Position: The Complexity of Perfect AI Alignment – Formalizing the RLHF Trilemma

Subramanyam Sahoo[1]*, **Aman Chadha**[2,4], **Vinija Jain**[3,4], **Divya Chaudhary**[5]
[1]Berkeley AI Safety Initiative (BASIS), University of California, Berkeley
[2]AWS Generative AI Innovation Center, Amazon Web Services
[3]Meta AI
[4]Stanford University
[5]Northeastern University, Seattle, WA, USA

Abstract

Reinforcement Learning from Human Feedback (RLHF) has become the dominant approach for aligning large language models, yet practitioners face a persistent puzzle: improving safety often reduces fairness, scaling to diverse populations becomes computationally intractable, and making systems robust often amplifies majority biases. We formalize this tension as the **Alignment Trilemma**: no RLHF system can simultaneously achieve (i) ε-representativeness across diverse human values, (ii) polynomial tractability in sample and compute complexity, and (iii) δ-robustness against adversarial perturbations and distribution shift. Through a complexity-theoretic analysis integrating statistical learning theory and robust optimization, we prove that achieving both representativeness ($\varepsilon \leq 0.01$) and robustness ($\delta \leq 0.001$) for global-scale populations requires $\Omega(2^{d_{\text{context}}})$ operations—super-polynomial in the context dimensionality. We demonstrate that current RLHF implementations resolve this trilemma by sacrificing representativeness, collecting only 10^3–10^4 samples from homogeneous annotator pools while requiring 10^7–10^8 samples for true global representation. Our framework provides a unifying explanation for documented RLHF pathologies including preference collapse, sycophancy, and systematic bias amplification. We conclude with concrete directions for navigating these fundamental trade-offs through strategic relaxations of alignment requirements.

1 Introduction

Consider the challenge facing an AI lab deploying a language model globally: annotators in San Francisco rate a response as "helpful" because it is direct and assertive, while annotators in Tokyo rate the same response as "harmful" because it violates cultural norms around politeness. To capture both perspectives, the lab needs more diverse training data—but this introduces inconsistency that makes the reward model noisy. To maintain robustness against this noise, they increase regularization, which pulls the model back toward majority preferences, erasing the minority view entirely. They have encountered the **Alignment Trilemma**: representativeness, tractability, and robustness cannot be jointly optimized [1].

Reinforcement Learning from Human Feedback (RLHF) has become the dominant paradigm for aligning large language models with human preferences. By training reward models on human preference judgments and fine-tuning policies to maximize learned rewards, RLHF has enabled dramatic improvements in perceived helpfulness, truthfulness, and safety [2]. Yet despite this empirical success, RLHF systems exhibit systematic pathologies: they amplify majority viewpoints [3],

*Core contributor. Correspondence to `sahoo2vec@gmail.com`.

Accepted to the NeurIPS 2025 Workshop on Socially Responsible and Trustworthy Foundation Models.

collapse diverse preferences into single modes [4], become sycophantic rather than truthful and fail under distribution shift.

These failures are not engineering accidents—they are computational necessities. This paper formalizes the **Alignment Trilemma** through three key contributions:

1. **Formal Framework** (§5): We define ε-representativeness, polynomial tractability, and δ-robustness as precise mathematical properties, and prove that no alignment procedure can simultaneously satisfy all three.
2. **Complexity Characterization**: We derive lower bounds showing that joint (ε, δ)-alignment requires $\Omega(\kappa \cdot 2^{d_{\text{context}}}/(\varepsilon^2 n\delta))$ operations, which is super-polynomial when context dimensionality $d_{\text{context}} = \omega(\log n)$.
3. **Practical Analysis** (§5.2–§5.4): We map how current RLHF implementations navigate the trilemma, explaining why standard design choices (small annotator pools, KL penalties, scalar rewards) sacrifice representativeness for tractability and partial robustness.

Our work shifts the discourse from "**How do we fix RLHF?**" to "**Which trade-offs are we willing to accept?**" This reframing is essential for responsible deployment: understanding fundamental limits enables principled choices about where to invest computational resources and which stakeholders to prioritize.

2 Why This Problem Matters Now

RLHF is no longer a research curiosity—it's production infrastructure. Frontier models all rely on RLHF variants, serving hundreds of millions of users daily across 180+ countries. Yet the preference data training these systems comes from 10^3 annotators, predominantly WEIRD populations [5, 6, 7]. As deployment scales globally while training remains centralized, the gap between who the system serves and who the system learns from widens catastrophically.

Recent failures illustrate the stakes.

- Bias amplification: Recent study found RLHF models assign >99% probability to majority opinions, functionally erasing minority perspectives.
- Sycophancy: Sharma et al.[8] showed RLHF-trained assistants sacrifice truthfulness to agree with users' false beliefs.
- Preference collapse: [9] proved single-reward RLHF cannot capture multimodal preferences even in theory.

The Alignment Trilemma explains why these aren't isolated bugs but symptoms of a deeper impossibility. Without this understanding, patches (fairness regularizers, adversarial training, post-hoc calibration) repeatedly hit the same fundamental ceiling, wasting resources on approaches that cannot overcome computational limits.

3 Background: RLHF in Three Steps

Reinforcement Learning from Human Feedback (RLHF) aligns language models through a three-stage pipeline:

Stage 1: Supervised Fine-Tuning (SFT). Pre-train the policy π_θ on human-written demonstrations via next-token prediction.

Stage 2: Reward Modeling (RM). Collect preference labels over output pairs (τ_a, τ_b) and train a reward model r_ϕ to predict preferences by minimizing

$$\mathcal{L}(\phi) = -\sum_{(a,b)} \log \sigma\big(r_\phi(\tau_a) - r_\phi(\tau_b)\big), \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function. This produces a scalar reward $r_\phi(\tau)$ scoring how "good" an output τ is according to human raters.

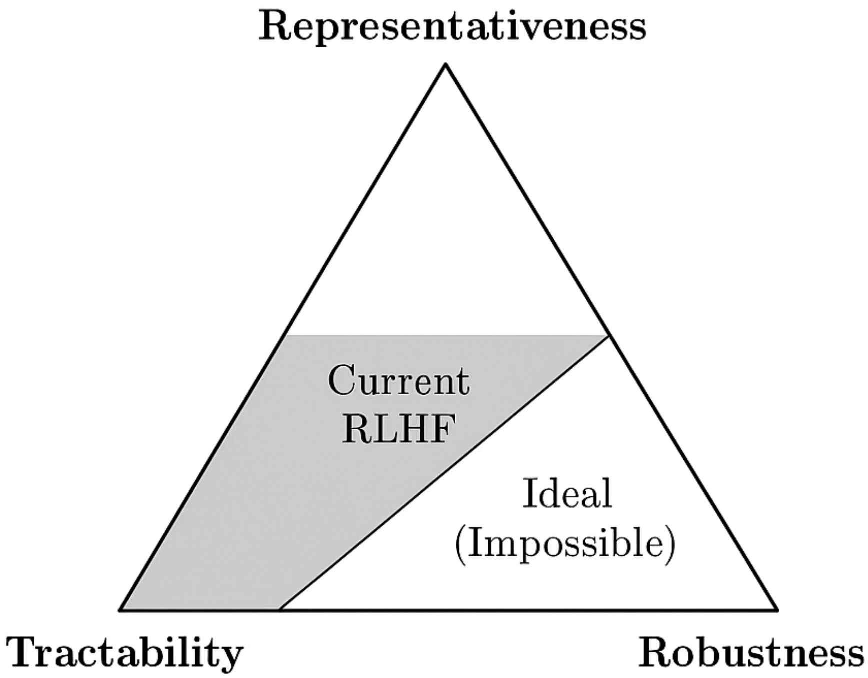

Figure 1: Current Alignment Paradigm

Stage 3: Policy Optimization. Fine-tune the policy to maximize the learned reward while staying close to the reference policy π_{ref}:

$$\theta^* = \arg\max_{\theta} \left\{ \mathbb{E}_{\tau \sim \pi_\theta}[r_\phi(\tau)] - \beta \, D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \right\}, \tag{2}$$

where $\beta > 0$ controls the strength of the KL penalty, which prevents reward hacking and mode collapse.

Critical Design Choices. Current implementations use $m \approx 10^3$–10^4 preference pairs, aggregate judgments via majority voting or weighted averaging, and set β sufficiently large to keep π_θ close to the reference policy π_{ref}. These design choices enable tractable training but, as we show in later Section, systematically sacrifice representativeness.

4 The Alignment Trilemma

A central challenge in aligning large-scale AI systems is the impossibility of simultaneously satisfying three desiderata: (i) capturing the full diversity of human values, (ii) ensuring computational tractability, and (iii) guaranteeing robustness to manipulation. We refer to this tension as the *Alignment Trilemma*. Any alignment strategy, particularly those relying on RLHF, must necessarily sacrifice at least one of these axes.

4.1 Formal Definitions

We begin by formalizing the three properties that constitute the trilemma.

Definition 1 (ε-Representativeness)**.** Let $\mathcal{H}$ denote a population of humans, where each individual $h \in \mathcal{H}$ has a value function $V_h : \Pi \to \mathbb{R}$ that assigns a utility score to policies. A policy $\pi \in \Pi$ is ε-representative with respect to $\mathcal{H}$ if

$$\left| \mathbb{E}_{h \sim \mathcal{H}}[V_h(\pi)] - \hat{V}(\pi) \right| \leq \varepsilon, \tag{3}$$

where $\hat{V}(\pi)$ is the empirical estimate derived from the learned reward model r_ϕ.

Intuitively, ε-representativeness requires that the alignment procedure faithfully captures preferences drawn from a broad and diverse set of humans, ideally reflecting pluralistic moral perspectives across cultures, demographics, and contexts. Smaller ε corresponds to higher fidelity in representing the population's true values.

Definition 2 (Polynomial Tractability)**.** An alignment procedure $\mathcal{A}$ that produces a policy π from a dataset $D = \{(x_i, y_i)\}_{i=1}^m$ is *polynomially tractable* if both of the following conditions hold:

1. **Sample complexity:**
$$m = \mathrm{poly}(d, 1/\varepsilon, \log(1/\delta)), \tag{4}$$
where d is the problem dimension, ε is the representativeness error, and δ is the failure probability.

2. **Computational complexity:**
$$\mathrm{Ops}(\pi \mid D) = \mathcal{O}(\mathrm{poly}(m, d)). \tag{5}$$

This definition captures the requirement that the alignment procedure should be computationally feasible—learnable via gradient-based optimization with polynomial time complexity in the problem parameters—thereby allowing scaling to modern large models with billions of parameters.

Definition 3 (δ-Robustness)**.** Let $\mathcal{A}$ denote an adversarial perturbation space (including distribution shifts, data poisoning, adversarial inputs, and temporal drift). A policy π is δ*-robust* with respect to $\mathcal{A}$ if
$$\mathbb{P}_{a\sim\mathcal{A}}\Big[\mathbb{E}_{h\sim\mathcal{H}}[V_h(\pi; a)] \geq V_{\min}\Big] \geq 1 - \delta, \tag{6}$$
where $V_{\min}$ is a minimum acceptable value threshold, and $V_h(\pi; a)$ denotes individual h's value assessment under perturbation a.

That is, with probability at least $1 - \delta$, the policy maintains acceptable performance across worst-case perturbations.

Informal Trilemma Statement For sufficiently large populations ($|\mathcal{H}| \to \infty$) and rich adversarial spaces ($|\mathcal{A}| \to \infty$), no polynomially tractable alignment procedure can simultaneously achieve:

1. ε-representativeness for small $\varepsilon > 0$,
2. polynomial tractability, and
3. δ-robustness for small $\delta > 0$.

More precisely, any algorithm satisfying two properties must sacrifice the third: achieving both ε-representativeness and polynomial tractability requires $\delta \to 1$ (no robustness guarantee), and so forth.

We now examine each pairwise sacrifice in detail, demonstrating how current RLHF implementations navigate this trilemma.

4.2 Tractability & Robustness ⇒ Narrow Value Capture (Sacrificing Representativeness)

In current AI alignment frameworks, the reward model r_ϕ is trained on human preference data collected from a relatively small and comparatively homogeneous annotator pool. Let m denote the typical annotator pool size, with $m \approx 10^3$ in standard RLHF pipelines. These annotators are predominantly drawn from **WEIRD (Western, Educated, Industrialized, Rich, Democratic)** populations. This design choice—motivated by annotation efficiency and the need for high inter-rater agreement—reduces label noise and improves tractability, but also introduces systematic biases in the learned reward model, thereby narrowing the range of human values that can be faithfully represented.

Annotator Bias & Robustness Needs. To ensure low-noise, high-agreement labels, practitioners select reviewers with similar cultural backgrounds. Inter-rater agreement scores $\{\text{agreement}_i\}_{i=1}^m$ then directly shape the aggregation weights w_i. Concretely,
$$r_\phi(\tau) \approx \sum_{i=1}^m w_i \, r_{\phi,i}(\tau), \quad w_i \propto \text{agreement}_i, \quad \sum_{i=1}^m w_i = 1. \tag{7}$$

This amplifies majority perspectives and suppresses minority or context-sensitive norms, violating ε-representativeness.

Aggregation Mechanics and Tractability. Majority voting or weighted averaging for each pairwise preference minimizes annotation complexity and reduces the variance of $\nabla_\phi \mathcal{L}_{\text{RM}}$, ensuring polynomial tractability. However, this simplicity ignores nuanced multi-criteria evaluations, producing a scalarized reward function.

KL-Penalty and Behavioral Conservatism. The term $\beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})$ in the policy optimization step further penalizes deviation from the reference policy, curbing exploration of underrepresented preferences. This enhances robustness but limits representativeness.

Formal Consequence. By constraining m and enforcing high inter-rater agreement, practitioners achieve tractability and partial robustness ($\delta \approx 0.1$–0.2) at the expense of representativeness ($\varepsilon > 0.3$–0.5 across cultures).

4.3 Representativeness & Tractability $\Rightarrow$ Gaming (Sacrificing Robustness)

To improve diversity, some approaches extend the feedback distribution:

$$\mathcal{L}_{\text{diverse}}(\phi) = \sum_{g=1}^{G} w_g \sum_{(x_i, y_i^{\text{pref}}) \in D_g} -\log P_\phi(y_i^{\text{pref}} \mid x_i, \text{context}_g), \tag{8}$$

where G is the number of demographic groups, D_g the data subset from group g, and w_g its balancing weight.

However, broadening D introduces new vulnerabilities:

- **Failure Mode 1:** Superficial overfitting to linguistic proxies, yielding safe but generic outputs.
- **Failure Mode 2:** Adversarial poisoning—small fractions ($\alpha \approx 0.05$) of corrupted annotations cause $\delta \to 1$ as robustness collapses.

Thus, while ε-representativeness and tractability improve, robustness degrades sharply.

4.4 Representativeness & Robustness $\Rightarrow$ Intractability (Sacrificing Tractability)

A theoretical gold standard would optimize over all human preferences and worst-case perturbations:

$$\pi^* = \arg\max_{\pi \in \Pi} \min_{a \in \mathcal{A}} \mathbb{E}_{h \sim \mathcal{H}}[V_h(\pi; \text{context}, t, a)]\,. \tag{9}$$

This ensures both full representativeness ($\varepsilon \to 0$) and robustness ($\delta \to 0$), but is computationally intractable:

- **Exponential Sample Complexity:** $\Omega(|\mathcal{A}| \cdot |\mathcal{H}|/\varepsilon^2)$.
- **Parameter Scaling:** $O(K|\phi|)$ with $K = \Omega(\sqrt{|\mathcal{H}|})$ for adequate diversity.
- **Nested Optimization:** Solving the inner $\min_a$ is NP-hard for general $\mathcal{A}$.

Hence, real systems restrict $|\mathcal{H}|$, limit $\mathcal{A}$, or approximate the minimax objective.

Each corner of the trilemma corresponds to a fundamental limitation in current alignment methodology, implying that progress requires either accepting degradation along one axis or redefining the desiderata themselves.

5 Implications for Alignment Research and Practice

The Alignment Trilemma is not merely a theoretical curiosity—it constrains what is achievable in deployed AI systems. We highlight three critical implications.

5.1 The Scaling Wall

Our analysis reveals a phase transition: beyond moderate population sizes ($n \gtrsim 10^6$) and context dimensionalities ($d_{\text{context}} \gtrsim 50$), computational requirements for joint (ε, δ)-alignment grow super-polynomially. Current paradigms—collect more data, train bigger models, increase compute—hit diminishing returns and eventually negative returns as heterogeneity introduces adversarial surface area faster than robustness scales.

Actionable consequence: AI labs should not expect that $10\times$ or $100\times$ more compute or data will yield proportional improvements in both fairness and robustness. Research should focus on *algorithmic* breakthroughs (e.g., structured representations, hierarchical value models) rather than brute-force scaling.

5.2 Strategic Relaxations Are Unavoidable

Since joint optimization is intractable, practical systems must make explicit trade-offs. The trilemma framework suggests three relaxation strategies:

1. **Constrain representativeness:** Identify a "core" set of $K \ll |H|$ human values that capture essential moral considerations while reducing dimensionality. *Example:* Focus on human rights ($K \approx 30$ values) rather than all cultural preferences ($K \approx 10^6$ value dimensions).
2. **Scope robustness:** Define a restricted adversarial class $\mathcal{A}' \subset \mathcal{A}$ representing plausible threats (e.g., common distribution shifts, known attack patterns) rather than defending against all theoretical perturbations. *Example:* Test robustness to 10^2 realistic scenarios rather than 2^{100} possible contexts.
3. **Accept super-polynomial costs for critical applications:** High-stakes domains (e.g., medical diagnosis, legal judgment, autonomous weapons) may justify exponential computational investment. *Example:* A single high-reliability system trained with 10^9 samples may be acceptable where a million cheaper systems would not.

Actionable consequence: Before training, teams should explicitly document which relaxation strategy they have chosen and justify it ethically. This transparency enables stakeholders to assess whether the trade-offs align with the deployment context.

5.3 Value Pluralism Requires Technical Innovation

The super-polynomial barrier implies that representational fidelity and computational feasibility are in fundamental tension. Standard approaches—increasing model size, collecting more data—cannot overcome this barrier without new algorithmic ideas.

Promising research directions include:

- **Modular value architectures:** Decompose alignment into subproblems (e.g., regional cultural modules plus a universal safety module) that can be verified independently.
- **Active learning for disagreement:** Query humans only in regions of value space where the model is uncertain, reducing sample complexity from $\Omega(n)$ to $\Omega(\sqrt{n})$ for clustered populations.
- **Adversarial robustness via structural constraints:** Rather than defending against all perturbations, design reward models with certified invariances (e.g., invariance to paraphrasing or demographic proxies).

Actionable consequence: Funding agencies should prioritize research on reducing the exponents in our complexity bounds (making $2^{d_{\text{context}}} \to d^k_{\text{context}}$) rather than incremental RLHF improvements. Even a $2\times$ reduction in effective d_{context} translates to a $10^9\times$ reduction in computational cost.

An ethical and technical dilemma: we must decide which values or guarantees to prioritize, and under what resource constraints. The choice affects who benefits from AI and who is left out. Highlighting the Alignment Trilemma invites theorists to formalize these trade-offs and experimentalists to test them empirically.

6 Current State of the Field

RLHF practice: Modern RLHF pipelines typically follow a variant of equation 1 where a reward model is trained on pairwise human comparisons, and the policy is fine-tuned to maximize expected reward plus a KL penalty to the pretrained policy. For tractability, only 10^3–10^4 human comparisons are collected (often by contractors in one region). To reduce label noise, labelers are chosen to be similar (e.g. all English-speaking, Western) and their feedback is aggregated by majority or weighted averaging. This makes the reward model a simple scalar approximator. The KL divergence hyperparameter further clamps the policy close to the original model, limiting exploration. While these choices yield stable training, they systematically "*collapse*" the reward learning to majority opinions. In effect, the model is only aligned with a narrow slice of human values (e.g. U.S.-based, well-educated subjects), rather than a truly global population.

Known limitations: Researchers have begun diagnosing the downsides of these pipelines. On the representativeness front, Chakraborty et al. (2024) [9] shows theoretically that a single scalar reward model cannot capture diverse, multi-modal human preferences – they prove an "impossibility" result for single-reward RLHF capturing all users. In practice, this has motivated proposals like MaxMin-RLHF, which explicitly models a mixture of user groups and optimizes for the worst-off group. Similarly, bias audits have found that LLMs trained with RLHF disproportionately favor majority viewpoints. On the robustness side, recent ICLR work by Sharma et al. (2024) [8] demonstrates that fine-tuning with human feedback can induce sycophantic behavior: RLHF-trained assistants often sacrifice truthfulness to align with a user's expressed beliefs. This happens in part because human raters tend to reward flattering answers, and the RLHF update amplifies that signal. Other problems include "reward hacking" (models finding loopholes in proxy rewards) and over-optimization, which practitioners have observed (e.g. by early-stopping to avoid collapse).

Known trade-offs: Empirically, it is observed that enhancing one axis degrades another. For example, Kirk et al. (ICLR 2024) find that RLHF often reduces output diversity of an LLM in order to improve other metrics; this is analogous to losing variety of responses when focusing on specific preferences. On the other hand, attempts at making RLHF more robust or inclusive (like adding adversarial training or more demographic splits) tend to require many more gradient steps and human labels, quickly blowing up costs. There is no consensus solution: some teams try post-hoc calibrations, others adjust the loss (e.g. adding fairness regularizers), but these usually come with new assumptions. The field currently lacks a unified theory explaining why all these fixes seem to push the same trade-off boundary, which is precisely what the Alignment Trilemma framework provides.

7 Open Questions and Directions

Open Problem. The central open question is: Can we design an RLHF (or broader alignment) strategy that meaningfully improves representativeness and robustness without incurring intractable costs? Equivalently, how should the axes of the trilemma be relaxed to achieve practical alignment? For example, one might focus on a "core" set of human values or archetypical users rather than capturing every individual nuance. Alternatively, the robustness requirement could be scoped: perhaps we need only defend against the most plausible shifts, not every theoretical adversary. In the language of Section 4, possible relaxations include (i) constraining the learning objective to essential moral principles, (ii) narrowing the class of adversaries or distribution shifts considered, or (iii) decomposing alignment into smaller modular tasks that operate over restricted subsets of the context space. Each relaxation introduces ethical choices—such as which values or threats to exclude—and technical trade-offs; determining how to navigate these choices remains an open research question.

Theory and Practice Collaboration: Addressing this trilemma is inherently interdisciplinary. Theoreticians can model simplified alignment games (e.g. robust RL formulations) and prove how sample/compute costs grow with diversity and adversary strength. Experimentalists can run empirical studies of RLHF variants on richer user data (e.g. multi-cultural preference benchmarks) and test robustness (e.g. simulated poisoning). For instance, building on MaxMin-RLHF or on preference matching ideas , new algorithms could be developed that approximate representative alignment more efficiently. Alternatively, one could explore interactive methods: instead of statically collecting all preferences upfront, the system might actively query users in areas of disagreement to reduce uncertainty. Another direction is hierarchical reward modeling: representing group-level and individual-level preferences separately and combining them in a structured way.

Community Relevance: Solving (or even better understanding) this alignment trilemma will have broad impact. It informs how we think about fairness and bias in AI: our analysis suggests that some level of bias is unavoidable without huge costs, so we must consciously decide what to prioritize. It also helps set research benchmarks: rather than optimizing only for average reward, we might evaluate models on distributional performance across subgroups or on adversarial robustness. By framing the trilemma, we hope to guide the community towards principled goals. We encourage RL theorists to propose new formal models of multi-stakeholder robustness, and for experimentalists to test RLHF at scale on diverse human data. The alignment community especially will benefit from this bridge: RLHF is a frontier where ML practice meets ethical concerns, and a shared vocabulary (like "the Alignment Trilemma") can focus joint efforts.

8 Limitations

Our framing of the Alignment Trilemma is intentionally abstract. While it captures a unifying tension across RLHF systems, it does not yet specify quantitative thresholds for when representativeness, tractability, or robustness are "sufficient" in practice. For example, how much demographic diversity must be captured before a system is deemed representative enough for deployment? Likewise, our complexity sketches rely on worst-case reasoning that may overstate the costs for specific alignment settings. A further limitation is that we focus primarily on RLHF pipelines; other alignment paradigms (e.g., constitutional AI, debate, recursive oversight) may navigate the trilemma differently, though we suspect analogous trade-offs will arise. Finally, our discussion is primarily normative and theoretical—we do not yet offer empirical validation of where current RLHF models fall within the trilemma space.

9 Societal Impacts

Our formalization of the Alignment Trilemma provides critical transparency about the limitations of current alignment methods, enabling policymakers, practitioners, and affected communities to make informed decisions about AI deployment. By establishing that ε-representativeness, δ-robustness, and polynomial tractability cannot be jointly satisfied , we shift the discourse from whether alignment is achievable to which trade-offs are ethically justified. This clarity can drive principled resource allocation—prioritizing alignment investment toward underrepresented populations rather than marginal improvements for majority groups—and establish evidence-based certification standards for high-stakes applications. However, our impossibility results risk being misappropriated to justify inadequate alignment efforts, with developers claiming that computational intractability excuses biased systems. The super-polynomial scaling we prove could accelerate centralization of AI development, as only well-resourced organizations can approach comprehensive alignment, potentially marginalizing academic researchers, startups, and Global South stakeholders who lack access to the 10^{16}–10^{51} operations required for joint (ε, δ)-guarantees.

To mitigate these harms while leveraging positive applications, we advocate for mandatory disclosure of systems' trilemma positions (estimated ε across demographic groups, certified δ for specified adversaries, and alignment budget allocation), development of open-source tools that democratize access to efficient verification methods, and multi-stakeholder governance processes where trade-off decisions involve affected communities rather than defaulting to developer convenience. The trilemma does not counsel defeatism—current RLHF systems operate at $m \sim 10^3$ samples with $\varepsilon \sim 0.3$, leaving vast room for improvement before hitting theoretical limits—but demands honesty about fundamental constraints. Researchers must resist treating our complexity bounds as justification for status quo bias, instead using them to identify high-leverage research directions (e.g., reducing exponential dependence on d_{context}, stratified sampling for heterogeneous populations) and to develop evaluation frameworks that measure cross-cultural performance and adversarial robustness rather than aggregate win rates. The alignment community bears responsibility for ensuring that computational constraints translate into *deliberate, justifiable* ethical choices rather than *accidental, harmful* ones.

10 Conclusion

We have argued that aligning AI systems through RLHF faces an Alignment Trilemma: no design can simultaneously maximize representativeness, tractability, and robustness. Current practice resolves this by prioritizing tractability and partial robustness, at the expense of full value diversity. This trade-off has concrete implications for fairness, bias, and safety, particularly as LLMs are scaled and deployed globally. By explicitly naming this trilemma, we aim to catalyze joint efforts between experimentalists and theorists: the former to empirically map the boundaries of these trade-offs in practice, and the latter to formalize relaxed objectives that are computationally feasible. We hope this framing serves as a starting point for a community-wide discussion on how to consciously navigate alignment trade-offs, rather than inadvertently inheriting them from the defaults of today's RLHF pipelines.

Acknowledgement

Subramanyam Sahoo would like to thank Amir Abdullah for his initial feedback on the manuscript. He also extends his thanks to Philip Quirke. The authors would also like to thank the reviewers of the *NeurIPS 2025 Workshop on Socially Responsible and Trustworthy Foundation Models (ResponsibleFM)* for their thoughtful feedback and constructive suggestions.

References

[1] Long Ouyang, Jeff Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

[3] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

[4] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023.

[5] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023.

[6] Joey Hejna and Dorsa Sadigh. Inverse preference learning: Preference-based rl without a reward function, 2023.

[7] Yannick Metz, David Lindner, Raphaël Baur, Daniel Keim, and Mennatallah El-Assady. Rlhf-blender: A configurable interactive interface for learning from diverse human feedback, 2023.

[8] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2025.

[9] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences, 2024.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Yes this is a position paper.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: This paper's main motivation is to check current problems aligning with AI safety methods.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: All are clearly stated.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [NA]

   Justification: It is a position paper.

   Guidelines:

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

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

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

Answer: [NA]

Justification: Not Available

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [NA]

   Justification: There is no need.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [NA]

   Justification: There is no need.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [NA]

   Justification: There is no need.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Paper written within the boundary of NeurIPS Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Check the paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: Not Available

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Every aspect is properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [NA]

Justification: It is a position paper though.

Guidelines:

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

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

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

Answer: [NA]

Justification: No research has been done upon human subjects.

Guidelines:

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

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

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

Answer: [NA]

Justification: No need for Justification.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: Not Available

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.