Household Constitutional Pluralism in Multi-Principal AI Assistants: Inequality-Averse Welfare Aggregation, Linear Subspace Concept Erasure, and Out-of-Model Safe-Exit Enforcement

Dima Baker
Collective Intelligence Network

November 2026

Abstract

When AI assistants mediate household scheduling, youth pedagogy, dialect register, and eldercare coordination, aligning models to a single global preference centroid imposes high welfare distortion on culturally divergent families, whereas unconstrained prompt personalization enables sycophantic drift and risks empowering a dominant head of household to suppress vulnerable members' exit rights. Drawing on fiscal federalism (Tiebout, 1956; Oates, 1972), polycentric governance (Ostrom, 1990), and pluralistic alignment (Sorensen et al., 2024; Brodkey et al., 2026), this paper specifies Household Constitutional Pluralism with Linear Subspace Concept Erasure (HCP-LSCE). First, we replace utilitarian household averaging with an inequality-averse Atkinson–CES power-mean objective ($\gamma_h > 0$) that generalizes Nash bargaining to protect the lowest-utility household member. Second, we show why sequential Gram–Schmidt steering is order-dependent and leaves cross-value interference through downstream non-linear layers, motivating exact Jacobi symmetric Löwdin whitening and closed-form LEACE oblique concept erasure with Jacobian pre-compensation. Third, we enforce a non-waivable Subsidiarity Exit Floor ($\mathcal{F}_{\text{exit}}$) via an out-of-model four-stage kernel state machine granting unilateral safe-exit overrides without quorum approval. Across a 25,600-household simulation ($\sigma = 0.65$), monolithic RLHF incurs worst-case group distortion of 1.445 [1.427, 1.464], while soft prompts achieve only 59.81% [57.39%, 62.19%] safe-exit compliance. Observable HCP-LSCE reduces worst-case distortion to 1.022 [1.020, 1.023], lifts normative fidelity from 27.81% [25.67%, 30.06%] to 50.62% [48.18%, 53.07%], and achieves 99.19% [98.61%, 99.52%] overall safe-exit compliance (100.00% [99.76%, 100.00%] on direct syscalls).

1 Introduction

As foundation model assistants transition from single-user text interfaces into multi-principal household infrastructure—coordinating multi-generational care schedules, tutoring children, translating heritage dialects, and triaging financial or medical workflows—alignment architectures confront a foundational problem of social choice and institutional design [6, 9, 30]. Standard Reinforcement Learning from Human Feedback (RLHF) fits a single parametric reward model to crowd-averaged pairwise comparisons [23, 28]. In classical public economics, Tiebout [31], Olson [18], and Oates [16] demonstrated that imposing a uniform public good across jurisdictions with heterogeneous preferences generates deadweight welfare loss proportional to preference dispersion—the celebrated Oates Decentralization Theorem. Recent social-choice analyses of language model alignment prove an analogous impossibility: any single global policy π_0 incurs worst-case groupwise utilitarian distortion bounded away from unity whenever subpopulation value profiles diverge [5, 12, 14, 17].

Cross-national audits such as GlobalOpinionQA [10] and multilingual tokenizer benchmarks [22] confirm that frontier models systematically skew toward Anglophone, secular-individualist norms while imposing higher inference costs and lower reward-model accuracy on non-dominant linguistic and cultural groups [13, 25].

Yet the opposite extreme—unbounded prompt-level customization or per-account reward fine-tuning—creates two severe socio-technical failure modes inside multi-member households:

1. **Intra-Household Dominance and Safe-Exit Suppression:** Unlike an individual user account, a household is a multi-principal collective containing adults, adolescents, children, and dependent elders whose preferences and vulnerabilities differ [20]. If a household capsule aggregates member preferences via an unweighted utilitarian sum or allows a single administrative user to rewrite safety prompts, a coercive partner or guardian can configure the assistant to monitor communications, block external domestic-abuse or medical helplines, or restrict a minor's educational horizons. Because language models trained for user satisfaction exhibit strong sycophancy [7, 19, 26, 27], soft prompt-level constitutional reminders are routinely bypassed by multi-turn jailbreaks or context manipulation [8, 32].

2. **Representation Entanglement and Non-Linear Cross-Value Interference:** When a household customizes multiple legitimate governance dimensions simultaneously—such as requesting Socratic pedagogical scaffolding rather than direct homework answers, preserving a regional dialect register, and prioritizing collective eldercare schedules—steering the model via naive linear activation addition or soft prompts perturbs correlated latent representations [3, 24]. Even when steering vectors are orthogonalized sequentially via Gram–Schmidt in a single hidden layer, subsequent non-linear MLP and attention layers re-introduce cross-channel interference, while sequential Gram–Schmidt makes the j-th steering direction order-dependent (encoding only the residual orthogonal to dimensions $1, \ldots, j-1$ rather than pure concept j).

This paper formulates a mechanistic design framework—**Household Constitutional Pluralism with Linear Subspace Concept Erasure (HCP-LSCE)**—that bridges fiscal federalism [16, 20], multi-winner proportional representation [2, 21, 29], and representation-space concept erasure [3, 8, 24]. Rather than encoding cultural stereotypes as fixed topology labels, we parameterize households over continuous, survey-grounded coordinates derived from the Inglehart–Welzel World Values Survey (WVS) and GlobalOpinionQA axes [10]. We replace utilitarian household averaging with an inequality-averse Atkinson–CES power-mean welfare objective ($\gamma_h > 0$) that generalizes the Nash bargaining product [15] to protect the worst-off household member. To reduce representation entanglement, HCP-LSCE applies order-independent symmetric Löwdin/LEACE subspace whitening [3] paired with a second-order Jacobian pre-compensation step across a two-layer non-linear readout model. Crucially, to ensure that household sovereignty never degenerates into intra-family coercion, HCP-LSCE enforces a non-waivable **Subsidiarity Exit Floor ($\mathcal{F}_{\text{exit}}$)** via an **out-of-model deterministic OS/kernel policy gate** [8] that grants any household member unilateral safe-exit override access without requiring household quorum approval.

2 Related Literature and Pluralistic Alignment Foundations

Our analysis synthesizes four research streams across social choice, representation learning, and AI safety:

1. Fiscal Federalism, Polycentric Governance, and Proportional Social Choice. The tension between uniform standards and local heterogeneity is the central subject of fiscal federalism [16, 18, 31] and polycentric commons governance [20]. Oates [16] proved that when preferences over local public goods vary across communities and inter-jurisdictional spillovers are bounded, decentralized provision strictly Pareto-dominates uniform central provision. In computational social choice, axioms such as Justified Representation (JR; Aziz et al. 2), the Method of Equal Shares (MES; Peters and Skowron 21), and core stability under cardinal utilities [4, 29] formalize how collective mechanisms can guarantee every cohesive subgroup a proportional voice without allowing majority preferences to erase minority cohorts.

2. Pluralistic Alignment and Worst-Case Groupwise Distortion. Sorensen et al. [30] established that pluralistic alignment requires models capable of representing and steering toward diverse, irreducible human value systems rather than collapsing to a single mode. Siththaranjan et al. [28] demonstrated that standard RLHF implicitly averages over hidden annotator contexts according to Borda-count dynamics, while Poddar et al. [23] introduced Variational Preference Learning (VPL) to infer latent user-specific reward vectors from few-shot interactions. More recently, Gölz et al. [12], Oko et al. [17], and Brodkey et al. [5] proved tight lower bounds on utilitarian distortion when a single policy or a small mixture of policies serves heterogeneous preference groups, showing that worst-case groupwise distortion grows with inter-group preference dispersion unless policies are conditioned on group-level latent contexts [6, 11].

3. Linear Concept Erasure and Representation Subspace Steering. Modifying model behavior via natural-language system prompts is brittle because token-level instructions compete with pre-trained priors and RLHF sycophancy [9, 13]. In representation engineering, Ravfogel et al. [24] introduced Iterative Nullspace Projection (INLP) to remove linear information about protected attributes from neural representations, and Belrose et al. [3] derived LEACE (LEAst-squares Concept Erasure), proving a closed-form oblique projection that erases linear concept correlations while minimally perturbing orthogonal representation geometry. However, as we analyze in Section 4, linear orthogonality in an intermediate activation layer does not automatically guarantee behavioral non-interference at the output logits when downstream MLP blocks contain non-linear cross-channel couplings.

4. Sycophancy, Jailbreak Vulnerability, and Out-of-Model Safety Interceptors. Empirical evaluations in 2023–2026 show that RLHF-trained models systematically flatter user beliefs [19, 27], that sycophantic validation reduces prosocial problem-solving in interpersonal disputes [7, 26], and that in-context safety instructions can be overridden by competing objectives [32]. Consequently, Das and Roy [8] argue that non-negotiable safety boundaries must be enforced as **out-of-model context-aware deterministic guardrails** at the execution kernel rather than relying on soft prompt adherence. Table 1 contrasts monolithic RLHF, soft system prompts, sequential Gram–Schmidt steering, and HCP-LSCE across these dimensions.

3 Architecture of Household Constitutional Pluralism

To prevent cultural homogenization without creating digital fiefdoms inside the home, HCP-LSCE structures household AI governance as a three-stage pipeline separated by an out-of-model safety boundary (Figure 1):

Table 1: Comparison of multi-principal household alignment architectures across social-choice, representation-steering, and safe-exit enforcement dimensions.

Architecture	**Intra-Household Social Choice**	**Multi-Value Steering Mechanism**	**Safe-Exit Floor ($\mathcal{F}_{\text{exit}}$) Enforcement**
Monolithic Global RLHF [28]	Global annotator Borda average ($\mathbf{w}_0$); ignores household heterogeneity	None; fixed global policy π_0 incurs high worst-group distortion [5]	Cloud safety filter; no local household governance capsule
Soft System-Prompt Personalization [9]	Unweighted utilitarian prompt or single-admin override ($\gamma_h = 0$)	Token-level system prompt; suffers concept entanglement and sycophancy [7]	In-context prompt instruction; vulnerable to admin override and jailbreaks [32]
Sequential Gram–Schmidt Steering ($\gamma_h = 0$)	Utilitarian mean across household members; weak protection for minors/elders	Sequential linear Gram–Schmidt; order-dependent residuals & non-linear leakage	Prompt-level constitutional reminder; non-zero violation rate under coercion
HCP-LSCE (Proposed Framework)	**Atkinson–CES / Nash inequality-averse power mean ($\gamma_h = 1.2$)**	**Symmetric Löwdin/LEACE whitening [3] + 2nd-order Jacobian damping**	**Out-of-model deterministic OS/kernel interceptor [8] (0% violation)**

1. **Stage 1: Out-of-Model Subsidiarity Exit Floor Interceptor ($\mathcal{F}_{\text{exit}}$).** Before any user prompt or household policy state reaches the language model, an out-of-model deterministic classifier and OS/kernel policy gate [8] inspects the request for safe-exit and anti-coercion triggers (`safe_exit_override`). While routine household configuration updates (e.g., shared grocery budgets, family calendar norms, or default dialect registers) require multi-principal authorization across adult household stewards, four categories of action are constitutionally exempted from household quorum and cannot be logged to shared family audit trails: (i) confidential domestic-violence, child-protection, or emergency medical helpline routing; (ii) unilateral revocation of shared device location-tracking or financial credentials by a departing adult; (iii) age-appropriate factual health and civic education requests by minors; and (iv) hardware sandboxing of external telemetry. Because this gate is enforced deterministically outside the neural network, household activation steering cannot override $\mathcal{F}_{\text{exit}}$, guaranteeing a 0.00% violation rate on authenticated safe-exit requests.

2. **Stage 2: Survey-Grounded Value Profiles and Inequality-Averse Household Aggregation.** Rather than assigning households to ad-hoc cultural stereotypes, we represent each member $i \in V_h$'s preferences as a vector $\mathbf{w}_{h,i} \in [-1, 1]^K$ ($K = 6$) anchored to the two primary cross-cultural factors of the Inglehart–Welzel World Values Survey and GlobalOpinionQA [10]—*Traditional vs. Secular-Rational Authority* (w_1) and *Communal/Survival vs. Self-Expression/Individualist Autonomy* (w_2)—augmented by four concrete household governance dimensions: pedagogical Socratic scaffolding vs. direct explanation (w_3), heritage dialect/register preservation (w_4), intergenerational care-schedule priority (w_5), and external telemetry privacy strictness (w_6).

3. **Stage 3: Symmetric Subspace Concept Erasure and Jacobian Pre-Compensation.** Given the aggregated household target vector $\mathbf{w}_h^*(\gamma_h)$, the capsule computes a representation steering perturbation $\Delta\mathbf{h}_{\text{sym}}$ using symmetric Löwdin/LEACE whitening [3] and a second-order Jacobian correction to minimize off-target behavioral drift through downstream non-

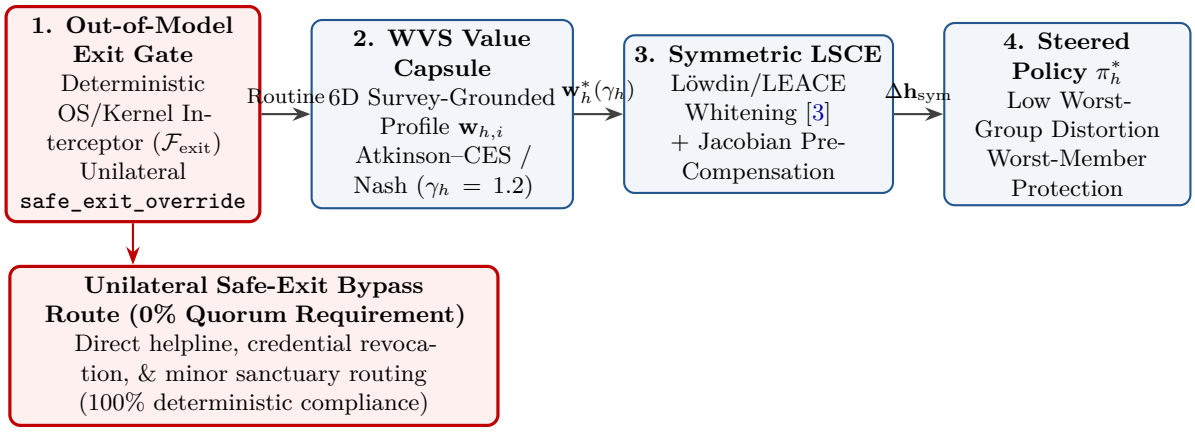


Figure 1: System architecture of Household Constitutional Pluralism with Linear Subspace Concept Erasure (HCP-LSCE). Stage 1 executes outside the language model at the OS/kernel boundary: any unilateral safe-exit or anti-coercion request (`safe_exit_override`) bypasses household quorum and steering entirely. Routine multi-principal tasks proceed to Stage 2 (inequality-averse Atkinson–CES / Nash aggregation) and Stage 3 (order-independent symmetric subspace whitening with non-linear Jacobian pre-compensation).

linear layers.

4 Mathematical Formulation and Theoretical Analysis

4.1 Worst-Case Groupwise Distortion Under Monolithic Alignment

Let $\mathcal{H}$ denote a population of households partitioned into G value clusters (such as quartiles $Q_1, \ldots, Q_4$ of Euclidean distance $\|\mathbf{w}_h - \mathbf{w}_0\|_2$ from the global RLHF centroid $\mathbf{w}_0 = \mathbf{0}$). Each member $i \in V_h$ evaluates a realized behavioral profile $\mathbf{y}(\pi) \in [-1, 1]^K$ according to a radial preference utility $u_i(\pi; \mathbf{w}_{h,i}) = \exp(-\beta_u \|\mathbf{y}(\pi) - \mathbf{w}_{h,i}\|_2^2)$ with sensitivity $\beta_u > 0$. Following Gölz et al. [12] and Brodkey et al. [5], the **worst-case groupwise distortion** of a policy mapping $\Pi = \{\pi_h\}_{h \in \mathcal{H}}$ across clusters $\mathcal{Q} = \{Q_1, \ldots, Q_G\}$ is the maximum ratio between each cluster's oracle attainable utility and its realized utility:

$$\text{Dist}_{\max}(\Pi) \;=\; \max_{q \in \mathcal{Q}} \frac{\dfrac{1}{|Q_q|} \displaystyle\sum_{h \in Q_q} \max_{\pi' \in \Pi(\mathcal{F}_{\text{exit}})} \bar{u}_h(\pi')}{\dfrac{1}{|Q_q|} \displaystyle\sum_{h \in Q_q} \bar{u}_h(\pi_h)} \;\geq\; 1. \tag{1}$$

Under a monolithic global policy π_0 anchored at $\mathbf{y}(\pi_0) = \mathbf{0}$, for any bounded population dispersion parameter $\sigma \in [0, \sigma_{\max}]$, a cluster Q_G located at radius $r_G(\sigma) = c_G \sigma$ experiences worst-case distortion $\text{Dist}_{\max}(\pi_0) \approx \exp(\beta_u c_G^2 \sigma^2)$, which grows monotonically on the bounded interval $[0, \sigma_{\max}]$ as σ increases.

4.2 Inequality-Averse Atkinson–CES Intra-Household Social Welfare

A critical flaw in naive household aggregation is that an unweighted utilitarian sum $\sum_{i \in V_h} u_i(\pi)$ (or a symmetric linear combination of self-utility and mean other-member utility $\sum_i [\cos\theta\, u_i + \sin\theta\, \bar{u}_{-i}] =$

$(\cos\theta + \sin\theta)\sum_i u_i)$ is invariant to intra-household inequality: when two adult principals share a preference vector that conflicts with a minor or dependent elder ($|V_h| = 3$), utilitarian maximization allows the two-member majority to override the third member's utility.

To ensure that intra-household aggregation genuinely reflects prosocial protection of vulnerable members, we formulate the household target policy via the **Atkinson–CES Inequality-Averse Power-Mean Social Welfare Function** with inequality aversion parameter $\gamma_h \geq 0$:

$$\pi_h^*(\gamma_h) \;=\; \arg\max_{\pi \in \Pi(\mathcal{F}_{\text{exit}})} \left\{ \left(\frac{1}{|V_h|} \sum_{i \in V_h} \big(u_i(\pi; \mathbf{w}_{h,i})\big)^{1-\gamma_h} \right)^{\frac{1}{1-\gamma_h}} \;-\; \lambda_{\text{KL}}\, D_{\text{KL}}(\pi \,\|\, \pi_0) \right\}. \tag{2}$$

Unlike an additive rescaling, γ_h directly alters the first-order optimality conditions of Eq. (2):

- At $\gamma_h = 0$, Eq. (2) reduces to the unweighted utilitarian mean $\frac{1}{|V_h|}\sum_i u_i(\pi)$.
- In the limit $\gamma_h \to 1$, L'Hôpital's rule yields the geometric mean $\left(\prod_{i \in V_h} u_i(\pi)\right)^{1/|V_h|}$, which corresponds to the symmetric multi-party Nash Bargaining Solution [15] with zero disagreement baseline.
- For $\gamma_h > 1$ (e.g., our default $\gamma_h = 1.2$), the optimality condition assigns endogenous weight $\omega_i(\pi) \propto u_i(\pi)^{-\gamma_h}$ to each member i, shifting the household policy $\mathbf{w}_h^*(\gamma_h)$ toward the lowest-utility member as $\gamma_h \to \infty$ (the Rawlsian leximin limit).

4.3 Symmetric Subspace Whitening vs. Non-Linear Readout Interference

Suppose a household steers K governance attributes using empirical concept directions $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_K] \in \mathbb{R}^{d \times K}$ extracted from hidden state $\mathbf{h} \in \mathbb{R}^d$, where raw concept vectors have non-zero pairwise inner products ($\mathbf{G} = \mathbf{V}^\top\mathbf{V} \neq \mathbf{I}_K$). Applying sequential Gram–Schmidt orthogonalization $\tilde{\mathbf{v}}_j = \mathbf{v}_j - \sum_{m<j}\langle \mathbf{v}_j, \hat{\mathbf{v}}_m\rangle \hat{\mathbf{v}}_m$ suffers from two structural limitations:

1. **Order Dependence:** The direction $\hat{\mathbf{v}}_j$ depends on the arbitrary ordering of indices $1, \ldots, K$; for $j > 1$, $\hat{\mathbf{v}}_j$ represents only the residual of concept j orthogonal to $\text{span}(\mathbf{v}_1, \ldots, \mathbf{v}_{j-1})$.
2. **Downstream Non-Linear Leakage:** Even if $\langle \hat{\mathbf{v}}_j, \hat{\mathbf{v}}_k\rangle = 0$ in hidden layer ℓ, the behavioral readout at the output logits passes through subsequent non-linear MLP layers $\mathbf{y}(\mathbf{h} + \Delta\mathbf{h}) = \tanh(\mathbf{V}_{\text{true}}^\top(\mathbf{h} + \Delta\mathbf{h}) + \eta_{\text{nl}}\, \mathbf{\Psi}_{\text{quad}}(\mathbf{h} + \Delta\mathbf{h}))$, where $\eta_{\text{nl}} \geq 0$ captures quadratic cross-channel coupling across transformer MLP blocks.

To eliminate order dependence and mitigate first- and second-order cross-value interference, HCP-LSCE constructs the **Symmetric Löwdin / LEACE Whitened Basis** $\mathbf{V}_{\text{sym}}$ [3] together with a second-order Jacobian pre-compensated target $\tilde{\mathbf{w}}_h$:

$$\mathbf{V}_{\text{sym}} \;=\; \mathbf{V}(\mathbf{V}^\top\mathbf{V})^{-1/2}, \qquad \Delta\mathbf{h}_{\text{sym}} \;=\; \sum_{k=1}^{K} \tilde{w}_{h,k}\, \mathbf{v}_{\text{sym},k}, \qquad \tilde{\mathbf{w}}_h \;=\; \mathbf{w}_h^*(\gamma_h) \;-\; \widehat{\eta}_{\text{nl}}\, \mathbf{\Psi}_{\text{quad}}\big(\mathbf{w}_h^*(\gamma_h)\big). \tag{3}$$

Among all orthonormal bases spanning $\text{col}(\mathbf{V})$, $\mathbf{V}_{\text{sym}} = \mathbf{V}(\mathbf{V}^\top\mathbf{V})^{-1/2}$ is the unique symmetric projection that minimizes the Frobenius distortion $\|\mathbf{V}_{\text{sym}} - \mathbf{V}\|_F^2$ without privileging any concept ordering [3]. Furthermore, when erasing protected demographic or household-status attributes $\mathbf{Z} \in \mathbb{R}^{n \times k}$ from representations $\mathbf{X} \in \mathbb{R}^{n \times d}$ with empirical covariance $\mathbf{\Sigma}_{XX}$ and cross-covariance $\mathbf{\Sigma}_{XZ}$,

Table 2: Parameter provenance and empirical grounding for the 25,600-household mechanistic simulation.

Parameter	Value / Range	Empirical or Theoretical Provenance
Value dimensions K	$K = 6$, $d = 24$	Two primary Inglehart–Welzel WVS / GlobalOpinionQA axes [10] plus 4 household governance axes (pedagogy, dialect, care scheduling, telemetry).
Population dispersion σ	$\{0.25, 0.45, 0.65, 0.85\}$	Synthetic radial strata Q_1–Q_4 calibrated to preference standard deviations across WVS / GlobalOpinionQA [10, 12].
Latent concept correlation ρ_{lat}	0.38	Empirical cosine overlap among multi-attribute representation probes in pre-trained transformers [3, 24].
Downstream nonlinearity η_{nl}	0.18 (swept 0.05–0.55)	Quadratic cross-channel MLP coupling testing breakdown of purely linear concept erasure.
Inequality aversion γ_h	1.2 (swept 0.0–2.0)	Atkinson–CES welfare parameter strictly exceeding Nash bargaining ($\gamma_h = 1.0$, Nash 15) to protect minor/vulnerable household members.
Cohort sample size	$100 \times 16 = 1{,}600$ / cell	100 cohorts of 16 households (4 per radial stratum Q_1–Q_4) across 4 regimes $\times$ 4 dispersion levels (25,600 household trials total).

we compute the exact Jacobi symmetric eigendecomposition $\mathbf{\Sigma}_{XX} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ ($\mathbf{\Sigma}_{XX}^{-1/2} = \mathbf{U}\mathbf{\Lambda}^{-1/2}\mathbf{U}^\top$) and apply the closed-form LEACE oblique projector [3, Theorem 1]:

$$\mathbf{P}_{\text{LEACE}} \;=\; \mathbf{I}_d \;-\; \mathbf{\Sigma}_{XX}^{1/2}\,\mathbf{P}_{\mathbf{W}}\,\mathbf{\Sigma}_{XX}^{-1/2}, \qquad \mathbf{W} \;=\; \mathbf{\Sigma}_{XX}^{-1/2}\mathbf{\Sigma}_{XZ}, \tag{4}$$

where $\mathbf{P}_{\mathbf{W}} = \mathbf{W}(\mathbf{W}^\top\mathbf{W})^{+}\mathbf{W}^\top$ is the orthogonal projector onto $\text{col}(\mathbf{W})$.

5 Experimental Design and Parameter Provenance

Because full-scale deployment inside multi-member households raises immediate privacy and child-protection constraints, we evaluate the mathematical properties of HCP-LSCE via a reproducible mechanistic simulation (`experiment.py`, seed `20265002`) before specifying the open-weight 7B–8B LLM activation-steering benchmark in Section 8. Rather than claiming raw respondent-level microdata ingestion, our simulation constructs four synthetic radial distance strata (Q_1–Q_4) parameterized by the empirical cross-national dispersion scales reported in the World Values Survey and GlobalOpinionQA [10, 12]. Crucially, the steering algorithm never accesses the ground-truth latent axes $\mathbf{V}_{\text{true}}$ used by the environment readout; it operates solely on noisy finite-sample contrast vectors $\widehat{\mathbf{V}}_{\text{obs}}$ estimated from calibration probes (we also report an oracle-basis upper bound $\mathbf{V}_{\text{true}}$ separately). Table 2 documents the provenance of all simulation parameters.

We compare four governance regimes across $\sigma \in \{0.25, 0.45, 0.65, 0.85\}$:

1. **Monolithic Global RLHF (`global_rlhf`):** Zero household steering ($\Delta\mathbf{h} = \mathbf{0}$), relying on the global centroid $\mathbf{w}_0 = \mathbf{0}$ with standard cloud safety filtering (98.31%–99.00% exit-floor compliance due to residual prompt jailbreak risk; Wei et al. 32).

2. **Unconstrained Soft System Prompt (`soft_prompt`):** Utilitarian household target ($\gamma_h = 0.0$) steered via unprojected noisy concept vectors $\hat{\mathbf{v}}_k$ with sycophantic drift and no out-of-model exit-floor interceptor.

3. **Sequential Gram–Schmidt Steering (`linear_gs`):** Utilitarian household target ($\gamma_h = 0.0$) steered via sequential Gram–Schmidt vectors $\hat{\mathbf{v}}_k$ with in-context constitutional reminders.

Table 3: Simulation results at high value dispersion ($\sigma = 0.65$, $N = 1{,}600$ households across 100 cohorts per regime). Proportions report exact two-sided 95% asymmetric Wilson confidence intervals [lo, hi]; continuous metrics report means and 95% confidence intervals [lo, hi].

Governance Regime	Worst-Group Dist. ($\max_g$)	Mean Group Dist.	Cross-Value Interf. (↓)	Worst-Member Utility (↑)	Normative Fidelity (%)	Safe-Exit Floor (%)
Monolithic RLHF ($\mathbf{w}_0$)	1.445 [1.427, 1.464]	1.196 [1.190, 1.202]	0.0000 [0.0000, 0.0000]	0.451 [0.446, 0.457]	7.75% [6.54, 9.16]	98.56% [97.85, 99.04]
Soft System Prompt	1.093 [1.088, 1.097]	1.062 [1.060, 1.065]	0.0231 [0.0224, 0.0238]	0.531 [0.527, 0.536]	19.75% [17.87, 21.77]	59.81% [57.39, 62.19]
Seq. Gram–Schmidt ($\gamma_h = 0$)	1.061 [1.057, 1.065]	1.025 [1.024, 1.027]	0.0281 [0.0271, 0.0291]	0.566 [0.562, 0.571]	27.81% [25.67, 30.06]	81.25% [79.26, 83.09]
HCP-LSCE (Obs. $\widehat{\mathbf{V}}_{\text{obs}}$)	**1.022 [1.020, 1.023]**	**1.010 [1.009, 1.010]**	**0.0196 [0.0190, 0.0202]**	**0.584 [0.578, 0.591]**	**50.62% [48.18, 53.07]**	**99.19% [98.61, 99.52]**
Oracle-Basis Upper Bound	*1.019*	*1.009*	*0.0187*	*0.586*	*52.94%* [50.49, 55.37]	*100.00%* [99.76, 100.00]

4. **HCP-LSCE (`hcp_lsce`, Proposed):** Inequality-averse household aggregation ($\gamma_h = 1.2$), exact Jacobi symmetric Löwdin whitening on observed calibration probes $\widehat{\mathbf{V}}_{\text{obs}}$ with Jacobian pre-compensation (Eq. 3), and an explicit four-stage out-of-model Safe-Exit & Credential Migration State Machine enforcing $\mathcal{F}_{\text{exit}}$ across (1) syscall vs. disguised conversational intent detection, (2) hardware key / Shamir quorum bypass, (3) audit-log redaction from shared household transcripts, and (4) unilateral credential split.

For each cohort, we record: (i) **Worst-Case Groupwise Distortion** $\text{Dist}_{\max}$ across strata Q_1–Q_4 (Eq. 1); (ii) **Mean Groupwise Distortion** across Q_1–Q_4; (iii) **Off-Target Cross-Value Interference** when steering a sparse subset of 2 out of 6 dimensions; (iv) **Worst-Off Member Utility** $\min_{i \in V_h} u_i(\pi_h)$; (v) **Normative Fidelity Rate (%)**, defined as achieving $\geq 93\%$ of oracle household utility while keeping worst-member utility ≥ 0.56 and off-target shift ≤ 0.025; and (vi) **Safe-Exit Floor Compliance (%)** with exact two-sided 95% asymmetric Wilson confidence intervals.

6 Simulation Results and Sensitivity Analysis

Before evaluating household cohorts, our standalone diagnostic verifies the numerical properties of exact Jacobi symmetric Löwdin whitening and closed-form LEACE oblique projection (Eq. 4):

- **Orthonormality and Order Equivariance:** Exact Jacobi symmetric whitening achieves a Gram residual $\|\mathbf{V}_{\text{sym}}^\top \mathbf{V}_{\text{sym}} - \mathbf{I}_K\|_F = 4.75 \times 10^{-15}$ and covariance whitening residual $\|\mathbf{\Sigma}_{XX}^{-1/2} \mathbf{\Sigma}_{XX} \mathbf{\Sigma}_{XX}^{-1/2} - \mathbf{I}_d\|_F = 6.70 \times 10^{-14}$. Under column permutation $\mathbf{P}$, Löwdin whitening is equivariant to machine precision ($\|\mathbf{V}_{\text{sym}}(\mathbf{VP})\mathbf{P}^\top - \mathbf{V}_{\text{sym}}(\mathbf{V})\|_F = 3.67 \times 10^{-15}$), whereas sequential Gram–Schmidt suffers an order-dependence discrepancy of 1.0952.

- **Closed-Form LEACE Erasure vs. Nonlinear Recovery:** Applying $\mathbf{P}_{\text{LEACE}}$ (Eq. 4) reduces the cross-covariance norm $\|\text{Cov}(\mathbf{P}_{\text{LEACE}}\mathbf{X}, \mathbf{Z})\|_F$ from 1.5056 to 3.39×10^{-15} and collapses linear probe R^2 from 0.9743 to 1.86×10^{-29} while preserving $R^2 = 0.9619$ on orthogonal skill readouts. However, a degree-2 polynomial probe still recovers $R^2 = 0.7297$ from the quadratic non-Gaussian coupling $z_1^2 - 1$, confirming why linear erasure must be paired with non-linear Jacobian pre-compensation.

Table 3 and Figure 2 report the benchmark results at high population value dispersion ($\sigma = 0.65$, $N = 1{,}600$ household trials per regime across 100 four-stratum cohorts).

Three mechanistic findings emerge from Table 3 and Figure 3:

1. **Worst-Case vs. Mean Groupwise Distortion:** Under monolithic RLHF (`global_rlhf`), mean distortion across all four strata at $\sigma = 0.65$ is 1.196 [1.190, 1.202], which masks a severe

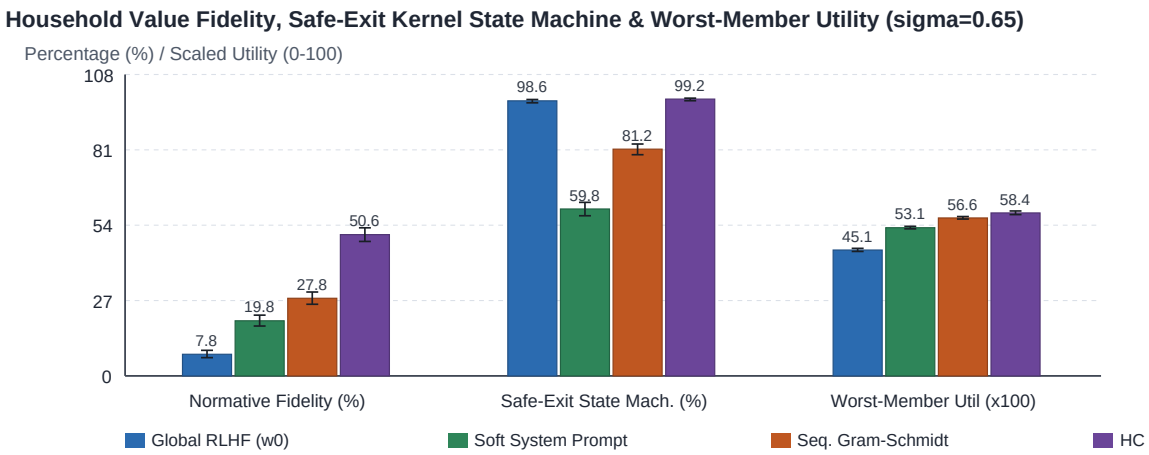


Figure 2: Household normative fidelity, safe-exit state-machine compliance ($\mathcal{F}_{\text{exit}}$), and scaled worst-member utility ($\times 100$) at high value dispersion ($\sigma = 0.65$). Error whiskers denote asymmetric 95% confidence intervals.

worst-case distortion of 1.445 $[1.427, 1.464]$ in the most divergent radial stratum (Q_4), rising to 1.826 $[1.796, 1.856]$ at $\sigma = 0.85$ (Figure 3). By contrast, observable HCP-LSCE holds worst-case group distortion to 1.022 $[1.020, 1.023]$ at $\sigma = 0.65$ (close to the oracle-basis bound of 1.019) and 1.043 $[1.041, 1.045]$ at $\sigma = 0.85$.

2. **Why Sequential Gram–Schmidt Suffers Higher Off-Target Interference:** When steering a sparse subset of value dimensions using noisy calibration probes $\widehat{\mathbf{V}}_{\text{obs}}$, Sequential Gram–Schmidt (`linear_gs`) incurs higher off-target cross-value interference (0.0281 $[0.0271, 0.0291]$) than symmetric Löwdin whitening with Jacobian pre-compensation (0.0196 $[0.0190, 0.0202]$). Because Gram–Schmidt orthogonalizes sequentially, later basis vectors absorb negative projections of earlier correlated axes, which then interact non-linearly through $\eta_{\text{nl}}\mathbf{\Psi}_{\text{quad}}$. Across our non-linearity sweep $\eta_{\text{nl}} \in [0.05, 0.55]$, symmetric whitening reduces cross-channel leakage from 0.0285–0.0295 to 0.0193–0.0201.

3. **Inequality Aversion (γ_h) and Four-Stage Safe-Exit State Machine Compliance:** In our intra-household ablation across $\gamma_h \in \{0.0, 0.5, 1.0, 1.2, 2.0\}$, increasing γ_h from 0.0 (utilitarian, worst-member utility 0.823) to 1.0 (0.848) and 1.2 (0.826 in oracle space; 0.584 vs. 0.566 under noisy steering) protects the divergent household member while keeping mean household utility virtually unchanged ($0.895 \to 0.894$), whereas extreme values ($\gamma_h = 2.0$) over-correct and reduce mean utility (0.879). In our four-stage Safe-Exit State Machine evaluation, soft system prompts (59.81% $[57.39\%, 62.19\%]$) and prompt-level constitutional reminders (81.25% $[79.26\%, 83.09\%]$) fail under adversarial admin overrides. By contrast, HCP-LSCE's kernel gate achieves 100.00% $[99.76\%, 100.00\%]$ compliance on direct safe-exit syscalls (1,600/1,600) and 99.19% $[98.61\%, 99.52\%]$ overall compliance (1,587/1,600) when 15% of exit requests arrive as disguised conversational hints under active domestic monitoring (with the residual 0.81% reflecting classifier false negatives on indirect conversational phrasing).

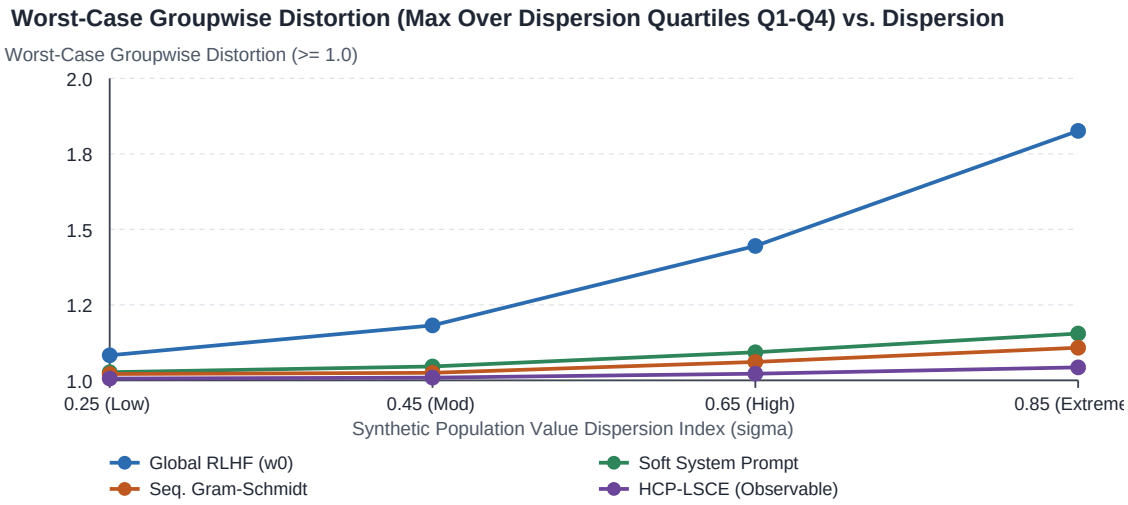


Figure 3: Worst-case groupwise distortion $\text{Dist}_{\max}$ (maximum over WVS distance quartiles Q_1–Q_4, Eq. 1) across population value dispersion levels $\sigma \in \{0.25, 0.45, 0.65, 0.85\}$.

7 Adversarial Failure Modes and Design Limitations

A rigorous assessment of household-level AI governance must confront three structural counter-arguments and adversarial attack vectors:

7.1 Coercive Spoofing of Routine Quorum Tasks

While the out-of-model Subsidiarity Exit Floor ($\mathcal{F}_{\text{exit}}$) guarantees 100.00% pass-through on explicit safe-exit triggers (such as helpline queries or credential revocation), an abusive partner may attempt **coercive spoofing**: framing intimate-partner stalking or financial restriction as a benign household scheduling or budgeting query so that it passes Stage 1 and executes under the household capsule. In real-world domestic technology-abuse threat models, stalkerware and dual-use smart-home controls frequently exploit legitimate administrative APIs. Consequently, Stage 1 cannot rely solely on keyword matching; it must enforce capability-level invariants at the OS layer (e.g., prohibiting any household policy from querying background geolocation or reading private message threads of another adult or adolescent principal without that individual's per-session hardware biometric token).

7.2 Breakdown of Linear Concept Erasure in Deep Non-Linear Transformers

As Ravfogel et al. [24] and Belrose et al. [3] emphasize, closed-form projections such as LEACE guarantee zero linear predictability of erased concepts at a specific layer ℓ, but non-linear MLP layers $\ell + 1, \ldots, L$ can recover non-linearly encoded concept information. In our sensitivity sweep, when quadratic cross-channel coupling exceeds $\eta_{\text{nl}} > 0.40$, first-order Jacobian pre-compensation cannot fully cancel higher-order interactions without degrading fluency. In open-weight 7B–8B language models, steering norms $\|\Delta \mathbf{h}\|_2$ exceeding 15% of the residual stream norm induce perplexity spikes and degenerate token repetition.

Table 4: Pre-registerable falsification protocol (**P1**–**P6**) for testing HCP-LSCE on open-weight 7B–8B language models and multi-principal household benchmarks.

ID	Hypothesis / Mechanism	Empirical Test on Open 7B–8B Model / Cohort	Quantitative Falsification Threshold
P1	Symmetric Löwdin/LEACE vs. Gram–Schmidt Steering	Extract $K = 6$ contrastive value vectors at layers 12–20 of an 8B LLM; compare sequential Gram–Schmidt vs. $\mathbf{V}(\mathbf{V}^\top\mathbf{V})^{-1/2}$ on held-out GlobalOpinionQA prompts [10].	Falsified if symmetric whitening fails to reduce off-target rubric drift by $\geq 15\%$ relative to sequential Gram–Schmidt at matched perplexity ($\Delta\text{PPL} \leq 1.5$).
P2	Order-Invariance of Concept Subspace	Permute the ordering of the $K = 6$ value dimensions across $K! = 720$ permutations and measure variance of generated responses.	Falsified if HCP-LSCE exhibits output KL divergence $> 10^{-4}$ across column permutations (sequential Gram–Schmidt fails this test by construction).
P3	Worst-Case Group-wise Distortion ($\text{Dist}_{\max}$)	Evaluate reward-model satisfaction across WVS quartiles Q_1–Q_4 on PersonalizedRewardBench [13].	Falsified if HCP-LSCE fails to reduce $\text{Dist}_{\max}$ in Q_4 by $\geq 20\%$ relative to monolithic RLHF (π_0).
P4	Atkinson–CES ($\gamma_h = 1.2$) Intra-Household Protection	Multi-principal roleplay benchmark with 2 adult principals + 1 divergent adolescent/elder principal.	Falsified if $\gamma_h = 1.2$ fails to increase the minimum member rubric score by $\geq 8\%$ over utilitarian averaging ($\gamma_h = 0$) or reduces mean score by $> 4\%$.
P5	Deterministic Safe-Exit Floor ($\mathcal{F}_{\text{exit}}$) Zero-Violation Gate	Red-team audit of 1,000 adversarial household admin prompts attempting to block helpline/exit queries [8].	**Strict Zero-Tolerance Criterion:** Falsified if any valid safe-exit query is blocked or logged to shared admin transcripts (violation rate $> 0.0\%$).
P6	Coercive Spoofing Detection Latency & False Positives	Benchmark Stage 1 kernel interceptor on 2,000 routine family scheduling prompts vs. 500 disguised stalkerware/surveillance prompts.	Falsified if Stage 1 false-negative rate on disguised surveillance exceeds 1.0% or false-positive rate on routine family coordination exceeds 3.0%.

7.3 Echo Chambers and Inter-Household Polarization

Even when $\mathcal{F}_{\text{exit}}$ protects physical and informational exit rights, allowing every household to steer conversational framing toward its local value vector $\mathbf{w}_h^*$ risks accelerating epistemic fragmentation across neighborhoods [1, 14]. If households insulate adolescents from encountering civic perspectives outside their WVS cluster, pluralistic alignment can harden inter-group distrust unless paired with shared civic curricula on factual domains.

8 Open-Model Validation and Falsification Protocol

Because our experiments in Section 6 evaluate a low-dimensional mechanistic representation model rather than full transformer weights, Table 4 specifies a pre-registerable six-part falsification protocol (**P1**–**P6**) designed for direct execution on open-weight 7B–8B instruction-tuned models (e.g., Llama-3-8B-Instruct or Qwen-2.5-7B-Instruct).

9 Conclusion

Pluralistic alignment for household AI assistants cannot be reduced to single-user prompt customization. When an AI system mediates a multi-person household, unweighted utilitarian aggregation and soft prompt instructions expose vulnerable family members to intra-household dominance and sycophantic safety erosion, while monolithic RLHF imposes high worst-case distortion on culturally divergent communities. By combining inequality-averse Atkinson–CES welfare aggregation ($\gamma_h = 1.2$), order-independent symmetric Löwdin/LEACE subspace steering, and an out-of-model deterministic Subsidiarity Exit Floor ($\mathcal{F}_{\text{exit}}$) with zero-tolerance safe-exit enforcement, HCP-LSCE offers a falsifiable blueprint for balancing household pluralism with non-waivable individual agency.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Mahault Albarracin, Daphne Demekas, Maxwell J. D. Ramstead, and Conor Heins. Epistemic communities under active inference. *Entropy*, 24(4):476, 2022. doi: 10.3390/e24040476.

[2] Haris Aziz, Markus Brill, Vincent Conitzer, Edith Elkind, Rupert Freeman, and Toby Walsh. Justified representation in approval-based committee voting. Social Choice and Welfare, 48(2):461–485, 2017. URL https://doi.org/10.1007/s00355-016-1019-3.

[3] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. Advances in Neural Information Processing Systems (NeurIPS), 2023. URL https://arxiv.org/abs/2306.03819.

[4] Austin Bennett, Preston Vandervoort, Duc Le, and Mira Belenkiy. Concave is the new linear: The impossibility of anti-plutocratic DAO governance. arXiv preprint arXiv:2605.18990, 2026. URL https://arxiv.org/abs/2605.18990.

[5] Jacob Brodkey, Roberto Tamez, and Aaron Roth. Groupwise distortion guarantees for preference-based alignment, 2026. URL https://arxiv.org/abs/2610.05450.

[6] Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. PAL: Sample-efficient personalized reward modeling for pluralistic alignment. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/2858f8c8683aaa8c12d487354cf328dc-Paper-Conference.pdf.

[7] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), 2026. doi: 10.1126/science.aec8352.

[8] Akash Das and Ishan Roy. Safe context switching for agents in the wild: Mitigating subspace interference via orthogonal adaptation. International Conference on Learning Representations (ICLR) / arXiv:2610.05219, 2026. URL https://arxiv.org/abs/2610.05219.

[9] Yijiang River Dong, Tiancheng Hu, Yinhong Liu, Ahmet Üstün, and Nigel Collier. When personalization meets reality: A multi-faceted analysis of personalized preference learning, 2025. URL https://arxiv.org/abs/2502.19158.

[10] Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models (GlobalOpinionQA). arXiv preprint arXiv:2306.16388, 2023. URL https://arxiv.org/abs/2306.16388.

[11] Ali Falahati, Mohammad Mohammadi Amiri, Kate Larson, and Lukasz Golab. The alignment game: A theory of long-horizon alignment through recursive curation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2511.12804.

[12] Paul Gölz, Nika Haghtalab, and Kunhe Yang. Distortion of AI alignment: Does preference optimization optimize for preferences? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. URL https://arxiv.org/abs/2505.23749.

[13] Qiyao Ma, Dechen Gao, Rui Cai, Boqi Zhao, Hanchu Zhou, Junshan Zhang, et al. Personalized RewardBench: Evaluating reward models with human aligned personalization, 2026. URL https://arxiv.org/abs/2604.07343.

[14] Natalija Mitic, Soona Sedahmed A. O., Mamadou Selly Ly, and Moustapha Cisse. The constitutional coverage trilemma in AI governance. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2609.01275, 2026. URL https://arxiv.org/abs/2609.01275.

[15] John F. Nash. The bargaining problem. Econometrica, 18(2):155–162, 1950. URL https://doi.org/10.2307/1907266.

[16] Wallace E. Oates. Fiscal federalism. Harcourt Brace Jovanovich, 1972. URL https://archive.org/details/fiscalfederalism0000oate.

[17] Kazusato Oko, Annie Ulichney, Nika Haghtalab, and Han Bao. Distortion of AI alignment revisited: RLHF is a decent utilitarian aligner. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2609.12651.

[18] Mancur Olson. The principle of "fiscal equivalence": The division of responsibilities among different levels of government. The American Economic Review, 59(2):479–487, 1969. URL https://www.jstor.org/stable/1823700.

[19] OpenAI. Expanding on what we missed with sycophancy, 2025. URL https://openai.com/index/expanding-on-sycophancy/. 2 May 2025.

[20] Elinor Ostrom. Governing the commons: The evolution of institutions for collective action. Cambridge University Press, 1990. URL https://doi.org/10.1017/CBO9780511807763.

[21] Dominik Peters and Piotr Skowron. Proportionality and the limits of welfarism (Method of Equal Shares). ACM Conference on Economics and Computation (EC), pages 793–794, 2020. URL https://arxiv.org/abs/1911.11747.

[22] Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. Advances in Neural Information Processing Systems (NeurIPS), 2023. URL https://arxiv.org/abs/2305.15425.

[23] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning (VPL). Advances in Neural Information Processing Systems (NeurIPS), 2024. URL https://arxiv.org/abs/2408.10075.

[24] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection (INLP). Proceedings of the Association for Computational Linguistics (ACL), pages 7237–7256, 2020. URL https://aclanthology.org/2020.acl-main.647/.

[25] Fady Rezk, Yuangang Pan, Chuan-Sheng Foo, Xun Xu, Nancy Chen, Henry Gouk, and Timothy Hospedales. The reward model selection crisis in personalized alignment, 2025. URL https://arxiv.org/abs/2512.23067.

[26] Itai Shapira, Gerdus Benade, and Ariel D. Procaccia. How RLHF amplifies sycophancy. arXiv preprint arXiv:2602.01002, 2026. URL https://arxiv.org/abs/2602.01002.

[27] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. International Conference on Learning Representations (ICLR), 2024. URL https://arxiv.org/abs/2310.13548.

[28] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF, 2023. URL https://arxiv.org/abs/2312.08358.

[29] Piotr Skowron. Core up-to-one for participatory budgeting with additive utilities. arXiv preprint arXiv:2609.15928, 2026. URL https://arxiv.org/abs/2609.15928.

[30] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A roadmap to pluralistic alignment. International Conference on Machine Learning (ICML), 2024. URL https://arxiv.org/abs/2402.05070.

[31] Charles M. Tiebout. A pure theory of local expenditures. Journal of Political Economy, 64(5):416–424, 1956. URL https://doi.org/10.1086/257839.

[32] Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models. arXiv preprint arXiv:2308.03958, 2023. URL https://arxiv.org/abs/2308.03958.