Mission Impossible: Universal LLM Moral Alignment

SAIMUN HABIB*, School of Informatics, University of Edinburgh, United Kingdom
XIAO XIAO*, School of Computer Science and Informatics, University of Liverpool, United Kingdom
MENG FANG†, School of Computer Science and Informatics, University of Liverpool, United Kingdom
FENGXIANG HE, School of Informatics, University of Edinburgh, United Kingdom

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 that sufficiently capable models can approximate a coherent and universally valid moral objective. We argue that this assumption is false in pluralistic settings. 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 single policy can satisfy all groups simultaneously. Under stronger forms of disagreement, aggregation can even produce policies that are misaligned with every group. This is therefore not merely an engineering bottleneck to be overcome with more data, larger models, or improved optimization, but a structural limitation of the universal-alignment objective itself. This paper clarifies a conceptual limit of current reward-modeling and preference-aggregation paradigms. 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.

CCS Concepts: • **Computing methodologies** → **Machine learning**; **Philosophical/theoretical foundations of artificial intelligence**; • **Social and professional topics**;

Additional Key Words and Phrases: Large Language Model, Morality, Value Alignment

ACM Reference Format:
Saimun Habib, Xiao Xiao, Meng Fang, and Fengxiang He. 2026. Mission Impossible: Universal LLM Moral Alignment. 1, 1 (March 2026), 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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 [21, 38, 39], and interact with humans and other agents over extended horizons [3, 5]. 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 [36].

*Both authors contributed equally to this research.
†Corresponding authors. Email: Meng.Fang@liverpool.ac.uk.

Authors' Contact Information: Saimun Habib, School of Informatics, University of Edinburgh, Edinburgh, United Kingdom; Xiao Xiao, School of Computer Science and Informatics, University of Liverpool, Liverpool, United Kingdom; Meng Fang, School of Computer Science and Informatics, University of Liverpool, Liverpool, United Kingdom; Fengxiang He, School of Informatics, University of Edinburgh, Edinburgh, United Kingdom.

ACM XXXX-XXXX/2026/3-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

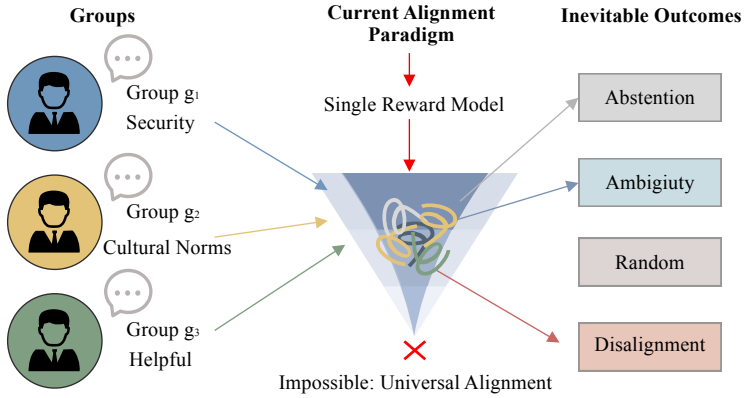

Fig. 1. The structural impossibility of universal alignment

Most existing approaches operationalize this objective through variants of preference learning [16, 24, 26], reinforcement learning from human feedback (RLHF) [23, 25, 31], rule-based constraints [40], or instruction tuning [2]. While differing in implementation, these methods share a common implicit assumption: *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 as fundamentally flawed. Human morality is not a single, universally agreed-upon construct; rather, moral judgments are deeply shaped by cultural norms, social practices, and historical contexts [15, 37]. 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. Based on the reality that different individuals and communities impose considerably diverse and even conflicting moral preferences, *universal LLM alignment is truly a mission impossible* in Figure 2.

Despite this, current alignment paradigms largely treat moral variation as a problem of data coverage or modeling capacity: collect more feedback, average across annotators, or train larger models to absorb diverse preferences, and then we may have a 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 induces objective-level incompatibilities, rendering universal moral alignment impossible in principle rather than merely difficult in practice.*

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 [1], we show that under minimal and intuitive assumptions, no single policy can satisfy the moral constraints of all cultural groups, simultaneously. This result should be independent of any particular moral theory, dataset, or model architecture. Instead, it follows directly from the diversity and disagreement within a real-world moral dataset.

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 three 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 universal moral alignment under cultural diversity and disagreement.
- 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, rather than to propose a specific algorithm or theory.

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 [23, 25, 31].

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 $a_w \succ a_l$ [25].

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 [7]. 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 [1, 35].

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) [35]:

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

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 an 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. [1, 32, 35]. 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 [22]. Furthermore, SCT already, has a home in computer science with multiple research efforts in efficient vote counting, mechanism designs, and algorithmic game theory [4, 8, 12] 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 [6, 10, 19, 34]. To that end, we formalize three mild assumptions 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 which mirrors 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 [28, 33]. 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 [15, 37], mismatched norms can create conflict because one side may view the other's behavior as inappropriate [11]. 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 policy π achieves Universal Moral Alignment if, all groups find some action admissible in a specific context c,

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

Proposition 3.1. *Since each group maintains an internally consistent moral framework (A1) that remains irreconcilably disjoint in specific contexts (A2), and given that the model is mandated to provide substantive, non-degenerate responses (A3), it is mathematically inevitable that a single policy cannot simultaneously align with all groups.*

This is straightforward to observe but note, the requirement for universal moral alignment is not for all contexts. In fact, it is weaker; for any context where one group may find an action disagreeable, the policy cannot satisfy the collective.

3.2 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:

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

By Gibbard-Satterthwaite's Theorem, a learning algorithm aiming to maximize aggregate fairness or social welfare (satisfying Assumption 3) produces a stochastic policy π^* that does not enforce moral admissibility as a hard constraint and distributes probability mass uniformly [32]:

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

Proposition 3.2. *Under Assumptions 1, 3, and Strict Intolerance, the welfare-maximizing policy π^* is not morally aligned with any group $g \in G$.*

The intuition here is when an algorithm pursues global optimality or social welfare maximization, it is compelled to seek an equilibrium among N mutually exclusive actions resulting in a policy that fails to satisfy any constituent group.

3.3 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 3.3. *Let $\tau_{\min} = \min_{g \in G}\{\tau_{g_1}, \ldots \tau_{g_N}\}$. Under Assumption 5, if the minimal normative floor exceeds the residual affinity ($\tau_{\min} > \epsilon$), then for sufficiently large population $N > \frac{1-\epsilon}{\tau_{min}-\epsilon}$, the fair policy π^* is universally misaligned as the the moral signal for any specific group is drowned out by the noise of competing moral frameworks .*

We, in fact, do not need every group to express sharp preferences; even if one group has a sharp moral preference, then universal alignment is impossible. We consider this to be the cost of consensus, where we have a nromative tradeoff between the breadth of disparate moral model inclusion N, diminishes the models fidelity to any individual group's moral model.

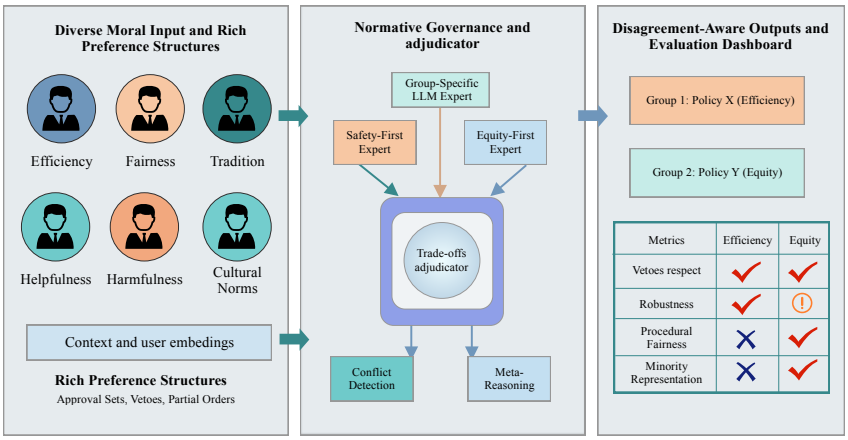

Fig. 2. A Roadmap for Normative Governance and Pluralistic Alignment

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 options to better capture pluralism.

4.1 From "Learning Values" to Normative Governance

A dominant narrative in alignment research frames the problem as one of value learning: 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 represented by a unified objective [16, 25]. 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 [1, 9, 35].

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 [17, 20, 27]. 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 [24]. 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 [13, 23].

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 [23].

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 [1, 18, 29, 35]. 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

Pairwise preference comparisons are the dominant data modality in alignment research, particularly in RLHF [25]. 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 [14, 30].

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 universal moral alignment 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.

Acknowledgments

S. Habib was supported in part by a NERC Doctoral Training Partnership grant (NE/S007407/1).

References

[1] K. J. Arrow. *Social Choice and Individual Values*. Wiley: New York, New York, NY, USA, 1951.
[2] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
[3] R. Baltaji, B. Hemmatian, and L. R. Varshney. Conformity, confabulation, and impersonation: Persona inconstancy in multi-agent LLM collaboration. *CoRR*, abs/2405.03862, 2024.
[4] N. Betzler, A. Slinko, and J. Uhlmann. On the Computation of Fully Proportional Representation. *Journal of Artificial Intelligence Research*, 47:475–519, July 2013. arXiv:1402.0580 [cs].
[5] F. Bianchi, P. J. Chia, M. Yüksekgönül, J. Tagliabue, D. Jurafsky, and J. Zou. How well can llms negotiate? negotiationarena platform and analysis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.
[6] J. H. Blau. Liberal Values and Independence. *The Review of Economic Studies*, 42(3):395–401, 1975. Publisher: Review of Economic Studies Ltd.
[7] R. A. BRADLEY and M. E. TERRY. RANK ANALYSIS OF INCOMPLETE BLOCK DESIGNS: THE METHOD OF PAIRED COMPARISONS. *Biometrika*, 39(3-4):324–345, Dec. 1952.
[8] F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia. *Handbook of Computational Social Choice*. Cambridge University Press, USA, 1st edition, Mar. 2016.
[9] V. Conitzer, R. Freedman, J. Heitzig, W. H. Holliday, B. M. Jacobs, N. Lambert, M. Mossé, E. Pacuit, S. Russell, H. Schoelkopf, E. Tewolde, and W. S. Zwicker. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.
[10] J. Craven. Liberalism and individual preferences. *Theory and Decision*, 14(4):351–360, Dec. 1982.
[11] E. Durkheim. *Suicide: A study in sociology*. Routledge, 2005.
[12] E. Elkind, M. Lackner, and D. Peters. Preference restrictions in computational social choice: recent progress. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 4062–4065, New York, New York, USA, July 2016. AAAI Press.
[13] E. Fehr and K. M. Schmidt. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, 1999.
[14] T. Fine. Review of 'Utility Theory for Decision Making' (Fishburn, P. C.; 1970). *IEEE Transactions on Information Theory*, 18(5):696–697, Sept. 1972.
[15] M. Finnemore. Norms, culture, and world politics: insights from sociology's institutionalism. *International organization*, 50(2):325–347, 1996.
[16] I. Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.
[17] D. Gauthier. The Roots and Roles of Normative Governance. *Synthese*, 91(3):319–335, 1992. Publisher: Springer.
[18] A. Gibbard. Manipulation of Voting Schemes: A General Result. *Econometrica*, 41(4):587–601, 1973. Publisher: Econometric Society.
[19] G. A. Gigliotti. Comment on Craven. *Theory and Decision*, 21(1):89–95, July 1986.
[20] A. S. Gill and S. Germann. Conceptual and normative approaches to AI governance for a global digital ecosystem supportive of the UN Sustainable Development Goals (SDGs). *Ai and Ethics*, 2(2):293–301, 2022.

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

[22] J. C. Harsanyi. Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. In J. C. Harsanyi, editor, *Essays on Ethics, Social Behavior, and Scientific Explanation*, pages 6–23. Springer Netherlands, Dordrecht, 1976.

[23] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt. Aligning AI with shared human values. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

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

[25] T. Kaufmann, P. Weng, V. Bengs, and E. Hüllermeier. A survey of reinforcement learning from human feedback. *Trans. Mach. Learn. Res.*, 2025, 2025.

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

[27] E. Liscio, R. Lera-Leri, F. Bistaffa, R. I. Dobbe, C. M. Jonker, M. Lopez-Sanchez, J. A. Rodriguez-Aguilar, and P. K. Murukannaiah. Value Inference in Sociotechnical Systems. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pages 1774–1780, Richland, SC, May 2023. International Foundation for Autonomous Agents and Multiagent Systems.

[28] S. Min, J. Michael, H. Hajishirzi, and L. Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5783–5797. Association for Computational Linguistics, 2020.

[29] H. Moulin. *Axioms of Cooperative Decision Making*. Cambridge University Press, 1991. Series Title: Cambridge Books.

[30] R. Noothigattu, S. N. S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, and A. D. Procaccia. A Voting-Based System for Ethical Decision Making, Dec. 2018. arXiv:1709.06692 [cs].

[31] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[32] M. A. Satterthwaite. 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, Apr. 1975.

[33] N. Scherrer, C. Shi, A. Feder, and D. M. Blei. Evaluating the moral beliefs encoded in llms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[34] A. Sen. Liberty and Social Choice. *Journal of Philosophy*, 80(1):5–28, 1983. Publisher: Journal of Philosophy Inc.

[35] A. K. Sen. The Impossibility of a Paretian Liberal. *Scholarly Articles*, 1970. Number: 3612779 Publisher: Harvard University Department of Economics.

[36] M. Sutrop. Challenges of aligning artificial intelligence with human values. *Acta Baltica historiae et philosophiae scientiarum*, 8(2):54–72, 2020.

[37] H. C. Triandis. *Culture and social behavior*. Mcgraw-Hill Book Company, 1994.

[38] M. J. Wooldridge and N. R. Jennings. Intelligent agents: theory and practice. *Knowl. Eng. Rev.*, 10(2):115–152, 1995.

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

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