag{16}$$

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. □

B.1 The Cost of Strategyproofness

Proposition 14 shows that the welfare loss of a random dictatorship can be arbitrarily large in the worst case. The following bound shows that the loss is controlled by the dispersion of preferences around the utilitarian direction, so strategyproofness is cheap when preferences are concentrated and expensive when they are dispersed.

Proposition 21. *For any $\lambda \in \Delta_N$, writing $U^\star = \bar{\theta}_w^\top \psi^\star$,*

$$U^\star - \mathbb{E}\big[U(\mu_\lambda^{\mathrm{sp}}(\theta))\big] \leq \operatorname{diam}(\bar{\Psi}) \sum_{n=1}^N \lambda_n \|\theta_n - \bar{\theta}_w\| \leq \mathbb{E}_{z\sim q_{\mathrm{dep}}}\big[\|\alpha_z\|\big] \sum_{n=1}^N \lambda_n \|\theta_n - \bar{\theta}_w\|.$$

Proof. By Lemma 7, $\mathbb{E}[U(\mu_\lambda^{\mathrm{sp}}(\theta))] = \sum_n \lambda_n \bar{\theta}_w^\top \psi_n^\star$. Since $\psi_n^\star$ maximizes $\theta_n^\top \psi$ over $\bar{\Psi}$, $\theta_n^\top(\psi^\star - \psi_n^\star) \leq 0$, hence

$$\bar{\theta}_w^\top(\psi^\star - \psi_n^\star) = (\bar{\theta}_w - \theta_n)^\top(\psi^\star - \psi_n^\star) + \theta_n^\top(\psi^\star - \psi_n^\star) \leq \|\bar{\theta}_w - \theta_n\| \operatorname{diam}(\bar{\Psi}).$$

Averaging over λ gives the first inequality; the second uses $\operatorname{diam}(\bar{\Psi}) \leq 2 \sup_{\psi\in\bar{\Psi}} \|\psi\| \leq \mathbb{E}_{z\sim q_{\mathrm{dep}}}\big[\|\alpha_z\|\big]$ (Theorem 2(3)). □

To measure agreement over deployment choices empirically, we use the average majority margin $M = \mathbb{E}_{z\sim q_{\mathrm{dep}}}\big|\frac{1}{N}\sum_n \operatorname{sign}(\alpha_z^\top \theta_n)\big|$, the expected gap between the shares of agents favoring the majority and minority options, which is 1 under unanimity and 0 under an even split. Table 3 in Appendix E reports the welfare loss of the uniform random dictatorship against M on our four datasets; the loss falls as M rises.

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 16) and manipulability at finite labeling precision (Corollary 22). We then develop the marginal welfare effect of reweighting training sources and the resulting contrast between RLHF stationarity and utilitarian efficiency (Theorem 25). 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 15. The anonymous RLHF objective is the cross-entropy loss

$$-\mathbb{E}_{(n,z)\sim q_{\mathrm{train}}}\, \mathbb{E}_{c|n,z}\big[c \log \sigma(\alpha_z^\top \theta) + (1-c)\log\big(1-\sigma(\alpha_z^\top \theta)\big)\big]\,, \tag{17}$$

whose stationarity condition is

$$\mathbb{E}_{z\sim q_{\mathrm{train}}}\Big[\big(\mu_z - \sigma(\alpha_z^\top \hat{\theta}^{\mathrm{RLHF}})\big)\,\alpha_z\Big] = 0, \qquad \mu_z = \mathbb{E}_{n\sim q_{\mathrm{train}}(\cdot|z)}\big[\mathbb{E}[c \mid n, z]\big]. \tag{18}$$

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,

$$\psi(\hat{\theta}^{\mathrm{RLHF}}) = \mathbb{E}_{z\sim q_{\mathrm{dep}}}\Big[\big(\sigma(\alpha_z^\top \hat{\theta}^{\mathrm{RLHF}}) - \tfrac{1}{2}\big)\alpha_z\Big] = \mathbb{E}_{z\sim q_{\mathrm{dep}}}\big[\big(\mu_z - \tfrac{1}{2}\big)\alpha_z\big]\,. \tag{19}$$

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 16. 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 $\sum_n q_{\mathrm{train}}(n \mid z) = 1$,

$$\mu_z - \tfrac{1}{2} = \tfrac{1}{2}\sum_n q_{\mathrm{train}}(n \mid z)\, \operatorname{sign}(\alpha_z^\top \theta_n) = \tfrac{1}{2}\, f_z(s_{z,\cdot}), \qquad f_z(s_{z,\cdot}) = \sum_n q_{\mathrm{train}}(n \mid z)\, s_{z,n}. \tag{20}$$

By Proposition 15 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 9 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 10. 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 12). □

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 22 (Finite-temperature RLHF is manipulable)**.** *Suppose, as in Corollary 16, 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 Bradley–Terry–Luce 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 15 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 2(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 16 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.

The deviation is small, however, whenever labelers are decisive on every deployment query.

Corollary 23. *In the setting of Corollary 22, suppose labels are truthful at precision $\eta > 0$, $\mathbb{E}[c \mid n, z] = \sigma(\eta\, \alpha_z^\top \theta_n)$, and the profile is generic. Let $A = \mathbb{E}_{z\sim q_{\mathrm{dep}}}\big[\|\alpha_z\|\big]$ and $m_n = \min\{|\alpha_z^\top \theta_n| : z \in \mathrm{supp}(q_{\mathrm{dep}})\}$. Then*

$$\Big\|\psi\big(\hat{\theta}_\eta^{\mathrm{RLHF}}\big) - \sum_{n=1}^{N} \lambda_n \psi_n^\star\Big\| \le A \sum_{n=1}^{N} \lambda_n e^{-\eta m_n}.$$

If the cross-entropy objective is augmented with an ℓ_2 penalty $\frac{\rho}{2}\|\theta\|^2$, the deployed impact is shifted by exactly $-\rho\,\hat{\theta}^{\mathrm{RLHF}}$, a vector of norm $\rho\|\hat{\theta}^{\mathrm{RLHF}}\|$.

Proof. By Corollary 22 with $\hat{\theta}_n = \eta\theta_n$, $\psi(\hat{\theta}_\eta^{\mathrm{RLHF}}) = \sum_n \lambda_n \psi(\eta\theta_n)$. By Theorem 2(5), $\psi_n^\star = \frac{1}{2}\mathbb{E}_z[\alpha_z \operatorname{sign}(\alpha_z^\top \theta_n)]$, so with $a_z = \alpha_z^\top \theta_n$, $\psi(\eta\theta_n) - \psi_n^\star = \mathbb{E}_z\big[\alpha_z\big(\sigma(\eta a_z) - \frac{1}{2} - \frac{1}{2}\operatorname{sign}(a_z)\big)\big]$. The bracket has magnitude $\sigma(-\eta|a_z|) \le e^{-\eta|a_z|} \le e^{-\eta m_n}$, giving $\|\psi(\eta\theta_n) - \psi_n^\star\| \le A e^{-\eta m_n}$; averaging over λ gives the bound. With the penalty, the stationarity condition in the proof of Proposition 15 becomes $\mathbb{E}_z[(\mu_z - \sigma(\alpha_z^\top \hat{\theta}^{\mathrm{RLHF}}))\alpha_z] = \rho\,\hat{\theta}^{\mathrm{RLHF}}$, so $\psi(\hat{\theta}^{\mathrm{RLHF}}) = \mathbb{E}_z[(\mu_z - \frac{1}{2})\alpha_z] - \rho\,\hat{\theta}^{\mathrm{RLHF}}$. □

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 $q_{\mathrm{train}} \in \Delta([N] \times Z)$, 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 24 (Marginal Alignment Externality)**.** For a training source (n, z), let $q_\varepsilon = (1-\varepsilon)\, q_{\mathrm{train}} + \varepsilon\, \delta_{(n,z)}$, where $\delta_{(n,z)}$ is a point mass on (n, z). The **marginal alignment externality** of source (n, z) on agent m is

$$\gamma_{n,z}^m = \frac{d}{d\varepsilon} U_m(\hat{\theta}_{q_\varepsilon})\Big|_{\varepsilon=0} = \theta_m^\top \frac{d\,\psi(\hat{\theta}_{q_\varepsilon})}{d\varepsilon}\Big|_{\varepsilon=0}, \tag{21}$$

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 $\hat{\theta}^{\mathrm{RLHF}}$ to pooled data, ignoring agent identities at training time:

$$\hat{\theta}_{q_{\mathrm{train}}}^{\mathrm{RLHF}} \in \arg\max_{\theta\in\mathbb{R}^k} -\mathbb{E}_{(n,z)\sim q_{\mathrm{train}}}\, \mathbb{E}_{c|n,z}\left[-c\log\sigma(\alpha_z^\top\theta) - (1-c)\log\big(1-\sigma(\alpha_z^\top\theta)\big)\right]. \tag{22}$$

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 $S_{\mathrm{train}} = \mathrm{span}\{\alpha_z : z \in \mathrm{supp}(q_{\mathrm{train}})\}$, with $S_{\mathrm{dep}} \subseteq S_{\mathrm{train}}$ required.) The first-order condition is $g(\hat{\theta}^{\mathrm{RLHF}}; q_{\mathrm{train}}) = 0$ 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 $g(\hat{\theta}^{\mathrm{RLHF}}; q_{\mathrm{train}}) = 0$ to cancel the distributional term gives the influence

function

$$\left.\frac{d\,\hat{\theta}_{q_\varepsilon}}{d\varepsilon}\right|_0 = \left(\sigma(\alpha_z^\top \theta_n) - \sigma(\alpha_z^\top \hat{\theta}^{\mathrm{RLHF}})\right) H^{-1}\alpha_z, \qquad H = \mathbb{E}_{(n',z')\sim q_{\mathrm{train}}}\left[v(\alpha_{z'}^\top \hat{\theta}^{\mathrm{RLHF}})\,\alpha_{z'}\alpha_{z'}^\top\right]. \tag{23}$$

Applying the Jacobian $\nabla\psi(\hat{\theta}^{\mathrm{RLHF}}) = \mathbb{E}_{z'\sim q_{\mathrm{dep}}}[v(\alpha_{z'}^\top \hat{\theta}^{\mathrm{RLHF}})\alpha_{z'}\alpha_{z'}^\top]$ (Theorem 2) 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):

$$\gamma_{n,z}^W = \underbrace{\left(\sigma(\alpha_z^\top \theta_n) - \sigma(\alpha_z^\top \hat{\theta}^{\mathrm{RLHF}})\right)}_{\text{preference deviation on query } z} \underbrace{\left(\alpha_z^\top \eta\right)}_{\text{welfare sensitivity of query } z} \tag{24}$$

where

$$\eta = \mathbb{E}_{(n',z')\sim q_{\mathrm{train}}}\left[v(\alpha_{z'}^\top \hat{\theta}^{\mathrm{RLHF}})\,\alpha_{z'}\,\alpha_{z'}^\top\right]^{-1}\,\mathbb{E}_{z\sim q_{\mathrm{dep}}}\left[(\alpha_z^\top \bar{\theta}_w)\,v(\alpha_z^\top \hat{\theta}^{\mathrm{RLHF}})\,\alpha_z\right] \tag{25}$$

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

Theorem 25.

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

$$\mathbb{E}_{(n,z)\sim q_{\mathrm{train}}}\left[\gamma_{n,z}^W\right] = 0. \tag{26}$$

   Individual sources may have $\gamma_{n,z}^W > 0$ (upweighting would improve welfare) or $\gamma_{n,z}^W < 0$, 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:*

$$\gamma_{n,z}^W = 0 \qquad \forall\,(n,z) \in \mathrm{supp}(q_{\mathrm{train}}). \tag{27}$$

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

Proof of Theorem 25. (1). By the closed form (24), $\mathbb{E}_{(n,z)\sim q_{\mathrm{train}}}[\gamma_{n,z}^W] = \big(\mathbb{E}_{(n,z)\sim q_{\mathrm{train}}}[(\sigma(\alpha_z^\top \theta_n) - \sigma(\alpha_z^\top \hat{\theta}^{\mathrm{RLHF}}))\,\alpha_z]\big)^\top \eta = 0$, 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 $\gamma_{n,z}^W$ is the directional derivative toward the vertex (n,z) of the simplex, $\gamma_{n,z}^W = \frac{d}{d\varepsilon}W(q_\varepsilon)\big|_0 = \partial_{n,z}W - \mathbb{E}_{q_{\mathrm{train}}}[\partial W]$. 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 $\gamma_{n,z}^W = \lambda - \lambda = 0$ 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 26 (Query-level decomposition). The impact gap decomposes across deployment queries as

$$\bar{\theta}_w^\top \Delta\psi = \mathbb{E}_{z\sim q_{\text{dep}}}\left[(\bar{\theta}_w^\top \alpha_z)\Big(\tfrac{1}{2}\operatorname{sign}(\alpha_z^\top \bar{\theta}_w) - \big(\sigma(\alpha_z^\top \hat{\theta}^{\text{RLHF}}) - \tfrac{1}{2}\big)\Big)\right]. \tag{28}$$

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

Definition 27. 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 28.

1. *A parameter $\theta \in \Theta^{\text{indiv}}$ if and only if $\alpha_z^\top \theta \in \big[\min_n \alpha_z^\top \theta_n,\ \max_n \alpha_z^\top \theta_n\big]$ 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 28.

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

$$\sigma(\alpha_z^\top \theta) = \sum_{n=1}^N \pi(n|z)\, \sigma(\alpha_z^\top \theta_n) \tag{29}$$

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

$$\sigma(\alpha_z^\top \theta) = (1 - t_z)\, \sigma(\alpha_z^\top \theta_{\underline{n}}) + t_z\, \sigma(\alpha_z^\top \theta_{\overline{n}}). \tag{30}$$

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 15 exactly, with zero population loss. Anonymous weighting, by contrast, is far more rigid:

Proposition 29. *If $S \neq \{0\}$, then $P_S\,\Theta^{\text{anon}}$ is either empty or a singleton.*

Proof of Proposition 29. When $\theta \in \Theta^{\text{anon}}$,

$$\alpha_z^\top \theta = \sigma^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}\sigma(\alpha_z^\top \theta_n)\right) \qquad \forall z \in Z \tag{31}$$

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 30 (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 30. The proof of Proposition 28(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 (31), 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:

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

Proposition 31. *Suppose every query is represented in training, $\text{supp}(q_{\text{train}}) = Z$. If $\hat{\theta}^{\text{RLHF}}_{q_{\text{train}}}$ achieves zero population training loss, then $P_S\hat{\theta}^{\text{RLHF}}_{q_{\text{train}}} \in P_S\Theta^{\text{indiv}}$.*

Proof of Proposition 31. If $\hat{\theta}^{\text{RLHF}}_{q_{\text{train}}}$ achieves zero excess loss, then every KL term vanishes on the support of q_{train}, hence

$$\sigma(\alpha_z^\top \hat{\theta}^{\text{RLHF}}_{q_{\text{train}}}) = \sum_{n=1}^{N} q_{\text{train}}(n \mid z)\,\sigma(\alpha_z^\top \theta_n) \qquad \forall z \in \text{supp}(q_{\text{train}}) \tag{32}$$

Since $\text{supp}(q_{\text{train}}) = Z$, this is exactly the mixture-recovery condition of Proposition 28(1) at every $z \in Z$. As $\hat{\theta}^{\text{RLHF}}_{q_{\text{train}}}$ is only identifiable in the subspace S, applying Proposition 28 gives $P_S\hat{\theta}^{\text{RLHF}}_{q_{\text{train}}} \in P_S\Theta^{\text{indiv}}$. □

D Additional Material for Section 6

We first define alignment externalities, then give the proofs and interpretation of constrained welfare and harm bounds. The public-spirit and participation-externality examples follow, each with its proof.

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

$$\Gamma_{n,m} = U_m(\hat{\theta}) - U_m(\hat{\theta}_{-n}) = \theta_m^\top(\psi - \psi_{-n})\,. \tag{33}$$

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

$$\Gamma_{n,-n} = \sum_{m \neq n} w_m\, \Gamma_{n,m} = (1 - w_n)\, \bar{\theta}_{-n}^\top \Delta_n, \tag{34}$$

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.

D.2 Constrained Welfare and Harm Bounds

Proof of Theorem 17. Define the Lagrangian

$$\mathcal{L}(\psi, \eta) = \bar{\theta}_w^\top \psi + \sum_{n=1}^{N} \eta_n(\gamma_n^\top \psi - c_n), \qquad \eta_n \geq 0. \tag{35}$$

Equivalently,

$$\mathcal{L}(\psi, \eta) = \left(\bar{\theta}_w + \sum_{n=1}^{N} \eta_n \gamma_n\right)^\top \psi - \sum_{n=1}^{N} \eta_n c_n. \tag{36}$$

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

By Theorem 2, $\bar{\Psi}$ 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

$$\psi^{\mathrm{wel}} \in \arg\max_{\psi \in \bar{\Psi}} \left(\bar{\theta}_w + \sum_{n=1}^{N} \eta_n \gamma_n\right)^\top \psi. \tag{37}$$

Complementary slackness gives

$$\eta_n(\gamma_n^\top \psi^{\text{wel}} - c_n) = 0 \qquad \forall n. \tag{38}$$

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

$$\bar{\theta}_w + \sum_{n \in \mathcal{B}} \eta_n \gamma_n \tag{39}$$

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

$$\frac{\partial V}{\partial c_n} = -\eta_n. \tag{40}$$

□

Proof of Corollary 19. Apply Theorem 17 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

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

□

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

Proposition 33 (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 32. Then the welfare floor constraint $\theta_m^\top \psi \geq b_m$ is equivalent to*

$$\Gamma_{n,m} \;\geq\; b_m - \theta_m^\top \psi_{-n} \qquad \forall\, n. \tag{41}$$

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 $\Gamma_{n,m} \geq b_m - \theta_m^\top \psi_{-n}$. □

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.3 traces the same effect as a function of the choice-precision parameter β.

A harm-bounded mechanism allows us to regulate individuals' participation externalities (Figure 4 in Appendix E.4). The bright rows in the harm-bounded 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.

D.3 Additional 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.[8]

[8]This is a special case of Theorem 17 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).

Definition 34 (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 35 (of Theorem 17)**.** *The public-spirit optimum satisfies*

$$\psi_\gamma^{\mathrm{ps}} \in \arg\max_{\psi \in \bar{\Psi}} \quad \left(\left(1 + \gamma \sum_{n \in \mathcal{B}} \eta_n\right) \bar{\theta}_w + (1-\gamma) \sum_{n \in \mathcal{B}} \eta_n\, \theta_n\right)^\top \psi, \tag{42}$$

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

Proof of Corollary 35. Apply Theorem 17 with

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

The effective welfare direction is

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

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

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

□

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.

D.4 Additional Example: Bounding Participation Externalities

The welfare floor (Corollary 19) 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,

$$\bar{\theta}_{-n} = \frac{1}{1 - w_n} \sum_{m \neq n} w_m \theta_m, \tag{43}$$

which is well defined whenever $w_n < 1$.

Definition 36 (Externality-bounded alignment)**.** For a reference impact ψ_0 and tolerances $b \in \mathbb{R}^N_{\geq 0}$: $\psi_b^{\mathrm{ext}} \in \arg\max_{\psi \in \bar{\Psi}} \; \bar{\theta}_w^\top \psi \quad \text{s.t.} \quad \bar{\theta}_{-n}^\top \psi \geq \bar{\theta}_{-n}^\top \psi_0 - b_n \quad \forall\, n.$

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 37 (of Theorem 17)**.** *The externality-bounded optimum satisfies*

$$\psi_b^{\text{ext}} \in \arg\max_{\psi \in \bar{\Psi}} \left(\left(1 + \sum_{n \in \mathcal{B}} \frac{\eta_n}{1 - w_n} \right) \bar{\theta}_w \; - \; \sum_{n \in \mathcal{B}} \frac{w_n \, \eta_n}{1 - w_n} \, \theta_n \right)^\top \psi, \tag{44}$$

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

Proof of Corollary 37. Apply Theorem 17 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 effective direction shifts weight in the opposite direction from the welfare floor. In harm-bounded alignment (Corollary 19), binding agents receive *additional* weight: $w_n \to w_n + \eta_n$. Here, binding agents have their weight *relatively* reduced. Collecting the coefficient on θ_n for a binding agent n, the amplification of $\bar{\theta}_w$ contributes $w_n \eta_n / (1 - w_n)$ and the subtraction removes exactly the same amount, so agent n's effective weight is $w_n \big(1 + \sum_{m \in \mathcal{B},\, m \neq n} \eta_m / (1 - w_m)\big)$, while a non-binding agent's is $w_n \big(1 + \sum_{m \in \mathcal{B}} \eta_m / (1 - w_m)\big)$. No agent's weight falls below w_n (the effective direction is $\bar{\theta}_w + \sum_{n \in \mathcal{B}} \eta_n \bar{\theta}_{-n}$, a positive combination), but binding agents are weighted less than everyone else. 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 relative to the rest of the population.

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 $1/N$.

E Additional Experiments

E.1 Dataset Information and Experimental Setup

All experiments were run locally on commodity hardware. An A100 was rented to compute the 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 of 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 answered ~ 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 (Qwen3-Embedding-8B) 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 Robustness to Nonlinear Utilities

Our framework assumes that each individual's utility is linear in a fixed feature map, $u_n(x) = \varphi(x)^\top \theta_n$. It does not assume that utility is linear in the raw attributes of an outcome. If φ is replaced with a richer map $\tilde{\varphi}$ that includes nonlinear combinations of the attributes, the result is a more expressive utility function that the framework handles without change. We redefine the impact vector, the impact polytope, and the participation externality construction using $\tilde{\varphi}$, and every result carries over. Our baseline analyses use the raw attributes as features, which amounts to assuming that preferences are linear in those attributes. In this section we measure how sensitive those analyses are to that choice by augmenting the features with second-order terms. We ask two questions. First, does a linear or a nonlinear utility fit human choices better on our datasets? Second, where a nonlinear utility fits better, do our welfare and externality conclusions change?

The nonlinear models. For a comparison between options A and B with feature difference $\alpha = \varphi_A - \varphi_B$, the linear model is the Bradley–Terry choice probability $\sigma(\theta^\top \alpha)$. The quadratic model adds a squared term for each feature, $\sigma\big(\theta^\top \alpha + \gamma^\top (\varphi_A^2 - \varphi_B^2)\big)$. This lets the return to each feature increase or diminish while utility stays additive across features. For Moral Machine we also fit an interaction model, which adds all $\binom{k}{2}$ products $\varphi_{A,i}\varphi_{A,j} - \varphi_{B,i}\varphi_{B,j}$. It captures value that depends on which features occur together. This is the harder test, because the quadratic model cannot represent it. Every model is a logistic regression with no intercept and ℓ_2 penalty $C = 1$, matching our main analyses, on features standardized across all options in the dataset. Moral Machine and Community Alignment record only the preferred option, so for those datasets we add each training comparison in both orientations.

Experiment 1: Does a nonlinear utility fit better? We fit each model separately for every agent and compare the models by 5-fold cross-validation on that agent's own comparisons. Agents are those with at least 20 comparisons, except that Kidney Allocation uses the 77 agents with at least 100 comparisons and Moral Machine uses a random sample of 200 annotators. Table 1 reports held-out log loss and accuracy averaged across agents.

Table 1: Held-out fit of linear and quadratic utilities, from per-agent 5-fold cross-validation averaged across agents. Lower log loss is better.

Dataset	Comparisons per agent	Log loss, linear	Log loss, quadratic	Accuracy, linear	Accuracy, quadratic
Kidney Allocation	480	0.285	0.263	87.8%	89.1%
412 Food Rescue	45	0.447	0.542	79.5%	77.6%
Moral Machine	69	1.005	2.024	60.1%	58.6%
Community Alignment	57	0.512	0.544	74.5%	77.3%

Kidney Allocation is the only dataset where the quadratic utility fits better on both log loss and accuracy, and it helps 56 of 77 agents. In this dataset, 23% of comparisons repeat a question the agent has already seen. Agents choose differently on 6.9% of these identical comparisons, so the best any model of an individual can do on them is 93.1% accuracy. The linear and quadratic models reach 89.9% and 92.5%. The quadratic utility therefore predicts human behavior better here, and comes close to the limit set by the agents' own consistency.

Elsewhere the nonlinear models overfit at the amount of data we have per agent. On 412 Food Rescue the quadratic utility helps only 3 of 19 agents. On Moral Machine it helps 1 of 200. The interaction model is no better, with log loss 1.976 and accuracy 58.1%, so even structure that a quadratic model cannot represent does not improve prediction. On Community Alignment the quadratic utility improves accuracy but worsens log loss, and it helps about half the agents (1,079 of 2,387). This is what we expect, because those features are already a learned nonlinear projection of text.

Experiment 2: Where utilities are nonlinear, do our results change? A quadratic utility is a linear utility over the augmented basis $\tilde{\varphi} = [\varphi, \varphi^2]$. We therefore re-ran the whole pipeline in that space, including the impact vectors, the impact polytope and the participation externality matrix, and compared the results with the linear ones. We also report the welfare regret of utilitarian deployment. This is the share of welfare given up by deploying as if utilities were linear when they are really quadratic, computed for deterministic deployments ($\beta \to \infty$). Table 2 reports the results.

Table 2: Agreement between results computed under linear and quadratic utilities. Correlations are Pearson; the externality correlation uses all off-diagonal entries of Γ. Split-half reliability is the correlation between Γ estimated from two random halves of each agent's comparisons under linear utilities. Regret is the share of quadratic-utility welfare lost by deploying the linear utilitarian decisions.

Dataset	Welfare correlation	Externality correlation	Split-half reliability of externalities	Welfare regret
Kidney Allocation	0.94	0.97	0.90	1.1%
412 Food Rescue	0.95	0.92	0.46	0.9%
Moral Machine	0.91	0.58	0.08	1.7%
Community Alignment	0.79	0.47	0.39	4.0%

Kidney Allocation is the most informative test, because it is the only dataset where the quadratic utility fits better. There, per-agent welfare correlates at 0.94 and participation externalities at 0.97, the best-off and worst-off agents are identical, and the welfare regret is about 1%. The two utilities make different choices only on queries where their utilitarian means disagree about which option is better, and they produce the same welfare everywhere else. A 1% regret therefore means they rarely disagree, and little welfare is lost when they do.

On the other datasets, welfare regret is at most 4.0% (Community Alignment), and per-agent welfare correlates between 0.79 and 0.95. The externality correlations are lower on Moral Machine (0.58) and Community Alignment (0.47), but this reflects how precisely externalities can be measured on those datasets, not a difference between the utility models. When we estimate the externality matrix from two random halves of each agent's data, the two estimates correlate at only 0.39 on Community Alignment and 0.08 on Moral Machine, even when both use linear utilities. Switching to quadratic utilities therefore perturbs the externalities no more than redrawing the data does. On Kidney Allocation, where externalities are precisely estimated (split-half reliability 0.90), the two utility models agree closely.

On three of the four datasets the linear model is the better estimator. Either the nonlinear models overfit at realistic per-agent sample sizes, or the features are already nonlinear. On the fourth, where a nonlinear utility does fit better, our framework absorbs it as a linear utility over a larger basis, and the welfare and externality conclusions barely move. The framework is therefore robust to the nonlinear utilities we test, and using them does not change what we find. The script `scripts/linearity_robustness.py` in our code repository reproduces every number in this appendix.

E.3 Welfare Dispersion vs. Choice Precision

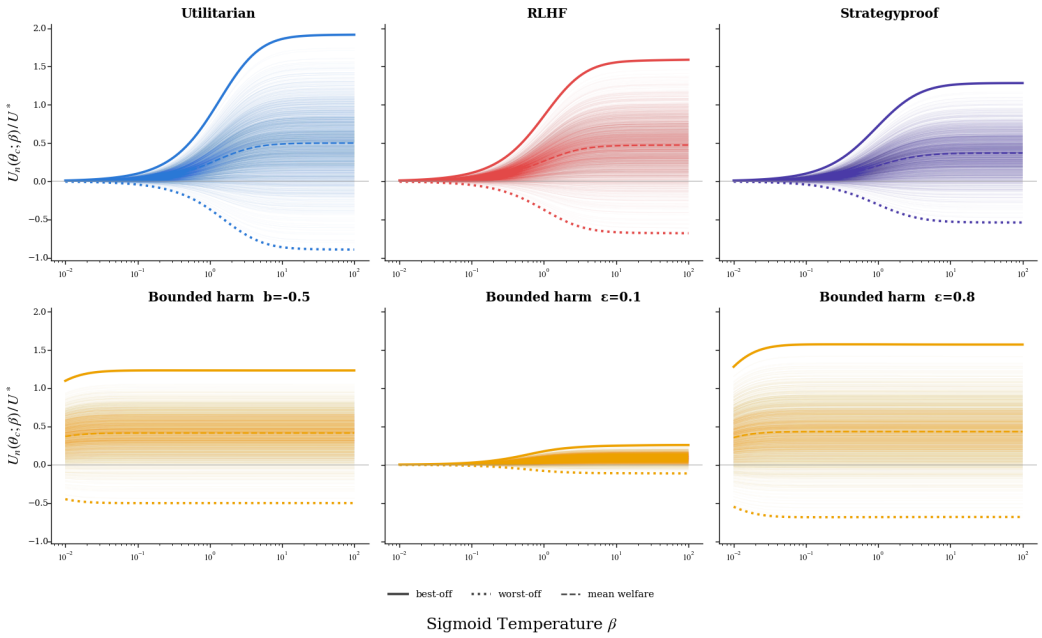


Figure 3: Community Alignment dataset: per-agent welfare $U_n/U^\star$ as a function of the choice-precision parameter β, for each mechanism. The shaded band is the density of all 2387 agents; solid, dashed, and dotted lines mark the best-off agent, mean welfare, and worst-off agent. As β grows, choices 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 harm bounds are empirically satisfied.

E.4 Per-Agent Harm Bound

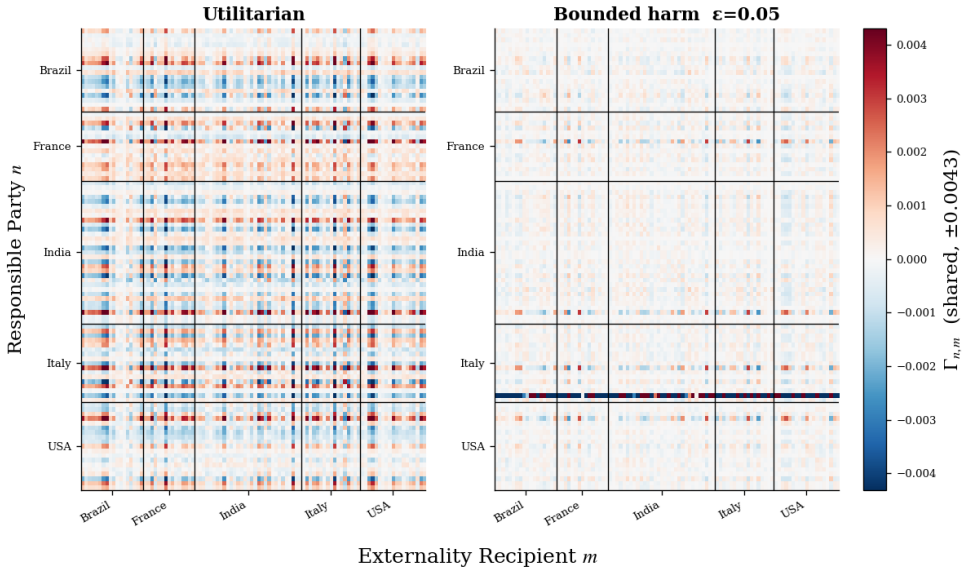


Figure 4: Participation externalities under the *per-agent* harm bound. Here each agent's welfare floor is a fixed fraction of their own maximum achievable welfare, $b_n = -\epsilon\, h_{\bar{\Psi}}(\theta_n)$ with $\epsilon = 0.05$ (lose at most 5% of one's own best case), rather than a single absolute 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.5 Person-on-Person Participation Externalities

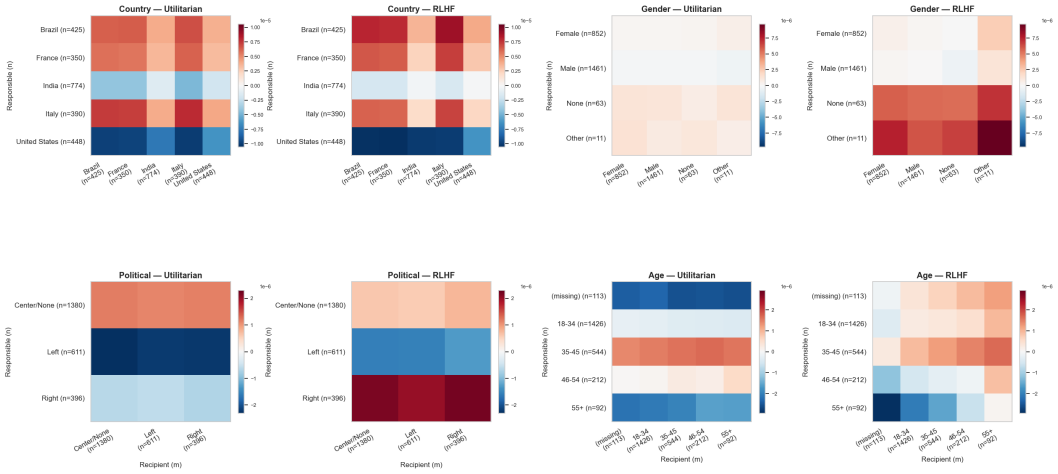


Figure 5: Community Alignment Dataset: Utilitarian vs. RLHF participation alignment externalities by demographic. The figure includes 2387 annotators, with participation externalities averaged within each group.

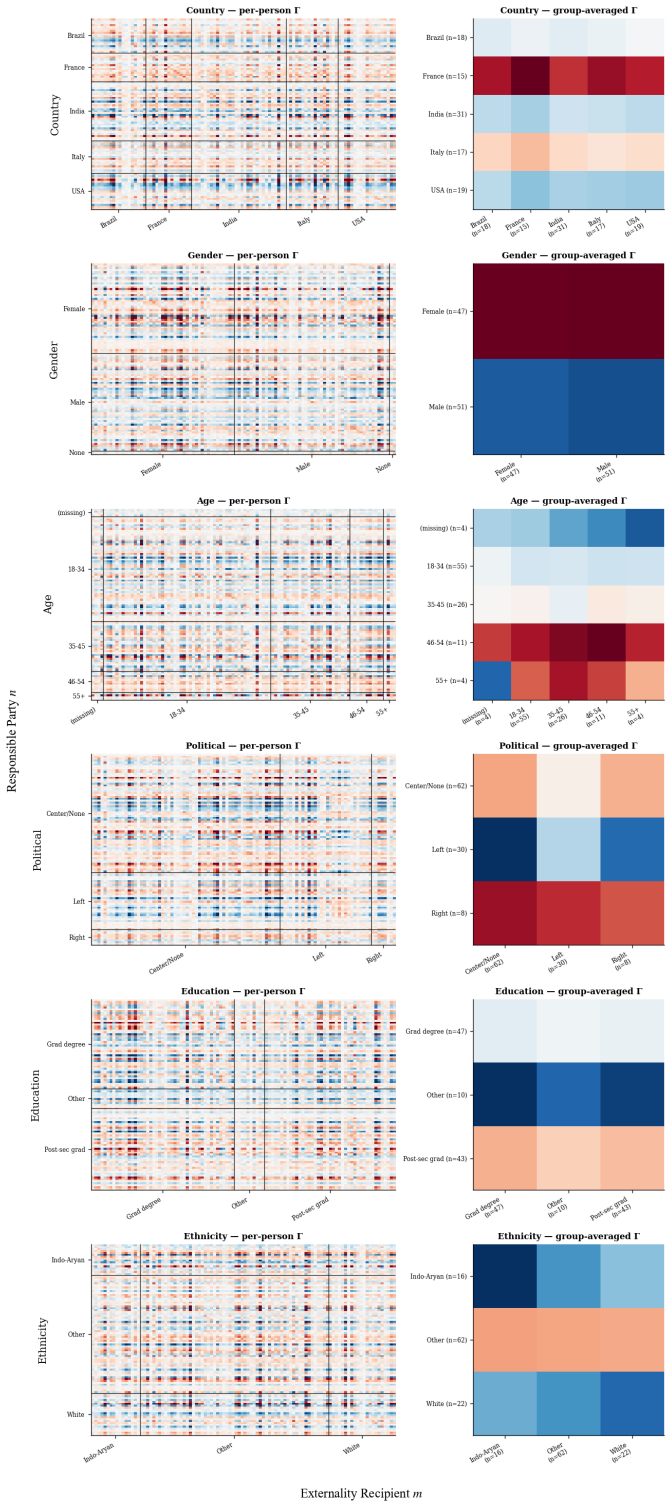


Figure 6: Community Alignment Dataset: 100-person sample with granular person-on-person participation externalities compared to the group-averaged participation externalities. Our participation externality formalization gives us a microscope to analyze where the per-individual welfare effects are coming from.

E.6 Four Dataset Comparisons

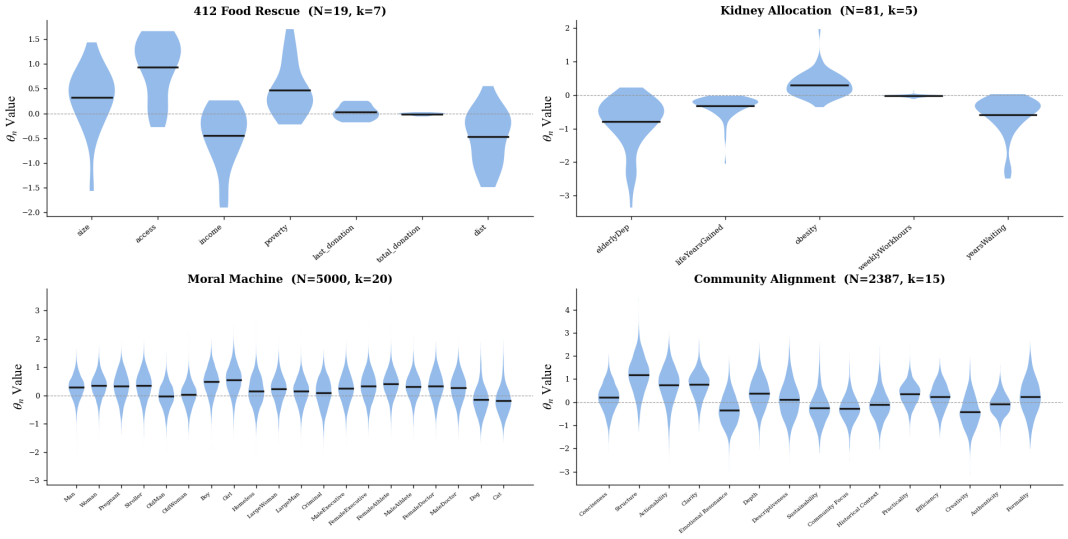


Figure 7: Preference distribution per feature dimension for four datasets. Moral Machine and Community Alignment have high disagreement, with a high number of preferences on either side of 0, while Kidney Allocation and 412 Food Rescue have less disagreement: people disagree on the magnitude of preference,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

The empirical welfare losses of the uniform random dictatorship, together with the average majority margin of Proposition 21, are reported in Table 3.

Table 3: Relative welfare loss of the uniform random dictatorship, $1 - \mathbb{E}[U(\mu^{\mathrm{sp}})]/U^\star$, against the average majority margin M (Proposition 21).

Dataset	Relative welfare loss	Average majority margin
Kidney Allocation	15.7%	0.66
Community Alignment	26.0%	0.60
412 Food Rescue	31.0%	0.52
Moral Machine	41.8%	0.41

Proposition 38 (Population welfare guarantee)**.** *Suppose $\theta_1, \dots, \theta_N$ are drawn i.i.d. from a distribution P on $\mathbb{R}^k$ with mean μ_P and finite second moments, and let $w_n = 1/N$. Population welfare at impact ψ is $\mathbb{E}_{\theta\sim P}[\theta^\top \psi] = \mu_P^\top \psi$, so the population problem is (1) with $\bar\theta_w$ replaced by μ_P. If $\hat\psi \in \arg\max_{\psi\in\bar\Psi} \bar\theta_w^\top \psi$ is the sample optimum, then*

$$\max_{\psi\in\bar\Psi} \mu_P^\top \psi - \mu_P^\top \hat\psi \le 2\,\|\bar\theta_w - \mu_P\| \sup_{\psi\in\bar\Psi} \|\psi\| \le \|\bar\theta_w - \mu_P\|\, \mathbb{E}_{z\sim q_{\mathrm{dep}}}\big[\|\alpha_z\|\big], \tag{45}$$

and $\mathbb{E}\big[\|\bar\theta_w - \mu_P\|\big] \le \sqrt{\operatorname{tr}\operatorname{Cov}_P(\theta)/N}$, so the expected regret is $O(N^{-1/2})$.

Proof. Let $\psi_P \in \arg\max_{\psi\in\bar\Psi} \mu_P^\top \psi$ be a population optimum. Since $\hat\psi$ maximizes $\bar\theta_w^\top \psi$ over $\bar\Psi$, we have $\bar\theta_w^\top(\psi_P - \hat\psi) \le 0$, and therefore

$$\mu_P^\top(\psi_P - \hat\psi) = (\mu_P - \bar\theta_w)^\top(\psi_P - \hat\psi) + \bar\theta_w^\top(\psi_P - \hat\psi) \le \|\mu_P - \bar\theta_w\|\,\|\psi_P - \hat\psi\| \tag{46}$$

by the Cauchy–Schwarz inequality. Since $\psi_P, \hat\psi \in \bar\Psi$, the triangle inequality and Theorem 2(3) give

$$\|\psi_P - \hat\psi\| \le 2 \sup_{\psi\in\bar\Psi} \|\psi\| \le \mathbb{E}_{z\sim q_{\mathrm{dep}}}\big[\|\alpha_z\|\big], \tag{47}$$

which establishes the first display. For the moment bound, $\bar{\theta}_w = \frac{1}{N}\sum_{n=1}^{N}\theta_n$ is the mean of N i.i.d. draws from P, so $\mathbb{E}[\bar{\theta}_w] = \mu_P$ and $\mathrm{Cov}(\bar{\theta}_w) = \mathrm{Cov}_P(\theta)/N$; taking the trace,

$$\mathbb{E}\big[\|\bar{\theta}_w - \mu_P\|^2\big] = \mathrm{tr}\,\mathrm{Cov}(\bar{\theta}_w) = \frac{\mathrm{tr}\,\mathrm{Cov}_P(\theta)}{N}, \tag{48}$$

which is finite by the second-moment assumption. Since $x \mapsto \sqrt{x}$ is concave, Jensen's inequality gives $\mathbb{E}\big[\|\bar{\theta}_w - \mu_P\|\big] \le \sqrt{\mathbb{E}\big[\|\bar{\theta}_w - \mu_P\|^2\big]} = \sqrt{\mathrm{tr}\,\mathrm{Cov}_P(\theta)/N}$. □

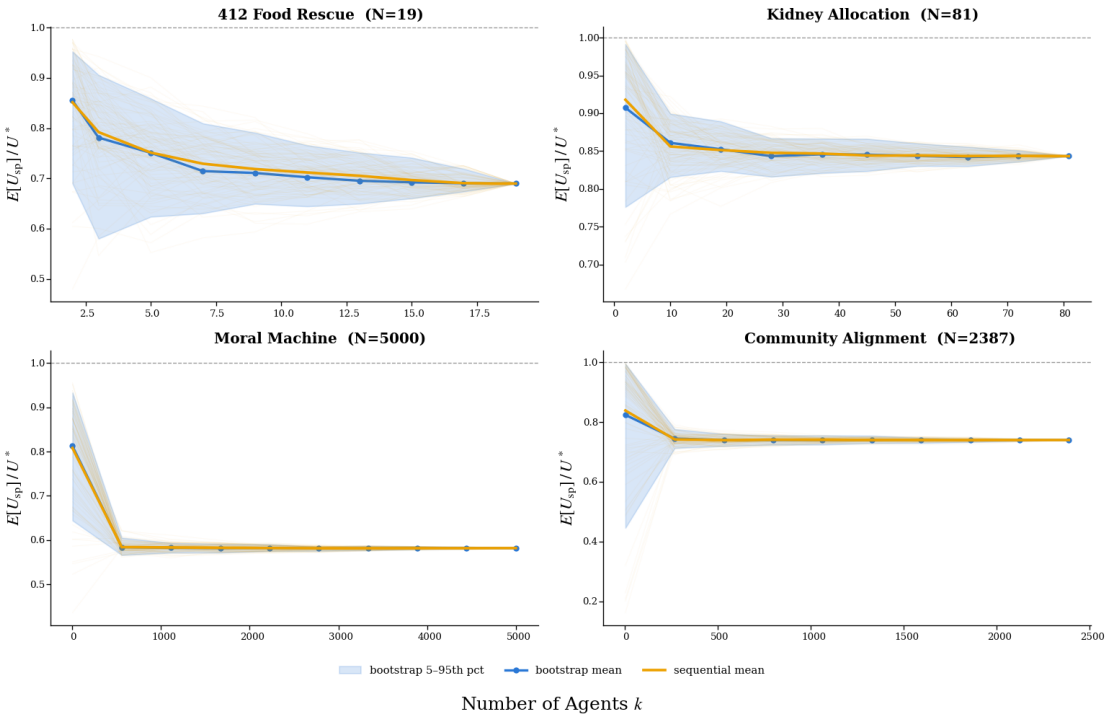


Figure 8: Empirical finite-sample stability of the strategyproof mechanism's welfare; Proposition 38 gives the corresponding theoretical population guarantee under the assumption that annotators are representative of the target population. Expected welfare for four datasets. Bootstrap means and standard deviations were calculated by bootstrapping l agents 100 times, for each subset size l, 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.

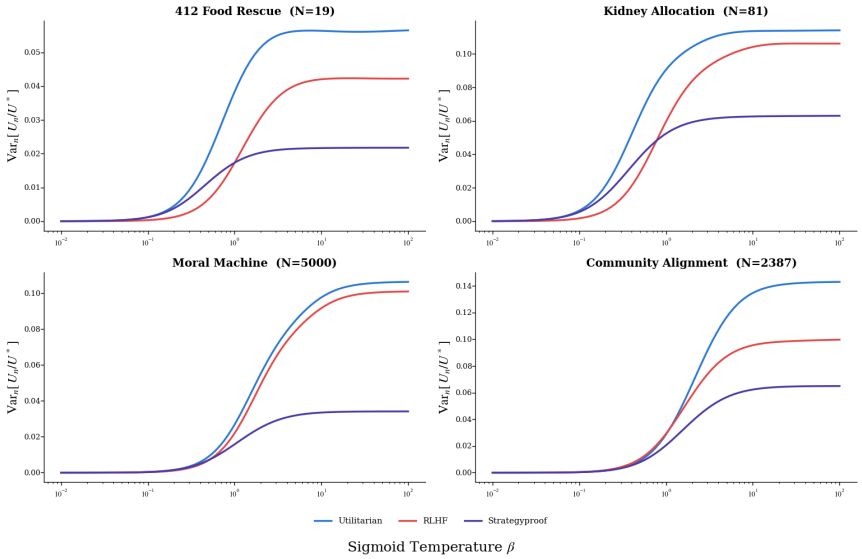


Figure 9: Welfare variance over four datasets. The strategyproof mechanism leads to the least welfare dispersion, while RLHF follows as a distant second for larger values of β.

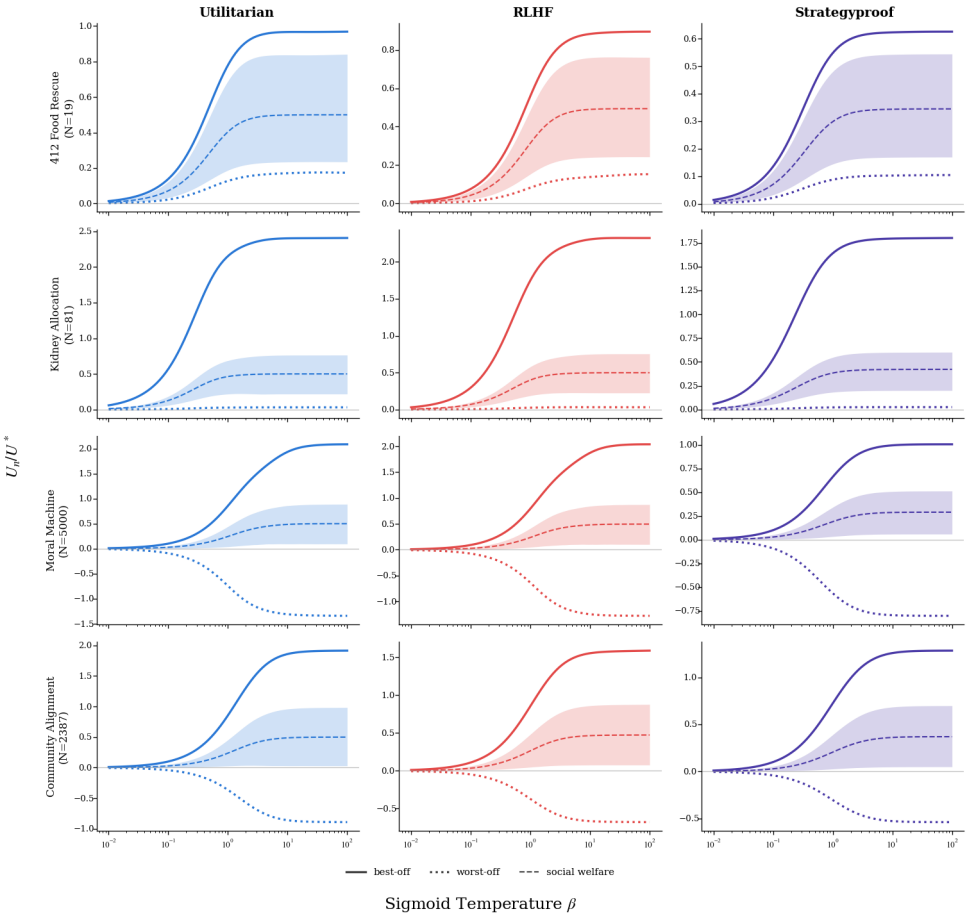


Figure 10: Welfare of the best- and worst-off agents in four datasets for three model parameters.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims in the abstract and introduction about the impact framework and what the framework accomplishes are reflected in Sections 4–7 of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The final section, Limitations and Discussion, addresses the limitations of the work regarding the stringency of assumptions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theorems and proofs are numbered and cross-referenced. All proofs are in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All details about experiments, where the data came from, and how the data was cleaned before we applied our results are described in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, the link to the code is given in Appendix E.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, all relevant information is in Appendix E. We simply implemented the calculations detailed in our theorems to get our charts.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Bootstrap standard deviations are reported for the expected-welfare figure (Figure 8 in Appendix E).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All compute information is in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, we have reviewed the Code of Ethics and comply with it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, impacts are discussed in the Discussion section.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper does not pose such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Sources and access conditions for the data are given in Appendix E. Only the 412 dataset is not public.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets are released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We use only previously collected, publicly released or permission-granted datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We use only previously collected, publicly released or permission-granted datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We describe the use of LLMs in the AI Disclosure paragraph following the acknowledgments.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.