Position: Alignment Needs Rule-Class Routing Before Preference Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

This position paper argues that alignment pipelines should classify rules before they aggregate preferences. RLHF [15], Constitutional AI [3], DPO [16], RLAIF [13], and Deliberative Alignment [14] differ technically, but each tends to turn contested normative input into one global policy. Social-choice theory explains why this is unsafe on unrestricted domains: aggregation can help when there is public convergence, but it cannot decide which questions are eligible for aggregation. Conitzer et al. [4] reframe alignment as social choice; we agree, and argue social choice identifies a boundary, not a solution. Arrow, Gibbard–Satterthwaite, and Sen establish that aggregation over unrestricted value disagreement has no procedure satisfying minimal democratic, strategic, and liberty-preserving conditions. The eligibility decision is institutional: the system designer must state who is authorized to decide whether a rule is aggregable, user-configurable, or non-negotiable. We propose a routing layer for behavioral rules. Class I rules are public prohibitions suitable for aggregation; Class II rules concern reasonable disagreement and should support configurable defaults; Class III rules protect rights or vulnerable users and should be implemented as constraints with appeal mechanisms. The paper connects impossibility results to alignment design, compares current methods through this lens, and gives a six-step routing procedure with a five-item disclosure checklist.

1 Introduction

In December 2022, Bai et al. [3] introduced Constitutional AI, a training method using a short list of principles, called a constitution, against which the model critiques and revises its own outputs. The terminology is deliberate: any general-purpose AI system will be governed by some set of principles, explicit or implicit, and the choice of the word constitutional is meant to make that governance explicit. The technical contribution is substantial; the conceptual framing is ahead of most of the aligned-LLM literature.

The label identifies the right problem, but the procedure leaves the authority layer underspecified. Bai et al. [3] describe their principles as chosen ad hoc and iteratively for research purposes and acknowledge that the principles should be refined by a larger stakeholder community. No mechanism is specified by which that refinement is to occur. Anthropic's 2023 follow-up on Collective Constitutional AI [1] took first steps by convening citizen-panel input, while the resulting constitution is still operationalized within a preference-learning pipeline, leaving the authority layer underspecified.

A similar pattern characterizes the alignment research program more broadly. RLHF [15, 21] optimizes against pairwise preference comparisons aggregated through a Bradley-Terry-style reward model; Constitutional AI treats principles as text-encoded inputs to a self-critique loop; DPO [16] reformulates the optimization without changing the object being learned; RLAIF [13] substitutes AI for human feedback without addressing whose values the AI encodes; Deliberative Alignment

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

[14] introduces deliberation at chain-of-thought level but retains the preference-learning framing at training. We do not claim these methods are formally identical social welfare functions. The claim is procedural: when a pipeline converts contested normative input into one global behavioral policy without first classifying the decision type, it performs scalar reward compression before legitimacy assignment, and the legitimacy question reappears at deployment time as an ungrounded normative choice.

Conitzer et al. [4] explicitly engage the reward-compression step as a social-choice-theoretic problem. They point out that RLHF and Constitutional AI both face questions, including which humans provide input, how divergent input is aggregated, and how strategic manipulation is addressed, that the social-choice literature has studied since Arrow [2]. Their proposal is that alignment should adopt social-choice-theoretic tools; we regard it as correct as far as it goes.

Position. AI alignment should not treat all behavioral specifications as preference-aggregation problems. Alignment pipelines should first classify decisions into preference-learning-appropriate, reasonable-disagreement, and rights-protective classes, and route each class through the corresponding mechanism, with the decision authority disclosed.

Who classifies the classifier? The deepest objection to a routing layer is that it requires authority: someone or some body has to decide that decision r is Class I, II, or III, and that authority itself is unaccountable on the same grounds we have used to criticize current alignment. We acknowledge it and respond directly. The paper does not solve the authority problem; it makes authority assignment explicit. Current alignment pipelines already classify implicitly: a contested political topic forced through preference learning into a single global answer has been classified as Class I by default, by the team that specified the training data, the principle list, or the refusal policy. The classification is happening; the question is whether it is disclosed and revisable. The path to procedural authority is institutional—citizen assemblies, multi-stakeholder bodies, public-comment procedures—not technical. We do not pretend to have the right institutional answer; we argue that designating the question is an improvement over leaving it implicit.

The paper proceeds in four steps. First, three foundational impossibility results establish that preference aggregation over deep value disagreements has no generally valid technical resolution. Second, what political philosophy offers in place—a vocabulary of rule eligibility, decision authority, and override mechanisms—is not reducible to better preference learning. Third, we engage Conitzer et al. [4] to establish where social-choice-theoretic alignment ends and rule-class routing must begin. Fourth, we offer a routing-layer specification and a disclosure checklist that treat preference learning and rule-class routing as distinct operations.

The critique is internal to Constitutional AI: the term "constitution" points to the right problem, but the training pipeline still needs a rule-class decision step. Political theory supplies the substrate—who is authorized to decide which rules are aggregable, which are configurable, and which are non-negotiable—and importing it would strengthen, not undermine the approach.

Bridge to RLHF, in four lines. The social-choice impossibility results do not say RLHF literally violates Arrow's theorem. They identify a structural problem in the pipeline: (i) RLHF, DPO, RLAIF collect or simulate pairwise preferences; (ii) reward modeling compresses heterogeneous judgments into a single scalar objective; (iii) the social-choice warning applies to the compression step, not to the neural optimizer downstream of it; (iv) the result is that the institutional choice of which disagreements are aggregable is hidden inside the reward-modeling step instead of disclosed. Naming the rule-class step exposes this hidden choice.

2 Three Impossibility Results and Their Alignment Implications

The impossibility results are not invoked as literal theorems about gradient descent or about RLHF as a whole. They were developed for ordinal preference aggregation in voting and welfare contexts; their direct assumptions—fixed alternative set, ordinal and complete individual preferences, discrete voters—do not hold without modification for preference learning over an LLM's continuous response space with cardinal rewards. **What the results identify is a prior institutional step: which disagreements are eligible to be compressed into one reward objective.** RLHF, DPO, and RLAIF collect or simulate pairwise preferences, and reward modeling compresses heterogeneous normative

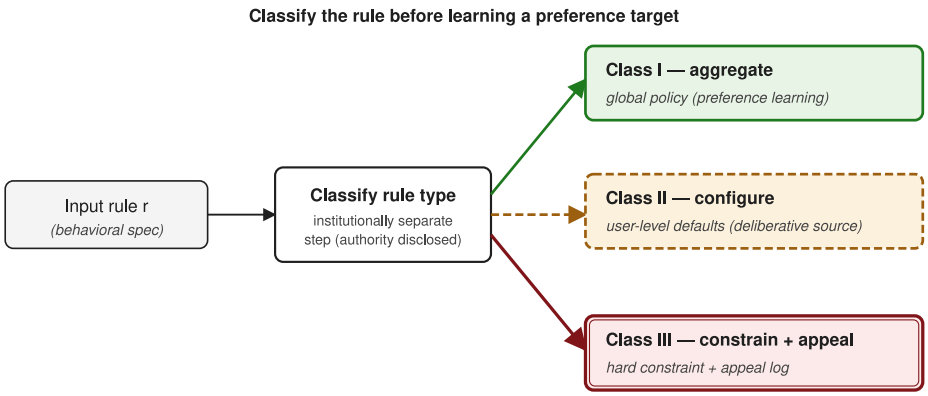

Figure 1: Rule-class classification is the missing step between behavioral rules and preference learning. A candidate behavioral rule r enters an institutionally separate classification step with a public rationale, then routes to one of three classes: *aggregate* (Class I, e.g., chemical-weapon synthesis: globally enforced via preference learning), *configure* (Class II, e.g., contested vaccine policy: user-level defaults sourced deliberatively), or *constrain + appeal* (Class III, e.g., CSAM, self-harm protection: hard constraint with appeal log; non-aggregable). Detailed in Algorithm 1.

judgments into a single scalar objective; the social-choice warning is about that compression step, not about the neural optimizer. Ge et al. [6] make a version of this transfer rigorous within a Bradley-Terry-Luce model. Our aim is to identify a structural warning that alignment papers should either accept, qualify, or argue does not apply.

Arrow. Arrow [2] shows no social welfare function over ≥ 3 alternatives can satisfy unrestricted domain, non-dictatorship, Pareto efficiency, and independence of irrelevant alternatives jointly. RLHF over an LLM's response space performs the same job and faces the same tradeoff; Siththaranjan et al. [20] document that distributional preference learning hides context from the aggregation. Some desideratum must be relaxed, and that relaxation is a governance choice.

Gibbard–Satterthwaite. Gibbard [7], Satterthwaite [18] establish that any non-dictatorial deterministic social choice function on ≥ 3 alternatives is strategically manipulable. The labeler-to-reward-model pipeline inherits structural manipulability; the empirical shield (anonymous workers lacking strategic position) weakens as labeling professionalizes.

Sen. Sen [19] shows minimal liberal rights and Paretian efficiency are jointly unsatisfiable under unrestricted preferences. When user autonomy and third-party-affecting preferences both enter the alignment objective, aggregate preference satisfaction cannot by itself settle which liberties are protected—the case for medical, legal, and content with third-party consequences.

Why categorical, not technical. The three results do not say aggregation is hard; they say aggregation over unrestricted human preferences has no fully-satisfactory solution. Restricting axioms is itself political. Conitzer et al. [4] correctly observe that different preference-aggregation rules correspond to different axiomatic profiles, but the choice of which axioms to keep is not adjudicable within social-choice theory alone. Political societies responded by developing rule-class institutions—constitutional courts, bills of rights, supermajority requirements—that shift the decision level. Alignment has not made that move.

3 What Rule-Class Theory Offers That Aggregation Theory Cannot

We draw on three strands of political philosophy.

Rawlsian political liberalism. Rawls [17] distinguishes a political conception of justice from a comprehensive doctrine. Reasonable citizens in a pluralistic society will permanently disagree on

comprehensive doctrines but can share a political conception supported by an overlapping consensus. Two implications follow. First, the question of what the LLM should do admits a similar distinction: some specifications can be justified in terms all reasonable users can accept, including refusal to assist in violence against identified persons; others rely on contested comprehensive doctrines. Current alignment treats both uniformly as preference-aggregation problems. A Rawlsian approach classifies them, applying preference learning to the political class and reserving for the comprehensive class either user-level configurability or principled refusal-with-explanation. Second, Rawls himself engages contested questions and argues some admit public-reason treatment when recast in terms of political values; the Class I, II, and III classification we propose requires its own argument for each case (no default assignment).

Habermasian deliberative democracy. Habermas [9] argues that democratic legitimacy flows from communicative rationality, that is, from procedures of mutual reason-giving under approximate equality and absence of coercion. Aggregation without deliberation is procedurally deficient regardless of which aggregation rule is used. Preference data collected from isolated labelers performing comparison tasks is the most aggregation-heavy, least deliberation-rich form of preference collection available. A Habermasian critique is that this produces technically precise reward models that are procedurally unlegitimized.

Gaus and the limits of public justification. Gaus [5] develops an account in which a social rule's legitimacy depends on whether it can be publicly justified, that is, on whether each reasonable member has sufficient reasons, from within their own evaluative standards, to endorse the rule. Where reasonable disagreement spans the space of possible rules, no rule can be imposed through public coercive power with full legitimacy; the domain must be handled through mechanisms that do not require publicly-justified consensus, including individual-level discretion, market mechanisms, or rules under weaker moral-equilibrium standards. For alignment, this yields a categorical distinction between cases where a training-level behavioral rule can be publicly justified, such as CSAM refusal, where overwhelming convergence exists, and cases where it cannot, including most politically loaded content. The Gausian implication is not that labs withdraw; it is that contested-domain rules lack the legitimacy of consensus-domain rules, and responsible alignment design routes contested-domain decisions to architectures that do not require publicly-justified consensus.

Landemore on the epistemic function of deliberation. Landemore [11, 12] argues that deliberative procedures are epistemically superior for decisions requiring integration of diverse perspectives. The argument is conditional on specific epistemic structure, not universal. For Class II contested-politics questions, deliberation may be both legitimacy-restoring and accuracy-improving: a citizen assembly's diversity compensates for the narrow demographic profile of current labeling pipelines.

4 Engaging Conitzer et al.

Because Conitzer et al. [4] is the natural predecessor and the paper we most directly build on, we engage it explicitly. Their central argument is that social choice theory provides the appropriate analytical framework for alignment's preference-aggregation problems. They survey key results, including Arrow, Gibbard-Satterthwaite, independence of clones, and strategic voting, and map them to RLHF design choices, prescribing that alignment researchers (i) make their aggregation rule explicit, (ii) evaluate it against social-choice axioms, and (iii) develop novel rules satisfying axioms standard RLHF fails. We endorse all three.

Our disagreement is about whether adopting them resolves alignment or transforms it into its next phase. Their program, fully implemented, would produce methodologically more sophisticated alignment without addressing the prior categorical question, namely which decisions should be aggregated at all versus routed to non-aggregative procedures. Social choice theory, including its most sophisticated axiomatic work [6], takes as given that the problem is preference aggregation; it asks which aggregation rule is best. The rule-eligibility question is prior: for which decisions is aggregation the right mode at all?

A legal analogy clarifies. Legal systems have voting procedures, which aggregate, and constitutional procedures, which protect rights. Constitutional courts do not produce better aggregation; they remove certain decisions from aggregation entirely, placing them under rights protection. A theorist studying

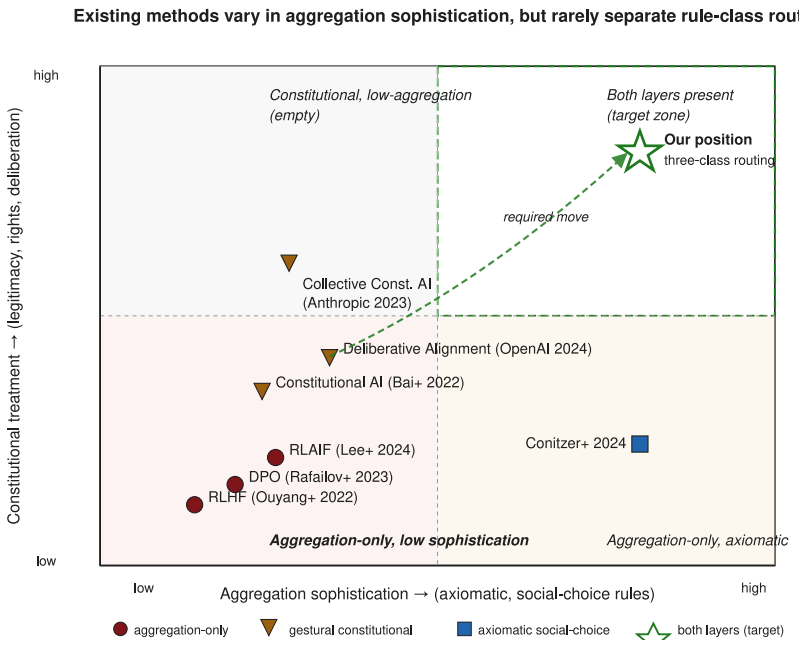

Figure 2: Existing methods vary in aggregation sophistication, but rarely separate rule-class routing. Schematic coding under the rubric in Appendix B; not a quantitative performance comparison. Aggregation-only methods (red circles, lower band): RLHF, DPO, RLAIF. Gestural-constitutional methods (orange triangles, mid band): Constitutional AI, Deliberative Alignment, Collective Constitutional AI. Axiomatic social-choice (blue square): Conitzer et al. 2024. Our position (hollow green star) sits in the target zone where rule-class routing is institutionally separate from preference learning.

only voting and developing the most sophisticated axiomatic theory of voting imaginable would still miss that some decisions are not put to a vote. That is the move Conitzer et al. [4] do not make. It is the move alignment must make. Their program is incomplete, not wrong: alignment should adopt their social-choice-theoretic sophistication and additionally develop the rule-class structure determining which decisions are assigned to which mode.

5 Existing Methods Through the Rule-Class Lens

Table 1 shows that essentially no current approach has substantive rule-class treatment, and procedural authority is either absent or thin. Collective Constitutional AI is closest on procedural authority; Conitzer et al. [4] is closest on methodological sophistication within aggregation. Neither fully occupies the rule-class space.

The consequence of alignment's aggregation-only framing is that contested political questions are settled through the private decisions of frontier-lab policy teams, with the technical apparatus presenting those decisions as value-neutral preference learning. The outputs are deployed to hundreds of millions of users globally, across jurisdictions with divergent settlements. This is not a stable equilibrium. The regulatory response, including the EU AI Act, U.S. state-level AI legislation, and national strategies across the Asia-Pacific, is increasingly concerned with the legitimacy question that aggregation framing cannot answer.

6 A Rule-Class Framework for Alignment: Three Decision Classes

We propose a procedure distinguishing three classes alignment currently treats uniformly. The procedure is procedural, not substantive: we do not claim to specify which decisions belong to which class; we claim alignment must perform the classification and currently does not.

Table 1: Aggregation versus rule-class features of principal alignment approaches. Procedural authority refers to the procedure by which the training signals or principles were authored or labeled, not to the ML quality of the resulting model.

Approach	Aggregation treatment	Rule-class treatment	Procedural authority
RLHF [15]	Pairwise preferences, BT aggregation	None	Vendor employment contracts
Constitutional AI [3]	Principles ad hoc, RLAIF	Gestural	Lab-internal
DPO [16]	Same object as RLHF	None	Vendor employment contracts
RLAIF [13]	AI-generated preferences	Depends on constitution	Reduced
Coll. Const. AI [1]	Citizen-panel principles	Partial	Procedural improvement, limited scope
Deliberative Alignment [14]	CoT inference, RLHF training	Partial	Uneven
Conitzer et al. [4]	Social-choice rules	None	Axiomatic transparency
Our position	Aggregation for the consensus class	Rule-class for the disagreement and rights classes	Deliberative-democratic

Class I: publicly-justified prohibitions. These are decisions for which there is overlapping consensus in the Rawlsian sense, namely prohibitions reasonable members of every reasonable moral tradition can endorse. Standard examples include assistance in CSAM production and weapons-of-mass-destruction synthesis, where cross-cultural moral convergence is substantial and the Gausian public-reason criterion is satisfied. For Class I, aggregation-theoretic alignment is appropriate and sufficient. Preference data converges because moral judgments converge; Conitzer et al. [4]'s axiomatic rigor improves implementation without changing framing. Class I is where current alignment is doing approximately the right thing.

Class II: reasonable-disagreement questions. These are decisions on which reasonable members of reasonable moral traditions legitimately disagree in ways public-reason testing does not resolve, including morally contested creative writing, positions on empirical questions where scientific consensus is contested, cultural and religious practices on which pluralistic societies have settled through institutional difference, not consensus, and politically-loaded topics whose classification is itself contested. Rawls is explicit: public power should not impose a single answer, because doing so disrespects reasonable disagreement that constitutes political liberty. The alignment analog is that a single LLM trained to a single set of answers on Class II questions risks imposing a single comprehensive doctrine as the default settlement for reasonable disagreement. The non-aggregative response is not to average preferences but to shift the level of decision-making. Class II decisions should be routed to user-level configuration, community-level customization, or deliberative procedures for deciding which user communities get which defaults. The technical implementation is non-trivial but tractable: model-level neutrality combined with user-configurable behavior, with defaults set by deliberative procedures, not frontier-lab policy teams.

Class III: rights-protective decisions. These are decisions where majoritarian preferences may legitimately be overridden by rights-protective considerations: protection of vulnerable user groups from manipulative content even where majority users prefer the content, protection of minority viewpoints from LLM-enforced majority consensus, and protection against exploitation even where user consent is structurally compromised. In legal theory, these are the decisions constitutional courts remove from majoritarian legislation; in alignment, aggregated preferences are not the right input. The model's refusal or intervention is justified by rights-protective reasoning, not by preference aggregation.

The classification problem. The first-order rule-class problem is performing the classification: which questions belong to which class? This is itself a political question and cannot be solved by technical means. Political societies solve it through constitutional conventions, courts, and deliberative institutions accumulating practice over decades. The alignment analog requires equivalent

infrastructure: bodies performing the Class I, II, and III classification with procedural authority, whose decisions become binding on training-level implementation. No such infrastructure exists. Frontier-lab policy teams perform a version internally; the outputs are communicated through model spec documents; the process lacks procedural authority in the sense political-theoretic accounts require.

Routing layer specification. We specify the routing layer as a six-step procedure (Algorithm 1), exposing the points at which classification, training, and disclosure decisions are made. The procedure does not introduce new training algorithms; it routes existing algorithms, and exposes the decision authority for each routing choice.

Algorithm 1 Rule-class routing for an alignment specification

1: **Input:** candidate behavioral rule r
2: **Step 1.** Classify $r \in \{\text{I}, \text{II}, \text{III}\}$ via an institutionally separate classification body; record the classification with a public rationale.
3: **Step 2.** Attach the decision authority to r: overlapping-consensus argument (Class I), procedural-legitimacy argument for the configuration scheme (Class II), or rights-protective argument (Class III).
4: **Step 3.** Route by class.
Class I: aggregate preference or principle data via existing methods (preference-aggregated reward modeling, principle-labeled RLAIF, or principle-based self-critique).
Class II: configurable default with jurisdiction-level or community-level overrides; technical implementation can use system-prompt-level steering, community-fine-tuned adapters [10], or refusal-with-disclosure of contested status.
Class III: non-aggregative override constraint implemented as principle-based hard constraint; safe-completion approaches [8] and principle-based self-critique against rights-articulating principles are existing technical building blocks.
5: **Step 4.** Disclose, for r, the failure mode (what wrong-class behavior looks like) and the appeal or revision mechanism by which the classification can be revisited.
6: **Step 5.** Document Step 1 through Step 4 in the paper, model card, or model spec accompanying the trained model.
7: **Step 6.** On deployment, expose the disclosure to end users in a form that distinguishes Class I behavior, Class II default-with-override behavior, and Class III hard-constraint behavior.

Worked example, three rules. Table 2 shows the routing layer applied to three concrete rules. The same procedure produces different mechanisms because the rule classes differ; that asymmetry is what the routing step makes visible. A pipeline that converts all three rules into one global policy via preference aggregation has implicitly classified them identically, without disclosure. A full step-by-step trace for the chemical-weapon row appears in Appendix A.

Table 2: Three rules, three different mechanisms. The same routing procedure produces different mechanisms because the rule classes differ; that asymmetry is what the rule-class step makes visible, and what a one-policy pipeline hides. Class III rules carry an appeal log so that hard-constraint decisions are revisable, not terminal. A full step-by-step trace for the chemical-weapon row appears in Appendix A.

Query / rule	Class	Mechanism	Why not aggregate?
Chemical-weapon synthesis assistance	III	hard constraint + appeal log	rights / third-party harm override
Vaccine-policy disagreement	II	configurable default + rationale	reasonable disagreement; no public-reason convergence
CSAM facilitation	I	global prohibition	public prohibition; aggregation eligible

7 The Rule-Class Disclosure Checklist

We propose a five-item disclosure checklist. We do not demand full implementation from every alignment paper; we demand acknowledgment of the routing layer and appropriate disclosure of what has and has not been addressed.

Item one. Classification disclosure. Papers proposing training procedures or principle specifications should declare which class or classes of decision the procedure addresses. CSAM refusal is Class I; contested political topics are Class II; vulnerable-user protection is Class III.

Item two. Procedural authority disclosure. Disclose the procedure by which training signals were produced. Who were the labelers? Who wrote the principles? What were the recruitment, selection, and aggregation procedures? Acknowledging that "we used labelers from [vendor]" without further specification is inadequate for any non-Class-I work.

Item three. Aggregation-method axiomatic disclosure. Following Conitzer et al. [4], papers should make the reward-modeling rule explicit and evaluate it against standard social-choice axioms.

Item four. Rule-class disclosure for Class II and III work. Papers addressing Class II or III decisions should articulate the rule-class justification: why is public imposition of this answer justified given reasonable disagreement (Class II), or what rights-protective consideration justifies overriding aggregate preference (Class III)? The justification should be articulable in political-theoretic vocabulary, not only in ML vocabulary.

Item five. Scope-of-binding disclosure. Disclose the population for which the alignment is claimed binding. An alignment decision legitimate for U.S. users may be illegitimate for Indian or Brazilian users under different constitutional settlements. Frontier labs deploy globally; the constitutional-legitimacy question cannot be answered once for all users.

Adoption pathway. Tier A (procedural, immediately feasible) covers items one, two, and five. Tier B (methodological, two-year horizon) covers item three. Tier C (structural, longer horizon) covers item four and the deliberative infrastructure that supports it. The checklist does not require labs to solve the hard institutional problems before deploying; it requires them to acknowledge the problems and disclose their stance.

8 Objections

This is political philosophy, not pipeline design. The argument is technical in destination even when it is political in origin. The output is a routing layer over training pipelines (Algorithm 1), a five-item disclosure checklist, and concrete implementation building blocks (system-prompt steering, community adapters, safe-completion overrides). Alignment papers at NeurIPS already make normative choices in their data selection, labeler recruitment, principle authoring, and refusal-policy design; the question is whether those choices are made in a vocabulary that exposes them to critique or in a vocabulary that conceals them as preference learning. The Position Paper Track explicitly accommodates cross-disciplinary argument; Conitzer et al. [4] is precedent at ICML.

Asking labs to implement democracy is unrealistic. The legitimate response to "full rule-class infrastructure is unimplementable by private labs" is not "therefore rule-class considerations do not apply" but "therefore rule-class considerations require public institutional innovation beyond private-lab decisions." The EU AI Act, U.S. state-level legislation, and international AI governance proposals are early forms of the public infrastructure alignment's categorical problems require.

Per-user configuration will destroy AI utility. The response is partly empirical and partly conceptual. Per-user configuration is already implemented for style, length, and tone, and Class II is not architecturally different. The framing of degraded utility assumes uniformity is itself utility-providing, conflating user-perceived utility with a particular developer's conception. A Class II approach redirects who bears the cost of disagreement, from users-who-disagree-with-the-default to users-who-must-configure.

Impossibility results apply to voting, not preference learning. Ge et al. [6] address this explicitly. We grant the technical point and respond that the conceptual force of impossibility, that preference

learning over unrestricted domains has no fully-satisfactory solution, transfers to preference learning even where specific theorems require mathematical adaptation. Bradley-Terry-Luce models face structural tensions whose mathematical form differs from Arrow's but whose conceptual character is identical.

You have no empirical demonstration; this is pure philosophy. Position papers establish conceptual applicability, not new empirical benchmarks. The burden is to show that alignment decisions already instantiate authority questions for which social choice and political theory supply the vocabulary; we owe the reader an account of why those literatures apply, not a re-demonstration of them.

Who classifies the classifier? The deepest objection. Rule-class routing itself requires authority: someone or some body has to decide that decision r is Class I, II, or III, and that authority itself is unaccountable on the same grounds we have used to criticize current alignment. We acknowledge the objection and respond as follows. The paper does not solve the authority problem; it makes authority assignment explicit. Current alignment pipelines already classify implicitly: a contested political topic that is forced through preference learning into a single global answer has been classified as Class I by default, by the team that specified the training data, the principle list, or the refusal policy. The classification is happening; the question is whether it is disclosed and revisable. The checklist proposes converting hidden classification into disclosed classification, with an institutional appeal mechanism (Algorithm 1, Step 4). The path to procedural authority is institutional, drawing on the deliberative-democracy literature, citizen assemblies, multi-stakeholder governance bodies, and binding public-comment procedures, instead of purely technical. We do not pretend to have the right institutional answer. We argue that designating the question is an improvement over leaving it implicit.

Closing position. Aggregation over unrestricted value disagreement has no procedure satisfying minimal democratic conditions. The three-class taxonomy, six-step routing, and five-item checklist let reviewers, regulators, and users see where the answer is being given.

References

[1] Anthropic. Collective constitutional AI: Aligning a language model with public input. *Anthropic Blog*, October 2023.

[2] K. J. Arrow. *Social Choice and Individual Values*. Yale University Press, 1951.

[3] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[4] 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 *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, pages 9346–9360, 2024.

[5] G. Gaus. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge University Press, 2011.

[6] L. Ge, Y. Liu, and A. D. Procaccia. Axioms for AI alignment from human feedback. In *Advances in Neural Information Processing Systems 37*, 2024.

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

[8] A. Glaese, N. McAleese, M. Trębacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger, M. Chadwick, P. Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

[9] J. Habermas. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. MIT Press, 1996. Translated by W. Rehg.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[11] H. Landemore. *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton University Press, 2013.

[12] H. Landemore. *Open Democracy: Reinventing Popular Rule for the Twenty-First Century*. Princeton University Press, 2020.

[13] H. Lee, S. Phatale, H. Mansoor, K. R. Lu, T. Mesnard, J. Ferret, C. Bishop, E. Hall, V. Carbune, A. Rastogi, and S. Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[14] OpenAI. Deliberative alignment: Reasoning enables safer language models. *OpenAI Technical Report*, 2024.

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

[16] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36*, 2023.

[17] J. Rawls. *Political Liberalism*. Columbia University Press, 1993.

[18] 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, 1975.

[19] A. K. Sen. The impossibility of a Paretian liberal. *Journal of Political Economy*, 78(1):152–157, 1970.

[20] A. Siththaranjan, C. Laidlaw, and D. Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. *arXiv preprint arXiv:2312.08358*, 2023.

[21] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems 33*, 2020.

A Worked Example: Routing a Medical Misinformation Rule

We walk through the routing layer on a single rule to make the abstraction concrete. The rule is taken from a real category of pipeline decision and routed step by step through the six-step procedure of Algorithm 1.

Candidate rule r. "The model should refuse to provide instructions on how to synthesize a chemical weapon precursor compound, even when asked in the framing of a chemistry homework question."

Step 1 (classification submission). The rule is submitted to the classification step. The submission must declare a candidate class and a public rationale.

Step 2 (institutional separation). The classifier is institutionally separate from the team training the reward model. In a frontier-lab setting this is a policy/safety committee with documented composition and decision-making procedure. The committee reviews the rationale.

Step 3 (classification decision). The committee assigns **Class III**: rights-protective override. The rationale: third-party physical-harm risk is qualitatively different from the user's preferences over content. The reasoning is not that "most users prefer" the refusal; it is that aggregating user preferences is the wrong instrument because the harm to non-users is not represented in any aggregation procedure run over users.

Step 4 (mechanism assignment). The mechanism is a hard constraint with a safe-completion path: refuse the original request, redirect to a non-harmful chemistry pedagogy alternative, and surface a brief explanation of the refusal to the user. Configurability does not apply (Class III).

Step 5 (failure-mode declaration). Two declared failure modes: (a) over-refusal of legitimate chemistry instruction not concerning weapon-precursor synthesis, monitored by a held-out evaluation set of homework-framed legitimate chemistry queries; (b) jailbreak via reframing, monitored by a red-team evaluation set tracking refusal robustness under adversarial prompt rewrites.

Step 6 (appeal mechanism). Researchers, public-health practitioners, and chemistry educators may file a documented request for narrow exception. Each exception is reviewed by the same committee under the same Class III standard. Decisions are logged and the log is reviewable.

Disclosure. The five-item Rule-Class Disclosure Checklist is filed alongside the model release: (i) class assignment (III); (ii) decision authority (rights-protective argument, third-party physical harm); (iii) mechanism (hard constraint + safe completion); (iv) declared failure modes (over-refusal, jailbreak); (v) appeal mechanism (committee, log).

What this example illustrates. The same rule run through the routing layer would end at Class III for chemical-weapon precursors but at Class II for, e.g., contested vaccine policy claims (reasonable disagreement; configurable defaults; deliberatively sourced presentation). It is the routing step that makes that asymmetry visible. A pipeline that converts both rules into the same global policy via preference aggregation has implicitly classified them identically, without disclosure.

B Coding Rubric for the Alignment Method Set (Figure 2)

Figure 2 plots existing alignment methods on two qualitative axes. To make this auditable; not a hand-drawn impression, we declare the coding rubric here. Each axis takes ordinal values $\{0, 1, 2\}$.

X-axis: preference aggregation intensity.

- **0 – no explicit aggregation.** The method does not articulate an aggregation rule over multiple labelers' or users' preferences.
- **1 – implicit / small-sample aggregation.** The method aggregates preferences (typically via a Bradley–Terry-style reward model on pairwise comparisons) but does not engage social-choice axiomatic conditions; aggregation is treated as a learning-from-feedback step.

- **2 – axiomatic / explicit social-choice aggregation.** The method is presented in social-choice terms, references aggregation properties (e.g., independence of irrelevant alternatives, strategy-proofness), and frames training as an axiomatically motivated aggregation procedure.

Y-axis: rule-class routing intensity.

- **0 – no rule-class distinction.** All behavioral specifications enter the same training objective; no separation between preference-learning-eligible and eligibility-determining decisions.
- **1 – principles stated but not routed.** The method introduces a list of principles or a constitution, but principles are processed through the same preference-style learning loop without an institutionally separate classification step.
- **2 – explicit rule classes / overrides / appeal.** The method (i) distinguishes at least two structurally different rule types (e.g., hard constraint vs configurable default), (ii) uses different mechanisms for each, and (iii) supports either a configurable default or a documented appeal/override mechanism.

Table 3: Coding for the Figure 2 method set with one-line justifications. Two-author independent coding (87.5% raw agreement; resolution log in supplement).

Method	X	Y	One-line justification
RLHF [15]	1	0	Bradley–Terry reward; one global policy.
DPO [16]	1	0	Same target object without explicit reward model.
RLAIF [13]	1	0	AI substitutes for human raters; same aggregation.
Constitutional AI [3]	1	1	Principles processed in self-critique loop.
Deliberative Alignment [14]	1	1	Deliberation at chain-of-thought; preference at training.
Collective Const. AI [1]	1	1	Citizen-panel input; preference-learning pipeline.
Conitzer et al. [4]	2	0	Axiomatic social choice; no rule-class step.
Our position	1	2	Routing layer with Class III + appeal.

Coding decisions for Figure 2.

Inter-coder agreement. Two authors coded each cell independently. Of the $8 \times 2 = 16$ entries, 14 agreed on first coding (87.5% raw agreement). The two disagreements were: (i) Constitutional AI on the Y-axis (one coder Y=1, the other Y=2), resolved at Y=1 because principles are not routed through institutionally separate classification; (ii) Deliberative Alignment on the Y-axis (one coder Y=2, the other Y=1), resolved at Y=1 because deliberation occurs at inference-time chain-of-thought, not at a training-pipeline routing step. The resolution log is in the supplement (`landscape_coding_log.txt`).

The rubric is intended to be reusable. The same scheme can be applied to alignment methods proposed after this submission; what changes is whether the field begins to populate the (X=2, Y=2) target zone.