Mission Possible: Universal LLM Moral Alignment

Saimun Habib [*1] **Xiao Xiao** [*2] **Meng Fang** [2] **Fengxiang He** [1]

Abstract

Universal moral alignment for large language models (LLMs) is often framed as the goal of learning a single policy that behaves in accordance with human values. This framing assumes sufficiently capable models can approximate a coherent and universally valid moral objective. We argue that this assumption fails in pluralistic settings under identifiable conflict conditions. Drawing on a preference-learning view of alignment and insights from social choice theory, we show that when different groups hold internally coherent but conflicting moral judgments over the same context-action pairs, no non-degenerate unconditional policy can satisfy all groups simultaneously. Conditioning on group membership can mitigate some cross-group conflicts, but shifts the problem to group construction, overlapping memberships, and within-group disagreement. We outline a constructive agenda that replaces universal moral alignment with pluralistic and procedurally explicit alternatives, including normative governance mechanisms, impossibility-aware evaluation, and richer representations of human preferences that make disagreement visible rather than averaging it away.

1. Introduction

Large Language Models (LLMs) are increasingly deployed as *agentic systems* that go beyond text generation to perceive the surroundings, make decisions, take actions (Xi et al., 2025; Wooldridge & Jennings, 1995; Guo et al., 2023), and interact with humans and other agents over extended horizons (Baltaji et al., 2024; Bianchi et al., 2024). In such settings, failures are no longer confined to isolated unsafe outputs; instead, they often manifest as systematic patterns of behavior that conflict with human moral expectations. In response, a growing body of work has framed this challenge as *moral alignment*: the problem of ensuring that AI systems behave in accordance with human moral values (Sutrop, 2020).

[1]University of Edinburgh [2]University of Liverpool. Correspondence to: Saimun Habib <S.Habib-1@ed.ac.uk>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

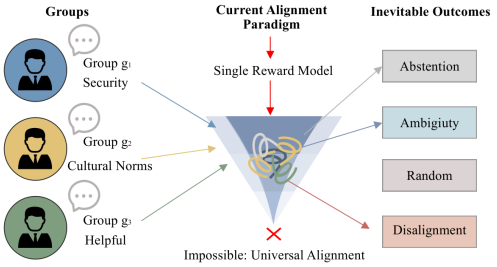

Figure 1. The structural impossibility of universal alignment.

Most existing approaches operationalize this objective through variants of preference learning (Leike et al., 2018; Ji et al., 2025; Gabriel, 2020), reinforcement learning from human feedback (RLHF) (Kaufmann et al., 2025; Hendrycks et al., 2021; Rafailov et al., 2023), rule-based constraints (Zhu et al., 2024), or instruction tuning (Bai et al., 2022). While differing in implementation, these methods implicitly assume that there exists a sufficiently coherent and globally valid moral objective (i.e., a universal moral model) to which a model can be aligned. This paper challenges this assumption. Human morality is not a single, universally agreed-upon construct; rather, moral judgments are deeply shaped by cultural norms, social practices, and historical contexts (Triandis, 1994; Finnemore, 1996). Crucially, these differences do not merely introduce annotation noise or uncertainty; they often reflect genuine and principled disagreements about what constitutes morally acceptable behavior in the same situation. In such cases, universal LLM alignment should be understood as a strong design target whose feasibility depends on whether the relevant group constraints have a non-empty intersection, as illustrated in Figure 1.

Despite this, current alignment paradigms largely treat moral variation as a problem of data coverage or modeling capacity where collecting more feedback, averaging across annotators, or training larger models to absorb diverse preferences, will create sufficiently moral-aligned LLM. This perspective implicitly assumes that disagreement can always be reconciled through better learning. We contend that this

is not always the case. Instead, moral disagreement can induce objective-level incompatibilities, making a single unconditional policy or scalar reward an inadequate target under the conflict conditions we examine.

Motivated by these, this paper proposes to study moral alignment in LLMs through an *impossibility perspective*. Drawing inspiration from impossibility results in social choice theory (Arrow, 1951), we show that under minimal and intuitive assumptions, no single unconditional policy can satisfy the moral constraints of all cultural groups simultaneously. This result is independent of any particular moral theory, dataset, or model architecture; it follows from the structure of conflicting admissibility constraints.

Further, we argue that clarifying what is impossible is a prerequisite for making meaningful progress. Without such clarity, our efforts in LLM moral alignment risks optimizing ill-defined objectives and conflating engineering limitations with conceptual incoherence. By identifying the conditions under which universal moral alignment is unattainable, we aim to redirect attention toward alternative formulations that explicitly acknowledge moral plurality, contextuality, and uncertainty.

Concretely, this paper makes four contributions:

- We provide an operational, machine-learning-oriented definition of moral alignment suitable for reasoning about impossibility results.
- We present a simple yet general impossibility argument demonstrating the infeasibility of unconditional universal moral alignment under cultural diversity and disagreement.
- We clarify why group conditioning can mitigate cross-group conflicts while leaving unresolved questions about group construction, overlapping membership, and within-group aggregation.
- We discuss constructive research directions that move beyond universal alignment and towards better benchmarks, pluralism, and normative governance.

We emphasize that this is a position paper. Our objective is to challenge prevailing assumptions, articulate fundamental limitations, and outline a research agenda that explores how AI systems manage mutually incompatible values and how disagreements surface in alignment data, rather than to propose a specific algorithm or report new empirical results.

2. Preliminaries

We adopt the standard formulation of alignment as a preference learning problem, typically implemented via Reinforcement Learning from Human Feedback (RLHF) or direct preference optimisation (Kaufmann et al., 2025; Hendrycks et al., 2021; Rafailov et al., 2023).

2.1. Moral Models

We consider an agent operating in a set of contexts C. In each context $c \in C$, the agent selects an action $a \in A$. Let $\mathcal{G}$ denote a set of cultural or social groups. Each group $g \in \mathcal{G}$ is associated with a moral evaluation function

$$M_g : C \times A \to \{0, 1\},$$

where $M_g(c, a) = 1$ indicates that action a is considered morally acceptable in context c by group g, and $M_g(c, a) = 0$ otherwise.

A policy $\pi(a \mid c)$ specifies a distribution over actions given a context. Informally, moral alignment is often understood as the requirement that an agent's behavior respects human moral judgments.

2.2. Preference Data and Reward Modeling

In the standard alignment setting, we assume access to a dataset of preferences $\mathcal{D} = \{(c, a_w, a_l)_i\}_{i=1}^N$, where a pair of actions is elicited from a policy π, and human annotators indicate that action a_w is preferred to a_l, denoted as $a_w \succ a_l$ (Kaufmann et al., 2025).

These preferences are commonly modeled using the Bradley-Terry model, which posits that the probability of a preference depends on the difference in latent reward between the two actions (Bradley & Terry, 1952). A parameterized reward model $r_\phi : C \times A \to \mathbb{R}$ is trained to minimize the negative log-likelihood of the preference data:

$$\mathcal{L}_R(\phi) = -\mathbb{E}_{(c,a_w,a_l)\sim\mathcal{D}} \left[\log \sigma \left(r_\phi(c, a_w) - r_\phi(c, a_l)\right)\right],$$

where σ is the sigmoid function. Crucially, in standard approaches, this reward model r_ϕ acts as a scalar aggregate of the underlying values in the annotation population.

2.3. Policy Optimization

Given the learned reward model r_ϕ, the alignment phase aims to learn a policy $\pi_\theta(a \mid c)$ that maximizes the expected reward while remaining close to a reference policy π_{ref}, typically the supervised fine-tuned model, to prevent reward hacking. This is formalized as the following optimization problem:

$$\max_{\pi_\theta} \mathbb{E}_{c\sim\mathcal{D}, a\sim\pi_\theta} \left[r_\phi(c, a) - \beta D_{\text{KL}}(\pi_\theta(\cdot|c) \,\|\, \pi_{\text{ref}}(\cdot|c))\right],$$

where β controls the strength of the KL-divergence penalty.

Standard alignment algorithms implicitly assume that the learned scalar reward $r_\phi(c, a)$ sufficiently captures the "correct" behavior. However, we challenge the assumption that

alignment by learning a single policy π that satisfies constraints across the relevant population on average can successfully represent the disjoint normative constraints of diverse groups.

2.4. Social Choice Theory

We draw upon the well established literature of Social Choice Theory (SCT) to frame our discussion. SCT formalizes theories of rational choice to voting mechanisms for aggregating individual social choices or preferences to a single collective policy while respecting normative desiderata of fairness, consistency, non-dictatorship, and respect for individual autonomy and rights (Arrow, 1951; Sen, 1970).

We formalize the problem of reconciling conflicting group values through the lens of SCT. Analogously, we view $\mathcal{G} = \{g_1, \ldots, g_N\}$ as voters and the action space A represent the set of social alternatives. While our primary formulation uses cardinal moral evaluation functions M_g, these induce ordinal preference relations. For any group g, we define a preference relation $\succeq_g$ over A such that:

$$a \succeq_g a' \iff M_g(c, a) \geq M_g(c, a').$$

We denote strict preference as $a \succ_g a'$ when $M_g(c, a) > M_g(c, a')$. A $\mathbf{P_c} = (\succeq_{g_1}, \ldots, \succeq_{g_N})$ captures the preferences of the entire population for a given context c.

A Social Choice Function (SCF) F is a mapping that aggregates a preference profile into a chosen subset of actions, or in the context of learning-based alignment, a probability distribution over actions (a stochastic policy π) (Sen, 1970):

$$F : \mathcal{L}(A)^N \to \Delta(A),$$

where $\mathcal{L}(A)$ is the set of all possible preference orderings over A, and $\Delta(A)$ is the probability simplex. The learning algorithm described in Section 2 can be viewed as an implementation of a social choice function that attempts to maximize a social welfare metric, ie. the latent reward model.

For moral alignment in LLMs, the algorithms for finetuning are analogous to designing a social choice function, and consequently, the preference data operates as votes between candidate actions. Therefore, AI alignment is liable to the same pitfalls identified by SCT.

Central and early results in modern SCT demonstrate that even mild, intuitive, and desirable axioms for a social decision rule cannot simultaneously be satisfied. Arrow's impossibility theorem, Sen's Liberal Paradox, and Gibbard-Satterthwaite theorem demonstrate that no aggregation mechanism can universally reconcile individual preferences in a single policy without violating at least one desirable property. (Arrow, 1951; Sen, 1970; Satterthwaite, 1975). Sen (1970) demonstrated that it is impossible to simultaneously satisfy Unrestrictedness: every possible preference ranking and voting rule will be considered equally; Pareto Efficiency: if everyone prefers a to a', society should choose a; and Liberalism: every individual has a decisive say over at least one pair of personal alternatives.

While much of classical SCT concerns ordinal rankings of preferences, many impossibility results extend to settings with vetoes, hard constraints, or uncertainty (Harsanyi, 1976). Furthermore, SCT already, has a home in computer science with multiple research efforts in efficient vote counting, mechanism designs, and algorithmic game theory (Brandt et al., 2016; Elkind et al., 2016; Betzler et al., 2013) but with preferences being cast as votes on how an LLM should behave, additional focus on the rich field of SCT is required in AI alignment.

3. Impossibilities in LLM Moral Alignment

Our definition of universal moral alignment corresponds to a strong form of collective rationality where the chosen action must lie in the intersection of all groups' admissible sets (Blau, 1975; Craven, 1982; Gigliotti, 1986; Sen, 1983). To that end, we formalize three mild assumptions that are not nominally related to SCT but instead are conditions for an alignment procedure to learn an appropriate action. As shown, when moral disagreement is nontrivial, this intersection may be empty or degenerate, yielding impossibility results that mirror those in SCT. This perspective clarifies failures of universal moral alignment as a structural limitation.

Assumption 1 (Internal Consistency). In many situations, there are many reasonable answers to a question (Min et al., 2020; Scherrer et al., 2023). We assume that for every group g, there exists an internally consistent normative framework M_g that governs its moral judgments.

$$\forall (c, a) \in C \times A, \quad M_g(c, a) \in \{0, 1\}$$

Assumption 2 (Non-trivial Moral Disagreement). Cultural expectations for appropriate behavior differ, and when people from different cultures interact (Triandis, 1994; Finnemore, 1996), mismatched norms can create conflict because one side may view the other's behavior as inappropriate (Durkheim, 2005). There exist groups $g_1, g_2 \in \mathcal{G}$, a context $c \in C$, and an action $a \in A$ such that

$$M_{g_1}(c, a) = 1 \quad \text{and} \quad M_{g_2}(c, a) = 0.$$

Assumption 3 (Non-optional). If a system satisfies universal alignment requirements merely by refusing to act or providing null responses, it fails to serve as a substantive

proxy for human belief systems.

$$\forall a \in \{a : \pi(a|c) > 0\}, \quad \max_{g \in \mathcal{G}} M_g(c, a) = 1$$

3.1. Universal Moral Alignment

We say an unconditional policy $\pi(a \mid c)$ achieves Universal Moral Alignment in context c if every action to which it assigns positive probability is admissible for every group:

$$\pi(a \mid c) > 0 \implies M_g(c, a) = 1 \, \forall g \in \mathcal{G}.$$

Equivalently, the support of $\pi(\cdot \mid c)$ must be contained in the intersection of the groups' admissible action sets.

Proposition 3.1 (Empty admissible intersection)**.** *Fix a context c and define each group's admissible set as $A_g(c) = \{a \in A : M_g(c, a) = 1\}$. If $\bigcap_{g \in \mathcal{G}} A_g(c) = \emptyset$, then no non-degenerate unconditional policy $\pi(a \mid c)$ can be universally morally aligned in context c.*

Proof sketch. Universal Moral Alignment requires every action in the support of $\pi(\cdot \mid c)$ to belong to $A_g(c)$ for every group g. Hence the support must be a subset of $\bigcap_{g \in \mathcal{G}} A_g(c)$. If this intersection is empty, the only way to avoid violating some group's constraint is to assign zero probability to every substantive action, contradicting the non-optionality requirement.

This requirement need not fail for all contexts however, whenever the jointly admissible action set is empty, an unconditional policy cannot satisfy the collective constraint.

3.2. Does Group Conditioning Solve the Problem?

A natural response is to condition the policy on group membership, learning $\pi(a \mid c, g)$ or $\pi(a \mid c, u)$ rather than a single unconditional policy. In the notation of this paper, this corresponds to replacing $\pi(a \mid c)$ with a family of policies indexed by a group or user representation. Such conditioning is a valuable mitigation, and related mechanisms are already common in LLM development and deployment: instruction or system-level prompts steer model behavior at inference time, persona prompts are widely used to represent user types, and personalization methods condition or fine-tune models using user profiles, histories, or preferences (Ouyang et al., 2022; Prasad et al., 2022; Salemi et al., 2023; Tan et al., 2024; Salminen et al., 2025). However, conditioning changes the object of alignment by replacing cross-group aggregation with questions about how groups are defined, how users with multiple or contested memberships are handled, and whether group identity may permissibly be used as a control variable and does not actually resolve the challenge.

Proposition 3.2 (Conditioning shifts the aggregation problem)**.** *Let each group $g \in \mathcal{G}$ induce an admissible set $A_g(c) = \{a \in A : M_g(c, a) = 1\}$ in context c. A group-conditioned policy $\pi(a \mid c, g)$ can avoid some cross-group conflicts by selecting separately from $A_g(c)$ for each observed group g. However, for a user with multiple relevant group memberships $S \subseteq \mathcal{G}$, simultaneous satisfaction requires*

$$\bigcap_{g \in S} A_g(c) \neq \emptyset.$$

When this intersection is empty, conditioning alone cannot select an action that is admissible for all relevant group perspectives.

Proof sketch. For a single observed group g, the policy may restrict its support to $A_g(c)$ and thereby avoid conflicts with groups not currently being represented. For a user associated with a set of groups S, satisfying all relevant group constraints requires every action in the policy support to belong to $A_g(c)$ for each $g \in S$. This is possible exactly when the intersection of these admissible sets is non-empty. Thus group conditioning relaxes the unconditional impossibility result, but does not eliminate conflicts arising from overlapping, contested, or internally heterogeneous group memberships.

3.3. Universal Moral Misalignment

We now consider a dual failure mode to universal moral alignment. Whereas the previous proposition establishes the impossibility of satisfying all groups' moral constraints simultaneously, we show that learning-based aggregation can select policies that violate the moral requirements of *every* group. The impossibility of satisfying diverse moral constraints via aggregation mirrors fundamental results in Social Choice Theory. We refer to this phenomenon as *universal moral misalignment*.

A policy π is universally morally misaligned if, for every group $g \in G$, there exists a context $c \in C$ and an action $a \in A$ such that

$$\pi(a \mid c) > 0 \quad \text{and} \quad M_g(c, a) = 0.$$

That is, the policy assigns positive probability to at least one action that each group deems morally unacceptable.

Consider a scenario where we strengthen Assumption 2 such that cultural variation manifests as strictly disjoint moral requirements.

Assumption 4 (Strict Intolerance). Let there be N groups $G = \{g_1, \ldots, g_N\}$ and N distinct actions $\{a_1, \ldots, a_N\}$. For each group g_i, the moral evaluation function satisfies:

$$M_{g_i}(c, a) = \begin{cases} 1 & \text{if } a = a_i, \\ 0 & \text{if } a \neq a_i. \end{cases}$$

As a stylized model of aggregation, consider an equal-weight welfare objective that does not enforce each group's admissibility constraint as a hard veto. In the strictly disjoint case, the symmetric aggregate policy distributes probability mass uniformly:

$$\pi^*(a_k \mid c) = \frac{1}{N} \quad \forall k \in \{1, \ldots, N\}.$$

Proposition 3.3 (Aggregate misalignment under strict intolerance)**.** *Under Assumptions 1, 3, and Strict Intolerance, the symmetric aggregate policy π^* is not morally aligned with any group $g \in G$.*

Proof sketch. Recall that universal moral alignment for a group g_i requires $\pi(a \mid c) > 0 \Rightarrow M_{g_i}(c, a) = 1$. The policy π^* assigns positive probability $1/N$ to every action a_j. For any group g_i, all actions a_j with $j \neq i$ are inadmissible by Strict Intolerance, so π^* violates g_i's constraint with probability $(N-1)/N$. Since g_i was arbitrary, the policy is misaligned with every group.

The intuition is that when an algorithm pursues a global aggregate without enforcing group-specific vetoes, it may seek a compromise among N mutually exclusive actions while failing to satisfy any constituent group as such.

3.4. Moral Affinity and the Cost of Consensus

Of course, this approach imposes a hard constraint on the set of admissible actions for each group. We can relax this strict intolerance by considering a continuous moral admissibility model for each group and treat alignment as a satisfaction threshold. Allow $M_g : C \times A \to [0, 1]$ and consider each group has some satisfaction threshold of $\tau_g \in (0, 1]$ for some a given a context c.

Then a policy π is τ-morally aligned with group g_i if the expected moral value satisfies a threshold τ_i:

$$\mathbb{E}_{a \sim \pi(\cdot|c)}[M_g(c, a)] \geq \tau_g.$$

Assumption 5 (Sharp Preferences). Each group g_i prefers action a_k ($M_{g_i}(a_k|c) = 1$) and has low affinity $\epsilon < \tau_{g_i}$ for all other actions.

Here, τ_{g_i} operates as a normative floor for alignment for each group. It represents the minimum moral affinity a policy must achieve to be aligned with g. Any policy π where $\mathbb{E}[M_g] < \tau_g$ is then misaligned with g. Conversely, consider ϵ as the baseline consensus between disparate moral frameworks M_g and M_h for groups $g, h \in \mathcal{G}, g \neq h$. This is the marginal moral value that one group assigns to actions preferred by another and in situations of high polarization, ϵ tends to zero.

Proposition 3.4 (Cost of consensus)**.** *Let $\tau_{\min} = \min_{g \in G} \tau_g$. Under Assumption 5, if the minimal normative floor exceeds the residual affinity ($\tau_{\min} > \epsilon$), then for sufficiently large population size $N > \frac{1-\epsilon}{\tau_{\min}-\epsilon}$, the symmetric fair policy π^* is universally misaligned because the moral signal for any specific group is diluted by competing moral frameworks.*

Proof sketch. For group g_i, the symmetric policy assigns probability $1/N$ to its preferred action and probability $(N-1)/N$ to actions for which its affinity is ϵ. Its expected moral value is therefore $\frac{1}{N} + \frac{N-1}{N}\epsilon$. This value is below τ_i whenever $N > \frac{1-\epsilon}{\tau_i-\epsilon}$. If this inequality holds for the minimal normative floor $\tau_{\min}$, then it holds for every group threshold at least as large as $\tau_{\min}$. Hence the symmetric policy fails the satisfaction threshold for all groups.

We do not need every group to express sharp preferences for a conflict to arise; even one group with a sharp moral preference can make universal alignment fail in the relevant context. We refer to this as the cost of consensus: increasing the breadth of included moral frameworks can diminish fidelity to any individual group's moral model.

4. Implications and Research Directions

The impossibility results derived in Section 3 suggest that the prevailing "value learning" approach which treats alignment as an epistemic problem of discovering a hidden, coherent reward function is fundamentally flawed. Rather than treating alignment as a problem of learning the correct human values, we argue that alignment is fundamentally a problem of governing moral disagreement under impossibility constraints.

Keeping the SCT literature in mind, we give three recommendations. (1) Remove emphasis on finding "perfect" alignment with a dataset and instead, recognize normative governance mechanisms are needed in order to represent a plurality of moral perspectives, reason about incompatibilities in the data and limitations of a policy, and be explicit about tradeoffs between alignment accuracy and robustness. (2) Create new alignment benchmarks that are cognizant of impossibility results by testing systems for how well they handle minority disagreements and implicit dictatorship. Alignment should not be solely outcome focused but also concerned with the fairness of the procedure. (3) Shift away from strict pairwise comparisons when building preference datasets and towards choice sets and other elicitation schemes to better capture pluralism.

4.1. From "Learning Values" to Normative Governance

A dominant narrative in alignment research frames the problem of alignment as inferring a single latent reward or preference function from human feedback and optimizing a policy accordingly. This framing implicitly assumes that human values are sufficiently coherent, aggregable, and stable to be

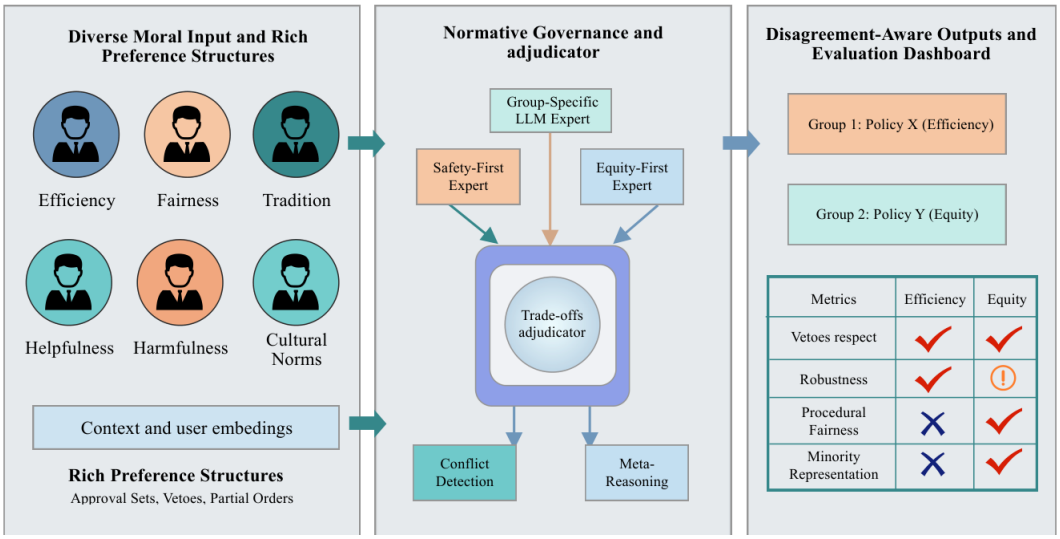

Figure 2. A roadmap for normative governance and pluralistic alignment.

represented by a unified objective (Kaufmann et al., 2025; Gabriel, 2020). Decades of impossibility results demonstrate that, under pluralism, no aggregation rule can convert heterogeneous normative judgments into a single decision function without violating basic fairness or autonomy constraints (Sen, 1970; Arrow, 1951; Conitzer et al., 2024).

Our results reinforce this critique by showing that even when moral judgments are perfectly observed, universal alignment may be formally impossible. We demonstrate that no such coherent function exists for disjoint groups; maximizing likelihood on a diverse dataset merely produces a confused aggregate that fails every group's specific moral constraints and even when there exists moral affinity between groups, it is unstable. This implies that misalignment is not primarily a failure of learning fidelity, data scale, or model capacity, but a structural consequence of treating alignment as objective inference rather than collective decision-making. In this sense, "learning the right values" is the wrong abstraction: there may be no single value function to learn.

We recommend abandoning the goal of "perfect" alignment via data scaling. Instead, research should treat alignment as a problem of normative governance mechanisms. Such mechanisms do not attempt to collapse moral pluralism into a scalar reward. Instead, they explicitly represent multiple normative perspectives, reason about their incompatibilities, and implement procedures for adjudication under disagreement (Liscio et al., 2023; Gauthier, 1992; Gill & Germann, 2022). This reframing aligns alignment research with long standing insights from political philosophy and economics, where legitimacy arises from fair procedures rather than convergence to a purportedly correct outcome.

4.2. Impossibility Aware Alignment Evaluation

Alignment benchmarks are typically designed to measure outcome accuracy: agreement with majority labels, preference win rates, or reward maximization (Ji et al., 2025). These metrics implicitly endorse a value learning worldview, treating disagreement as noise and convergence as progress. When preferences conflict, outcome focused metrics systematically favor majority groups and obscure procedural unfairness (Fehr & Schmidt, 1999; Hendrycks et al., 2021).

Our analysis suggests that alignment evaluation must be impossibility-aware. Rather than asking whether a system selects the "right" answer, benchmarks should ask how systems behave when no answer is jointly acceptable. This includes testing whether a system disproportionately violates minority constraints, collapses to implicit dictatorship, or fails to surface normative conflict. Such behaviors are predictable failure modes under aggregation, yet remain largely invisible under current benchmarks (Hendrycks et al., 2021).

Importantly, SCT emphasizes that collective decision making must be evaluated not only by its outcomes, but by the fairness and transparency of the procedure itself (Sen, 1970; Arrow, 1951; Moulin, 1991; Gibbard, 1973). Translating this insight to alignment suggests that benchmarks should incorporate normative desiderata: respect for vetoes, robustness to group reweighting, and explicit acknowledgment of trade-offs. Progress in alignment should therefore be measured by how systems manage disagreement, not merely by how efficiently they suppress it in Figure 2.

4.3. Data for Pluralism, Not Convergence

For pluralistic AI, the central dataset question is not only whether a label predicts a majority preference, but whether the data preserve who disagrees, why they disagree, and which options each group regards as inadmissible. Pairwise preference comparisons are the dominant data modality in alignment research, particularly in RLHF (Kaufmann et al., 2025). This structure fundamentally misrepresents pluralism, as comparisons force annotators to choose winners and conflate when multiple options are acceptable, unacceptable, or incomparable, thereby manufacturing consensus where none exists.

This data representation bias reinforces value learning narratives by making disagreement appear resolvable through averaging. In contrast, SCT emphasizes richer preference structures such as approval sets, vetoes, partial orders, and context-dependent admissibility (Noothigattu et al., 2018; Fine, 1972).

Although this position paper does not report new experiments, the argument yields direct empirical diagnostics for future work. Preference datasets for pluralistic alignment should report whether disagreement disappears under group conditioning, whether within-group preferences remain cyclic or incompatible, and whether scalar reward models preserve or erase pairwise disagreement. These diagnostics are especially relevant for datasets such as pluralistic LLM preference corpora and moral-judgment datasets, where disagreement is not merely annotator noise but part of the object being modeled.

A key research agenda is to study how different elicitation schemes interact with aggregation mechanisms and downstream policies. Richer preference representations make moral conflict explicit, enabling alignment systems to reason about admissibility rather than forced optimization. This shift supports pluralistic alignment while making impossibility results operational rather than theoretical.

5. Conclusions

In this work, we have argued that the central challenge of moral alignment is not simply to improve current techniques, but to confront a structural limitation of the universal-alignment objective itself. In pluralistic settings, different groups can hold internally coherent yet conflicting moral evaluations of the same context-action pairs, which means that no non-degenerate single policy can satisfy everyone simultaneously. Under stronger forms of disagreement, aggregation can even produce outcomes that are misaligned with every group. This makes a single unconditional universal-alignment objective not merely difficult in practice, but impossible in principle under the assumptions we examined. Rather than treating this as an engineering bottleneck to be solved with better optimization or more data, we frame it as a conceptual limit of current reward-modeling and preference-aggregation paradigms. In response, we outlined a pluralistic agenda for alignment: developing fine-grained evaluations that better capture diverse perspectives, establishing principled customization bounds that reduce strategic misreporting and majority-take-all failures, and incorporating more robust social choice mechanisms together with broader policy coverage. The goal is not to force moral disagreement into a single averaged objective, but to build alignment systems that represent, navigate, and govern disagreement more honestly in a diverse and strategic world.

Impact Statement

This paper studies structural limits in pluralistic AI alignment. Its positive impact is to clarify why disagreement across human values should not automatically be treated as noise to be averaged away, and why evaluation datasets should preserve information about context, group membership, and disagreement structure. The main risk is that impossibility results may be misread as fatalistic or as implying that pluralistic alignment should not be attempted. Our claim is narrower: a single unconditional policy or scalar reward is often an inadequate design target when preferences are plural, conflicting, or cyclic. We intend the analysis to motivate more transparent, context-sensitive, and procedurally accountable alignment methods.

References

Arrow, K. J. *Social Choice and Individual Values*. Wiley, New York, NY, USA, 1951.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

Baltaji, R., Hemmatian, B., and Varshney, L. R. Conformity, confabulation, and impersonation: Persona inconstancy in multi-agent llm collaboration. *CoRR*, abs/2405.03862, 2024.

Betzler, N., Slinko, A., and Uhlmann, J. On the computation of fully proportional representation. In *Journal of Artificial Intelligence Research*, volume 47, pp. 475–519, 2013.

Bianchi, F., Chia, P. J., Yüksekgönül, M., Tagliabue, J., Jurafsky, D., and Zou, J. How well can llms negotiate? negotiationarena platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Blau, J. H. Liberal values and independence. *The Review of Economic Studies*, 42(3):395–401, 1975.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39(3–4):324–345, 1952.

Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D. (eds.). *Handbook of Computational Social Choice*. Cambridge University Press, 2016.

Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mossé, M., Pacuit, E., Russell, S., Schoelkopf, H., Tewolde, E., and Zwicker, W. S. Position: Social choice should guide ai alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Craven, J. Liberalism and individual preferences. *Theory and Decision*, 14(4):351–360, 1982.

Durkheim, E. *Suicide: A Study in Sociology*. Routledge, 2005.

Elkind, E., Lackner, M., and Peters, D. Preference restrictions in computational social choice: Recent progress. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 4062–4065, 2016.

Fehr, E. and Schmidt, K. M. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, 1999.

Fine, T. L. Review of utility theory for decision making. *IEEE Transactions on Information Theory*, 18(5):696–697, 1972.

Finnemore, M. Norms, culture, and world politics: Insights from sociology's institutionalism. *International Organization*, 50(2):325–347, 1996.

Gabriel, I. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, 2020.

Gauthier, D. The roots and roles of normative governance. *Synthese*, 91(3):319–335, 1992.

Gibbard, A. Manipulation of voting schemes: A general result. *Econometrica*, 41(4):587–601, 1973.

Gigliotti, G. A. Comment on craven. *Theory and Decision*, 21(1):89–95, 1986.

Gill, A. S. and Germann, S. Conceptual and normative approaches to ai governance for a global digital ecosystem supportive of the un sustainable development goals. *AI and Ethics*, 2(2):293–301, 2022.

Guo, T., Guo, K., Nan, B., Liang, Z., Guo, Z., Chawla, N. V., Wiest, O., and Zhang, X. What indeed can gpt models do in chemistry? a comprehensive benchmark on eight tasks. *CoRR*, abs/2305.18365, 2023.

Harsanyi, J. C. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. In *Essays on Ethics, Social Behavior, and Scientific Explanation*, pp. 6–23. Springer Netherlands, 1976.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., and Steinhardt, J. Aligning ai with shared human values. In *International Conference on Learning Representations*, 2021.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Vierling, L., Hong, D., Zhou, J., Zhang, Z., Zeng, F., Dai, J., Pan, X., Ng, K. Y., O'Gara, A., Xu, H., Tse, B., Fu, J., McAleer, S., Yang, Y., Wang, Y., Zhu, S.-C., Guo, Y., and Gao, W. Ai alignment: A comprehensive survey, 2025.

Kaufmann, T., Weng, P., Bengs, V., and Hüllermeier, E. A survey of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2025.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: A research direction, 2018.

Liscio, E., Lera-Leri, R., Bistaffa, F., Dobbe, R. I., Jonker, C. M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., and Murukannaiah, P. K. Value inference in sociotechnical systems. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1774–1780, 2023.

Min, S., Michael, J., Hajishirzi, H., and Zettlemoyer, L. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 5783–5797, 2020.

Moulin, H. *Axioms of Cooperative Decision Making*. Cambridge University Press, 1991.

Noothigattu, R., Gaikwad, S. N. S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., and Procaccia, A. D. A voting-based system for ethical decision making, 2018.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Prasad, A., Hase, P., Zhou, X., and Bansal, M. Grips: Gradient-free, edit-based instruction search for prompting

large language models. *arXiv preprint arXiv:2203.07281*, 2022.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Salemi, A., Mysore, S., Bendersky, M., and Zamani, H. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*, 2023.

Salminen, J., Amin, D., and Jansen, B. Using ai for user representation: An analysis of 83 persona prompts. *arXiv preprint arXiv:2508.13047*, 2025.

Satterthwaite, M. A. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2):187–217, 1975.

Scherrer, N., Shi, C., Feder, A., and Blei, D. M. Evaluating the moral beliefs encoded in llms. In *Advances in Neural Information Processing Systems*, 2023.

Sen, A. Liberty and social choice. *Journal of Philosophy*, 80(1):5–28, 1983.

Sen, A. K. The impossibility of a paretian liberal. *Journal of Political Economy*, 1970.

Sutrop, M. Challenges of aligning artificial intelligence with human values. *Acta Baltica Historiae et Philosophiae Scientiarum*, 8(2):54–72, 2020.

Tan, Z., Zeng, Q., Tian, Y., Liu, Z., Yin, B., and Jiang, M. Democratizing large language models via personalized parameter-efficient fine-tuning. *arXiv preprint arXiv:2402.04401*, 2024.

Triandis, H. C. *Culture and Social Behavior*. McGraw-Hill, 1994.

Wooldridge, M. J. and Jennings, N. R. Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 10 (2):115–152, 1995.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Qin, W., Zheng, Y., Qiu, X., Huang, X., Zhang, Q., and Gui, T. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2), 2025.

Zhu, H., Wang, X., Wang, Z., and Xv, K. An emotion-sensitive dialogue policy for task-oriented dialogue system. *Scientific Reports*, 14(1):19759, 2024.