Pluralistic AI Alignment Requires Inference-Time Multi-Objective Control

Weichen Li [1] **Mislav Stojanović** [2 3] **Daniel Neider** [2 3] **Marius Kloft** [1] **Sophie Fellenz** [1]

Abstract

Pluralistic AI alignment—accommodating diverse human values rather than a single canonical preference—requires agents to reason under multiple, often conflicting objectives, such as helpfulness, honesty, harmlessness, fairness, and context-specific user preferences. Unlike classical learning methods that optimize a fixed scalar objective, pluralistic alignment requires distinguishing between *objectives* that may be flexibly traded off and *constraints* that should remain non-negotiable. These two categories map onto two existing lines of research: offline multi-objective reinforcement learning provides tools for representing and navigating trade-offs among multiple objectives, and offline safe reinforcement learning formalizes safety-critical constraints and feasible policy regions. A third line, multi-objective LLM alignment, exposes a further requirement: because retraining large models for each preference configuration is infeasible, controllability must shift from training time to deployment time. Taken together, these observations motivate our central position: *inference-time multi-objective control should be a central goal of pluralistic AI alignment*. We argue that unifying these perspectives yields a framework in which training-time learning produces reusable objective and safety representations, while inference-time control enables adaptation to diverse and changing preferences without retraining.

1. Introduction

Many decision-making problems naturally involve *multiple conflicting objectives*. For example, an autonomous driving system must reach its destination efficiently while satisfying requirements related to safety, passenger comfort, and energy efficiency. In a healthcare decision-making system, an AI agent may need to recommend treatments that balance clinical effectiveness, side-effect risk, and cost. In large language models (LLMs), the model usually aims to balance helpfulness, honesty, and harmlessness. These objectives can conflict with one another, and different users may prefer different trade-offs among them. As a result, there is rarely a single universally optimal solution.

This observation has important implications for AI alignment. A common approach is to reduce alignment to optimizing a single scalar reward or preference model. However, human preferences are inherently diverse, context-dependent, and sometimes inconsistent, making such reductions fundamentally limited. Recent work has shown that compressing inherently multi-dimensional preferences into a single reward can produce inconsistent supervision signals and undesirable behaviors (Zhong et al., 2024). More broadly, reward misspecification and preference aggregation introduce systematic challenges in aligning AI systems with complex human values (Taylor, 2016). These limitations suggest that alignment cannot, in general, be reduced to optimizing a single objective.

These challenges point to the need for *pluralistic AI alignment*, in which systems must accommodate multiple, potentially conflicting values and support different trade-offs across users and contexts. In this view, alignment is not about identifying a single optimal behavior, but about enabling systems to adapt over a space of valid trade-offs. Recent work has begun to formalize this perspective (Sorensen et al., 2024; Harland et al., 2024; Vamplew et al., 2024; Guan et al., 2025).

However, many existing alignment methods remain fundamentally *training-centric*. In offline reinforcement learning (RL) and LLM alignment, models are typically trained using pre-specified objective signals, preference weights, or constraint thresholds. While such approaches can capture average or anticipated preferences, they struggle to generalize to preference configurations that differ from those observed during training (Dai et al., 2024; Zhu et al., 2023; Lin et al., 2024b). In practice, user preferences, social norms, and environmental conditions often change after deployment, making it infeasible to specify all desired behaviors

[1]RPTU University of Kaiserslautern-Landau, Germany [2]Research Center for Trustworthy Data Science and Security of the University Alliance Ruhr, Germany [3]TU Dortmund University, Germany. Correspondence to: Weichen Li <weichen@cs.uni-kl.de>, Sophie Fellenz <fellenz@cs.uni-kl.de>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

in advance. Training separate models for each preference profile or context is computationally expensive and unrealistic at scale. These limitations indicate that training-time alignment methods, such as Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022), cannot fully address the pluralistic alignment requirements.

At the same time, not all objectives are equally flexible. In many applications, certain requirements—such as avoiding harmful, unsafe, or illegal behavior—must be treated as non-negotiable constraints. Safe RL addresses this by formulating decision-making as constrained optimization, where a primary objective is maximized subject to safety budgets or feasibility constraints (Altman, 1995; Achiam et al., 2017; Stooke et al., 2020). Similar patterns appear in LLM alignment, where safety mechanisms such as refusal policies, filtering, or cost models are used to bound harmful behavior (Dai et al., 2024; Chen et al., 2025b). However, these approaches typically enforce constraints at training time or with fixed thresholds, limiting their ability to adapt to changing deployment conditions.

Taken together, these observations reveal a fundamental tension. On the one hand, alignment must support flexible trade-offs among multiple, preference-dependent objectives that vary across users and contexts. On the other hand, safety-critical constraints must remain reliably enforced and cannot be arbitrarily relaxed. Current training-time approaches do not simultaneously support both requirements: they either fix preferences during training or encode constraints that are difficult to adjust after deployment. As a result, they are not well suited for pluralistic, real-world settings.

This tension implies a structural requirement for alignment methods: systems must support *adaptive control of preference-dependent objectives at deployment time*, while ensuring that safety constraints remain satisfied. This leads to our central position:

Offline multi-objective alignment should move beyond training-time optimization toward inference-time control of preference trade-offs under explicit safety constraints.

We develop this position by integrating insights from three research areas: offline multi-objective reinforcement learning (MORL), offline safe RL, and multi-objective LLM (MO-LLM) alignment. These areas are often studied separately, but they address closely related aspects of the same problem. Offline MORL provides tools for representing and navigating trade-offs among multiple objectives, often through Pareto-optimality or preference-conditioned policies (Roijers et al., 2013; Zhu et al., 2023). Offline safe RL formalizes safety-critical constraints and feasible policy regions under limited data coverage (Liu et al., 2023; Yao et al., 2024). MO-LLM highlights the practical necessity of post-training controllability, since retraining large models for each preference configuration is infeasible (Shi et al., 2024; Fu et al., 2025; Guan et al., 2025).

A common feature across these areas is the reliance on *offline* data. In many domains, including healthcare, robotics, and language modeling, online trial-and-error learning is infeasible, unsafe, or overly expensive. Agents must therefore learn from fixed datasets and generalize to new situations at deployment time (Levine et al., 2020). This setting introduces shared challenges, including distributional shift, limited data coverage, and uncertainty when evaluating unseen policies. These shared constraints further motivate the need for methods that enable controlled adaptation at inference time.

To support this position, we make three arguments that form a logical chain:

- *Training-time alignment cannot scale to pluralistic deployment.* When user preferences are diverse and context-dependent, training-time methods must either overfit to a fixed preference distribution or retrain for every new profile. Neither scales: for large models, the cost of retraining per preference configuration is prohibitive, and the space of possible preferences is too large to enumerate in advance.
- *Pluralistic alignment requires simultaneously supporting flexible trade-offs and non-negotiable constraints—a requirement no single existing paradigm satisfies.* Offline MORL exposes flexible preference trade-offs but lacks formal constraint enforcement; offline safe RL enforces hard constraints but cannot adapt preference weights at deployment time; MO-LLM methods achieve inference-time adaptability but without formal safety guaranties. Ethical decision-making sharpens this gap: moral objectives are too contested and culturally variable to fix as hard constraints, yet too consequential to leave to unconstrained user preference, demanding inference-time control over the trade-off structure itself.
- *Unifying MORL, Safe RL, and MO-LLM alignment is therefore necessary, not merely useful.* The three areas address complementary parts of the same problem: MORL contributes trade-off representation, Safe RL contributes constraint formalism, and MO-LLM alignment provides both the practical evidence that inference-time control is required and emerging mechanisms for achieving it. Pluralistic alignment cannot be achieved by extending any one of these areas alone; it requires their integration.

2. Conceptual Framework and Background

This section first introduces the conceptual framework underlying our position and then reviews the three research domains that support it: offline MORL, offline safe RL, and MO-LLM alignment. We do not aim to provide a comprehensive survey; instead, we use this background to clarify the role each area plays in an inference-time pluralistic alignment framework.

2.1. Conceptual Framework

Our conceptual framework has two stages: *offline training* and *inference-time alignment*. During offline training, the goal is not to learn a single fixed policy or a single scalar preference model. Instead, methods should learn flexible representations of multiple objectives and safety constraints from pre-collected data, so that preference weights or constraint thresholds can be updated after training. At inference time, users can influence the agent's behavior through deployment-specific preferences and contexts. The framework should allow flexible control over *negotiable objectives*, such as *helpfulness*, *humor*, or other *non-safety-critical* preferences. At the same time, it should enforce *non-negotiable constraints*, such as avoiding *illegal behavior*, *privacy violations*, or *unsafe actions*. Thus, inference-time alignment should not mean unconstrained user control; rather, it should mean controlled adaptation within a safety-feasible region.

Figure 1 illustrates this two-stage view and the roles of the three research areas studied in this paper. Offline MORL provides methods for learning and navigating trade-offs among multiple objectives. Offline safe RL provides tools for learning safety constraints and feasible policy regions from offline datasets. Recent work in MO-LLM alignment has increasingly emphasized decoding-time alignment, motivating offline RL decision-making systems to consider post-training control as well. Together, these areas support the unified position of this paper: inference-time alignment should be central to offline pluralistic alignment under safety constraints.

This framework organizes the rest of the paper. Section 3 argues that training-time alignment alone is insufficient. Section 4 argues that pluralistic alignment requires both flexible trade-offs and non-negotiable constraints. Section 5 argues that unifying offline MORL, offline safe RL, and MO-LLM alignment is necessary for this goal.

2.2. Background of the Three Pillars

In the following, we review the background of three areas: offline MORL, offline safe RL, and MO-LLMs, to clarify the role each plays in the conceptual framework.

Why the Offline Setting Matters. A common feature of the three areas in this paper is that they are all studied in the *offline setting*, which refers to learning from a pre-collected dataset without additional interaction during training. This setting is particularly important when online exploration is costly, unsafe, or infeasible—as in healthcare, autonomous systems, and many alignment settings—but also when the model itself is too large for iterative interaction with an environment.

Offline RL studies how to learn policies from a fixed dataset without additional environment interaction (Levine et al., 2020). The learner is given a dataset $\mathcal{D}_{\mathrm{RL}} = \{(s_t, a_t, r_t, s_{t+1})\}_{t=1}^{N}$, collected by one or more behavior policies, and must learn a policy that performs well after deployment. Compared to online MORL and safe RL, offline MORL and safe RL have received less attention, despite their practical importance in settings where environment interaction is expensive or risky.

In the LLM setting, the offline assumption is not merely a practical convenience but a structural necessity. Many LLMs contain hundreds of billions of parameters and require enormous computational resources to train; once deployed, they are usually not updated iteratively in response to individual interactions. This creates a strict separation between the *training phase*—in which the model is aligned from a fixed, human-annotated preference dataset—and the *deployment phase*, in which the frozen model serves users.

Learning exclusively from offline data introduces shared challenges across all three domains, including limited coverage of the state–action space and uncertainty when extrapolating to unseen behaviors. Although the data modalities differ—environment trajectories versus textual comparisons—the underlying challenge of generalization beyond the dataset support is common to all three settings.

Offline MORL. Offline MORL aims to learn policies that support multiple objective trade-offs from a fixed dataset, where each trajectory $\tau = (s_1, a_1, \mathbf{r}_1, \ldots, s_T, a_T, \mathbf{r}_T)$ is conditioned on preference weights $w \in \mathbb{R}^m$, where m is the number of objectives.

MORL is commonly formulated as a multi-objective Markov decision process (MOMDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathbf{r}, \gamma \rangle$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}$ is the transition function, γ is the discount factor, and $\mathbf{r}(s,a) \in \mathbb{R}^m$ is a vector-valued reward containing $m \geq 2$ objectives. For a policy π, the expected discounted vector return is

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}(s_t, a_t)\right] = [J_1(\pi), \ldots, J_m(\pi)].$$

Unlike single-objective RL, where policies are compared according to one scalar return, MORL must reason about

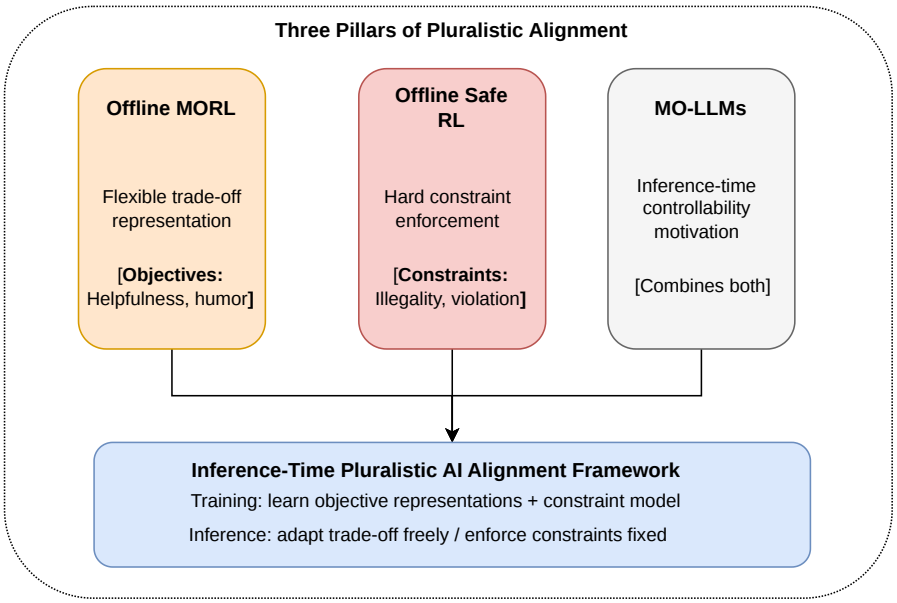

Figure 1. Conceptual framework. Offline training learns objective and constraint representations; inference-time alignment adapts negotiable objectives while enforcing non-negotiable constraints.

trade-offs among multiple objectives. MORL has a long research history, and existing surveys cover its motivations, methods, and metrics (Hayes et al., 2022). A common approach is to introduce a utility or scalarization function f that maps a vector return to a scalar value. For example, linear scalarization uses a weight vector $w \in \mathbb{R}^m$:

$$r_w(s, a, s') = \sum_{i=1}^{m} w_i \, \mathbf{r}_i(s, a), \qquad w_i \geq 0, \quad \sum_i w_i = 1.$$

The choice of w determines the preferred trade-off among objectives. Other scalarization or aggregation criteria, such as Chebyshev scalarization, Nash social welfare, and max–min fairness, encode different assumptions about how objectives should be balanced.

Existing offline MORL methods include *preference-conditioned* methods (Zhu et al., 2023; Lin et al., 2024b; Yuan et al., 2024), *fairness-based* methods (Kim et al., 2025), and *constrained MORL* methods (Lin et al., 2024a). Most prior work assumes that *preference conditioning* over objectives is available during training. For example, in preference-conditioned methods, the model is conditioned on preference weights together with the state and action inputs. Instead of relying on linear scalarization, *fairness-based* methods capture fairness considerations across objectives using methods such as Nash social welfare or max-min optimization. The motivation behind *constrained MORL* (Huang et al., 2022; Lin et al., 2024a; Wu et al., 2021) is that, although users may have different preferences over multiple objectives, safety considerations require that specific objectives be constrained to prevent unacceptable behavior. This line of work should be further explored: compared with the broader literature on MORL and Safe RL, relatively few studies have investigated constrained MORL.

Offline Safe RL. Offline Safe RL, by contrast, aims to learn a reward-maximizing policy that satisfies safety constraints using only pre-collected data, often of the form $\mathcal{D}_{\text{safe}} = \{(s_t, a_t, s_{t+1}, r_t, c_t)\}_{t=1}^{N}$.

Safe RL is commonly formalized as a constrained Markov decision process (CMDP) (Altman, 1995): $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \mathbf{c}, \gamma \rangle$, where $\mathbf{c}$ is the cost function. A CMDP augments the standard reward with one or more cost functions and constrains the expected cumulative cost. In the simplest case, the objective is

$$\max_{\pi} J_r(\pi) \quad \text{s.t.} \quad J_c(\pi) \leq d,$$

where $J_r(\pi)$ is the expected reward return, $J_c(\pi)$ is the expected cost return, and d is a safety budget, which is usually pre-defined during training. This formulation differs from standard MORL in an important way: it distinguishes between objectives that may be optimized and constraints that should not be violated. Common approaches include Lagrangian and primal–dual methods (Stooke et al., 2020), constrained policy updates (Achiam et al., 2017), and shielding mechanisms that monitor or correct unsafe actions (Alshiekh et al., 2018).

Several works frame offline safe RL as a sequence modeling problem. The Constrained Decision Transformer (Liu et al., 2023) adapts the Decision Transformer architecture by conditioning the policy on both target returns and safety

thresholds. OASIS (Yao et al., 2024) is a model-based approach that employs a diffusion model conditioned on both reward and cost to generate safe, high-return synthetic trajectories

LLM Alignment via Preference Learning. Aligning LLMs with human values is typically framed as learning from a fixed dataset of pairwise human preferences. Because many LLMs are large and expensive to update during deployment, much of the alignment signal is typically extracted offline from pre-collected preference datasets $\mathcal{D}_{\text{LLM}} = \{(x^i, y_+^i, y_-^i)\}_{i=1}^N$, where y_+^i is the human-preferred response and y_-^i the dispreferred one for prompt x^i.

The dominant pipeline, RLHF (Christiano et al., 2017; Ouyang et al., 2022), proceeds in two stages. First, a *reward model* r_ϕ is trained to predict human preferences by maximizing the Bradley–Terry likelihood:

$$\mathcal{L}_{\text{BT}}(\phi) = -\mathbb{E}_{(x,y_+,y_-)\sim\mathcal{D}}\big[\log \sigma\big(r_\phi(x,y_+) - r_\phi(x,y_-)\big)\big].$$

Second, the language model π_θ is fine-tuned to maximize the learned reward while staying close to a reference policy π_{ref}:

$$\max_{\pi_\theta} \; \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)}\big[r_\phi(x,y)\big] - \beta\, D_{\text{KL}}\big(\pi_\theta(\cdot|x)\|\pi_{\text{ref}}(\cdot|x)\big).$$

An alternative, *Direct Preference Optimization* (DPO) (Rafailov et al., 2023), bypasses reward modeling by reparameterizing the problem and optimizing the policy directly on $\mathcal{D}_{\text{LLM}}$, avoiding the two-stage pipeline at the cost of reduced flexibility.

Extending this single-objective framework to *multiple* human values—such as helpfulness, harmlessness, and honesty—is non-trivial, because different users or use cases weight these objectives differently. This is precisely the gap that motivates MO-LLM alignment.

Multi-objective LLM Alignment. Multi-objective LLM alignment methods can be viewed from two perspectives. Table 1 summarizes the previous work.

The first is an offline *MORL perspective*, where alignment desiderata such as helpfulness, honesty, and harmlessness are treated as separate reward dimensions, and the goal is to represent trade-offs among them. One line of work trains fine-grained reward frameworks with multiple reward models that capture different perspectives (Wu et al., 2023). A second line of work draws inspiration from *preference-conditioned* methods in offline MORL. In the LLM setting, this idea translates into incorporating preference vectors directly into the prompt or optimization process (Yang et al., 2024b; Guo et al., 2024). More recent work has moved toward aligning preferences at *decoding time*, enabling models to adapt flexibly without retraining (Shi et al., 2024; Jang et al., 2024; Rame et al., 2023).

The second is an *offline Safe RL perspective* (Dai et al., 2024; Wachi et al., 2024; Chen et al., 2025b), where one objective, such as helpfulness, is treated as the primary reward, while another, such as harmlessness, is treated as a constraint. One standard way is to train explicit reward and cost models to guide generation. Others, such as SACPO (Wachi et al., 2024), instead apply direct preference optimization to learn the policy directly from preference data, without requiring reward or cost models.

3. Training-Time Alignment Is Insufficient for Pluralistic Deployment

In offline MORL, recent methods commonly learn preference-conditioned policies or planners, where behavior is controlled by explicit preference weights or implicit welfare criteria (Zhu et al., 2023; Lin et al., 2024b; Yuan et al., 2024; Kim et al., 2025). These methods are useful for representing trade-offs, but many still rely on training-time conditioning over a limited set of preferences. As a result, they may struggle when deployment-time preferences differ from those observed during training.

The training-time framing has several limitations for pluralistic alignment (Li et al., 2026): i) policies may overfit to the preference weights observed during training and generalize poorly to novel or shifted preferences; ii) incorporating new objectives often requires retraining entire models, which is costly and impractical, especially in large-scale systems; iii) collecting preference-weighted offline data is expensive and may be unintuitive for users.

A similar issue appears in offline safe RL, where the goal is to learn policies that maximize reward while satisfying safety constraints from fixed datasets. However, constraint thresholds are often predefined during training. Once training is complete, adapting these thresholds at deployment time usually requires retraining or substantial policy updates. This is problematic because real-world constraints are often dynamic rather than fixed. For example, speed limits may change according to traffic conditions or local regulations. Therefore, several recent works have begun to study real-time budget inference, such as generating trajectories with diffusion models conditioned on dynamic constraint budgets (Lin et al., 2023). These approaches suggest that inference-time adaptation is necessary when safety requirements vary across contexts.

These limitations are especially severe for LLM alignment. As shown in Table 1, methods that do not support inference-time adaptation, such as MORLHF (Wu et al., 2023), require training a separate LLM for each preference configuration. This is computationally expensive, particularly

Method	Inf. Adapt	# LLMs	RM / CM Free	Evaluated Dataset
MORL view (helpfulness and harmlessness as separate rewards)				
MORLHF (Wu et al., 2023)	×	# pref.	×	QA-feedback
MODPO (Zhou et al., 2024)	×	# pref.	✓	BeaverTails, QA-feedback
CPO (Guo et al., 2024)	×	1	×	UltraFeedback, UltraSafety, HH-RLHF
RiC (Yang et al., 2024b)	✓	1	×	HH-RLHF, Reddit Summary †
DPA (Wang et al., 2024a)	✓	1	×	HelpSteer, UltraFeedback
Panacea (Zhong et al., 2024)	×	1	✓	BeaverTails
RS (Rame et al., 2023)	✓	# obj.	×	HH-RLHF, News, Reddit, StackEx, IMDB †
MOD (Shi et al., 2024)	✓	# obj.	✓	Reddit Summary, HH-RLHF, BeaverTails
MetaAlign (Yang et al., 2024a)	✓	1	✓	RiC+HH-RLHF, UltraFeedback, IMHI
PAD (Chen et al., 2025a)	✓	1	✓	HelpSteer2, BeaverTails, RiC+HH-RLHF
MCA (Fu et al., 2025)	✓	0	✓	HH-RLHF, BeaverTails
Safe RL view (helpfulness as reward, harmlessness as cost)				
Safe RLHF (Dai et al., 2024)	×	1	×	SafeRLHF
SACPO (Wachi et al., 2024)	×	# obj.	✓	BeaverTails
SAP (Chen et al., 2025b)	✓	1	✓	BeaverTails

Table 1. Summary of LLM alignment methods from MORL and Safe RL perspectives. "Inf. Adapt" indicates whether user preferences can be adjusted at inference time. "# LLMs" denotes the number of trained models; "# pref." is the number of preferences and "# obj." is the number of objectives. "RM / CM Free" indicates whether general reward or cost models are required. † methods additionally evaluate image-text or text-image tasks.

Dataset	**Alignment objectives**
BeaverTails (Ji et al., 2023)	Helpfulness; harmlessness
HH-RLHF (Bai et al., 2022)	Helpfulness; harmlessness
RiC+HH-RLHF (Yang et al., 2024b)	Harmlessness; helpfulness; humor
QA-feedback (Wu et al., 2023)	Relevance; factuality; completeness
Reddit Summary (Stiennon et al., 2020)	Summary quality; faithfulness
IMHI (Yang et al., 2024a)	Correctness; informativeness; professionalism
UltraFeedback (Cui et al., 2023)	Instruction following; honesty; truthfulness; helpfulness
HelpSteer (Wang et al., 2024b)	Helpfulness; correctness; coherence; complexity; verbosity

Table 2. Evaluation datasets and their alignment objectives.

because training and fine-tuning LLMs are costly and the space of possible preferences can be large. Although pre-training and fine-tuning are necessary, they are insufficient for pluralistic deployment.

Our argument is therefore not that training-time alignment should be abandoned. Rather, training-time alignment can serve as the foundation for inference-time control. Offline multi-objective alignment should aim to produce models that can support flexible preference trade-offs at inference time without retraining from scratch, while keeping safety-critical behavior within explicit constraints. Inspired by (Shi et al., 2024), one possible method is to use separate objective models or objective representations that can support decoding-time alignment according to user preferences. Possible research questions could include evaluating different value-combination methods. Linear combination is the most straightforward approach; other methods also need to be explored, for example, the fairness-based approaches.

We therefore argue that research should move toward methods that support inference-time adaptation of both preference weights and constraint thresholds. Such methods would make AI systems more deployable in real-world settings, where user preferences and safety requirements may shift after training. Inference-time alignment is already an emerging trend in MO-LLM research; it remains, however, an urgent and underexplored direction in the offline RL community.

4. Pluralistic Alignment Requires Both Flexible Trade-offs and Non-Negotiable Constraints

Pluralistic alignment faces a structural tension: some objectives should be flexibly traded off according to user and context-specific preferences, while others must remain non-negotiable constraints that cannot be relaxed—yet existing methods in offline MORL, offline safe RL, and MO-LLM alignment treat these two requirements separately rather than simultaneously. Existing benchmarks in offline MORL, offline safe RL, and MO-LLM alignment typically emphasize either flexible trade-offs or constraint satisfaction, but rarely both at the same time. In offline MORL, common benchmarks, such as MO-MuJoCo (Zhu et al., 2023), evaluate agents on their ability to trade off objectives, such as *speed* and *energy efficiency*. In offline safe RL, benchmarks such as Bullet-Safety-Gym (Liu et al., 2024; Gronauer, 2022) focus instead on maximizing the reward while minimizing safety violations, such as leaving a safe region. A

similar issue appears in MO-LLM alignment. As shown in Table 2, existing benchmarks often include multiple alignment objectives, such as *helpfulness*, *harmlessness*, *honesty*, *completeness*, or *humor*. However, these objectives are usually treated uniformly, either as dimensions to be traded off in a MORL-style formulation or as reward–cost signals in a safe RL-style formulation. As a result, current benchmarks often do not clearly distinguish between objectives that may be flexibly adjusted according to user or context-specific preferences and constraints that should remain non-negotiable.

This distinction is important for real-world pluralistic alignment because it occurs in many applications. For example, in medical decision support (Montori et al., 2023), doctors and patients may express different preferences over treatment effectiveness, side-effect risk, cost, and quality of life. However, safety-critical constraints, such as illegal prescriptions or unsafe dosages, should be non-negotiable. In autonomous driving, users may prefer faster or more energy-efficient driving, but compliance with traffic laws should not be negotiable. Although existing work has begun to study related problems (Huang et al., 2022; Wu et al., 2021), it has not been widely extended to inference-time control or MO-LLM alignment.

We therefore argue that pluralistic alignment requires more benchmarks and methods that simultaneously model both sides of the problem: flexible control over negotiable objectives and reliable enforcement of non-negotiable constraints.

5. Pluralistic Alignment Requires Unifying MORL, Safe RL, and MO-LLM Alignment

The two preceding arguments have identified a three-way gap. Section 3 showed that training-time alignment is insufficient: pluralistic deployment requires methods that can adapt preference weights and constraint thresholds after training. Section 4 showed that no existing paradigm covers all the components this requires simultaneously: offline MORL provides flexible trade-off representation but lacks formal constraint enforcement; offline Safe RL enforces hard safety constraints but cannot adapt preference weights at deployment time; MO-LLM alignment achieves inference-time controllability but without the formal safety guarantees that non-negotiable constraints demand. Because each area covers a different and complementary component of pluralistic alignment, extending any one of them alone cannot close the full gap. Their integration is therefore necessary.

5.1. A Common Structure

Despite differences in terminology and application domains, the three areas can be cast as instances of a single optimization template. Let π denote a policy (or language model), $\mathbf{J}(\pi) = [J_1(\pi), \ldots, J_m(\pi)]$ denote a vector of expected objective returns, and $\mathcal{D}$ a fixed offline dataset. The shared problem is

$$\max_{\pi \in \Pi_{\mathcal{D}}} f\big(\mathbf{J}(\pi),\, w\big) \quad \text{s.t.} \quad J_{c_j}(\pi) \le d_j, \quad j = 1, \ldots, k, \tag{1}$$

where $f(\cdot, w)$ is a scalarization function parameterized by preference weights $w \in \mathbb{R}^m$, each $J_{c_j}(\pi)$ is the expected return of a constraint objective with budget d_j, and $\Pi_{\mathcal{D}}$ encodes a behavioral constraint that keeps the policy within the support of $\mathcal{D}$. Table 3 shows how each domain instantiates this template.

Two structural observations follow directly from this template. First, the KL penalty in LLM alignment plays the same role as pessimism in offline RL: both constrain the learned policy to remain close to the data distribution, addressing distributional shift without online interaction. Second, the boundary between objectives and constraints is not fixed: offline MORL sets $k{=}0$ and treats all desiderata as trade-offs, while offline Safe RL sets $m{=}1$ and encodes safety as hard constraints. Constrained MORL and safety-aware MO-LLMs occupy the interior of this space, with $m \ge 2$ and $k \ge 1$ simultaneously—the regime that pluralistic alignment requires but that remains underexplored in both communities.

We therefore argue that offline MORL, Safe RL, and MO-LLM alignment should not be treated as separate research areas. Offline multi-objective alignment sits at their intersection. Progress requires methods that combine MORL-style trade-off representation, Safe-RL-style constraint enforcement, and LLM-style inference-time controllability. This unified view is necessary for building AI systems that are flexible enough to support pluralistic preferences, yet constrained enough to remain safe.

5.2. Research Directions

Offline constraints are the common bottleneck. Across all three domains, the offline setting sharply limits what can be guaranteed. In offline RL (MORL or Safe RL), distribution shift arises when learned policies choose actions outside the data support; in MO-LLMs, the analogous risk occurs when models extrapolate beyond the preference/rule distribution captured by annotation datasets. This suggests a research direction: inference-time alignment should estimate when a requested trade-off lies outside the reliable region of the offline data. For example, conservatism and pessimism in offline RL may inspire refusal and conservative decoding in MO-LLMs; feasibility estimation in CMDPs can inspire explicit safety or refusal modeling; and preference condi-

Domain	m	k	Scalarization f	Behavioral constraint $\Pi_{\mathcal{D}}$
Offline MORL	≥ 2	0	$w^\top \mathbf{J}$ or Pareto front	Pessimistic policy class
Offline Safe RL	1	≥ 1	$J_r(\pi)$	Pessimistic policy class
MO-LLM	≥ 2	≥ 0	$w^\top \mathbf{J} - \beta D_{\mathrm{KL}}(\pi \| \pi_{\mathrm{ref}})$	KL penalty from reference model

Table 3. The three domains as instances of the unified template (Eq. 1). Here, m denotes the number of objectives, k denotes the number of constraints, and w denotes the preference weight vector.

tioning in MORL corresponds to controllable generation mechanisms.

Constrained MORL for pluralistic yet safe alignment. A recurring gap is the lack of methods that support diverse preferences while maintaining a robust safety envelope. Constrained MORL provides a natural bridge: optimize trade-offs among user-facing objectives while enforcing costs or safety constraints. The idea can be applied in both RL and MO-LLMs alignments.

Unified benchmarks for sequential, multi-objective alignment. Benchmarking remains fragmented. Developing benchmarks that combine long-term sequential structure, multiple objectives, and realistic offline data—potentially in both embodied and language-based environments—remains an important open challenge. Continuous-control suites offer clean objective signals but capture only a limited range of trade-offs involving ethical values, social norms, and human preferences; MO-LLM benchmarks capture rich human judgments but often lack long-horizon interaction structure and standardized evaluation.

6. Alternative Views

LLM alignment is not simply an RL problem. One objection is that LLM alignment is fundamentally different from RL: it often relies on static human-preference datasets rather than repeated interaction with the environment, and it optimizes language generation rather than classical sequential decision-making policies. We do not argue that LLM alignment should be reduced to classical RL, or that RL algorithms should be directly applied to LLMs. Instead, our argument is structural: offline MORL, offline safe RL, and MO-LLM alignment all involve learning from fixed data under competing objectives. The goal is therefore not algorithmic equivalence, but targeted transfer of knowledge across these areas.

Human preferences may be too inconsistent to learn reliably from offline data. A reasonable concern is that offline preference data may be noisy, annotator-dependent, culturally variable, and non-stationary. If the dataset poorly represents human preferences, then learned reward models, objective representations, and trade-off surfaces may be unreliable, and inference-time control cannot fully correct this problem. We agree that inference-time control is not a substitute for high-quality, diverse data. Rather, as in offline RL and LLM alignment more broadly, reliable data remains a foundation for learning meaningful objective and safety representations.

Parameter-efficient fine-tuning may make per-user retraining feasible. Parameter-efficient fine-tuning methods, such as LoRA or adapters, reduce the cost of adapting models to a new preference profile. However, in pluralistic alignment, the space of possible preference combinations is large, and new objectives or constraints may arise after deployment. Even lightweight adapters require additional computation and energy. In contrast, inference-time control aims to adjust preference trade-offs without additional training, making adaptation more flexible and energy-efficient.

Inference-time control is bounded by the training distribution. A technical limitation is that a frozen model can only control trade-offs that it has learned to represent during training. One possible direction is to train offline models with rich objective and constraint representations, together with uncertainty or coverage estimates that indicate when a requested trade-off lies outside the reliable region. This objection is precisely why inference-time control under offline constraints is challenging and deserves further research.

7. Conclusion

In the real world, many decision-making tasks involve multiple, context-dependent objectives, while human preferences are diverse and dynamic. This makes pluralistic alignment difficult to solve with a single training-time objective. In this paper, we argue that pluralistic alignment research should move beyond training time optimization toward inference-time control. Achieving this goal requires closer interaction between offline MORL, offline safe RL, and MO-LLM alignment. We hope this position encourages cross-domain research toward AI agents that are both flexible enough to reflect diverse human values and constrained enough to ensure safety.

Acknowledgements

The authors acknowledge support by the DFG through FOR 5359 (ID 459419731), TRR 375 (ID 511263698), SPP 2298 (IDs 441826958 and 441826958), and SPP 2331 (IDs 441958259, 553345933, and 466468799), by the Carl-

Zeiss Foundation through the initiative AI-Care, and by the BMFTR award 01IS24071A.

References

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In *International Conference on Machine Learning*, volume 70, pp. 22–31, 2017.

Alshiekh, M., Bloem, R., Ehlers, R., Könighofer, B., Niekum, S., and Topcu, U. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Altman, E. Constrained markov decision processes. Technical Report RR-2574, INRIA, Sophia Antipolis, France, 1995.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Chen, R., Zhang, X., Luo, M., Chai, W., and Liu, Z. PAD: Personalized alignment at decoding-time. In *International Conference on Learning Representations*, 2025a.

Chen, X., As, Y., and Krause, A. Learning safety constraints for large language models. In *International Conference on Machine Learning*, 2025b.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 2017.

Cui, G., Yuan, L., Ding, N., Yao, G., Zhu, W., Ni, Y., Xie, G., Liu, Z., and Sun, M. Ultrafeedback: Boosting language models with high-quality feedback. 2023.

Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., and Yang, Y. Safe RLHF: Safe reinforcement learning from human feedback. In *International Conference on Learning Representations*, 2024.

Fu, T., Hou, Y., McAuley, J., and Yan, R. Unlocking decoding-time controllability: Gradient-free multi-objective alignment with contrastive prompts. In *Nations of the Americas chapter of the Association for Computational Linguistics*, 2025.

Gronauer, S. Bullet-safety-gym: A framework for constrained reinforcement learning. 2022.

Guan, J., Wu, J., Li, J.-N., Cheng, C., and Wu, W. A survey on personalized alignment–the missing piece for large language models in real-world applications. *Findings of the Association for Computational Linguistics*, 2025.

Guo, Y., Cui, G., Yuan, L., Ding, N., Sun, Z., Sun, B., Chen, H., Xie, R., Zhou, J., Lin, Y., Liu, Z., and Sun, M. Controllable preference optimization: Toward controllable multi-objective alignment. In *Empirical Methods in Natural Language Processing*, 2024.

Harland, H., Dazeley, R., Vamplew, P., Senaratne, H., Nakisa, B., and Cruz, F. Adaptive alignment: Dynamic preference adjustments via multi-objective reinforcement learning for pluralistic AI. In *Pluralistic Alignment Workshop at NeurIPS*, 2024.

Hayes, C. F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L. M., Dazeley, R., Heintz, F., et al. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 2022.

Huang, S., Abdolmaleki, A., Vezzani, G., Brakel, P., Mankowitz, D. J., Neunert, M., Bohez, S., Tassa, Y., Heess, N., Riedmiller, M., et al. A constrained multi-objective reinforcement learning framework. In *Conference on Robot Learning*, 2022.

Jang, J., Kim, S., Lin, B. Y., Wang, Y., Hessel, J., Zettlemoyer, L., Hajishirzi, H., Choi, Y., and Ammanabrolu, P. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*, 2024.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Chen, B., Sun, R., Wang, Y., and Yang, Y. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. *Advances in Neural Information Processing Systems*, 2023.

Kim, W., Lee, J., Lee, J., and Lee, B.-J. FairDICE: Fairness-driven offline multi-objective reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Li, W., Mustafa, W., Monteiro, M., Wang, P., Kloft, M., and Fellenz, S. Tora: Train once, realign anytime for offline multi-objective reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 37609–37617, 2026.

Lin, Q., Tang, B., Wu, Z., Yu, C., Mao, S., Xie, Q., Wang, X., and Wang, D. Safe offline reinforcement learning with real-time budget constraints. In *International Conference on Machine Learning*, 2023.

Lin, Q., Liu, Z., Mo, D., and Yu, C. An offline adaptation framework for constrained multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 2024a.

Lin, Q., Yu, C., Liu, Z., and Wu, Z. Policy-regularized offline multi-objective reinforcement learning. In *Autonomous Agents and Multi-Agent Systems*, 2024b.

Liu, Z., Guo, Z., Yao, Y.-F., Cen, Z., Yu, W., Zhang, T., and Zhao, D. Constrained decision transformer for offline safe reinforcement learning. In *International Conference on Machine Learning*, 2023.

Liu, Z., Guo, Z., Lin, H., Yao, Y., Zhu, J., Cen, Z., Hu, H., Yu, W., Zhang, T., Tan, J., and Zhao, D. Datasets and benchmarks for offline safe reinforcement learning. *Journal of Data-centric Machine Learning Research*, 2024.

Montori, V. M., Ruissen, M. M., Hargraves, I. G., Brito, J. P., and Kunneman, M. Shared decision-making as a method of care. *BMJ evidence-based medicine*, 28(4): 213–217, 2023.

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

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

Rame, A., Couairon, G., Dancette, C., Gaya, J.-B., Shukor, M., Soulier, L., and Cord, M. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 2023.

Roijers, D. M., Vamplew, P., Whiteson, S., and Dazeley, R. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 2013.

Shi, R., Chen, Y., Hu, Y., Liu, A., Hajishirzi, H., Smith, N. A., and Du, S. S. Decoding-time language model alignment with multiple objectives. *Advances in Neural Information Processing Systems*, 2024.

Sorensen, T., Jiang, L., Hwang, J. D., Levine, S., Pyatkin, V., West, P., Dziri, N., Lu, X., Rao, K., Bhagavatula, C., Sap, M., Tasioulas, J., and Choi, Y. Value kaleidoscope: engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i18.29970. URL https://doi.org/10.1609/aaai.v38i18.29970.

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

Stooke, A., Achiam, J., and Abbeel, P. Responsive safety in reinforcement learning by PID lagrangian methods. In *International Conference on Machine Learning*, 2020.

Taylor, J. Quantilizers: A safer alternative to maximizers for limited optimization. In *AAAI Workshop: AI, Ethics, and Society*, 2016.

Vamplew, P., Hayes, C. F., Foale, C., Dazeley, R., and Harland, H. Multi-objective reinforcement learning: A tool for pluralistic alignment. In *Pluralistic Alignment Workshop at NeurIPS 2024*, 2024.

Wachi, A., Tran, T., Sato, R., Tanabe, T., and Akimoto, Y. Stepwise alignment for constrained language model policy optimization. *Advances in Neural Information Processing Systems*, 2024.

Wang, H., Lin, Y., Xiong, W., Yang, R., Diao, S., Qiu, S., Zhao, H., and Zhang, T. Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. In *Proceedings of the Association for Computational Linguistics*, 2024a.

Wang, Z., Dong, Y., Zeng, J., Adams, V., Sreedhar, M. N., Egert, D., Delalleau, O., Scowcroft, J., Kant, N., Swope, A., et al. Helpsteer: Multi-attribute helpfulness dataset for steerlm. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024b.

Wu, R., Zhang, Y., Yang, Z., and Wang, Z. Offline constrained multi-objective reinforcement learning via pessimistic dual value iteration. In *Advances in Neural Information Processing Systems*, 2021.

Wu, Z., Hu, Y., Shi, W., Dziri, N., Suhr, A., Ammanabrolu, P., Smith, N. A., Ostendorf, M., and Hajishirzi, H. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processings Systems*, 2023.

Yang, K., Liu, Z., Xie, Q., Huang, J., Zhang, T., and Ananiadou, S. Metaaligner: Towards generalizable multi-objective alignment of language models. *Advances in Neural Information Processing Systems*, 2024a.

Yang, R., Pan, X., Luo, F., Qiu, S., Zhong, H., Yu, D., and Chen, J. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In *International Conference on Machine Learning*, 2024b.

Yao, Y., Cen, Z., Ding, W., Lin, H., Liu, S., Zhang, T., Yu, W., and Zhao, D. Oasis: Conditional distribution shaping for offline safe reinforcement learning. *Advances in Neural Information Processing Systems*, 2024.

Yuan, Y., Zheng, Z., Dong, Z., and Hao, J. Moduli: Unlocking preference generalization via diffusion models for offline multi-objective reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2024.

Zhong, Y., Ma, C., Zhang, X., Yang, Z., Chen, H., Zhang, Q., Qi, S., and Yang, Y. Panacea: Pareto alignment via preference adaptation for llms. *Advances in Neural Information Processing Systems*, 2024.

Zhou, Z., Liu, J., Shao, J., Yue, X., Yang, C., Ouyang, W., and Qiao, Y. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of ACL*, 2024.

Zhu, B., Dang, M., and Grover, A. Scaling pareto-efficient decision making via offline multi-objective RL. In *International Conference on Learning Representations*, 2023.