Research Statement

Inference-Time Pluralistic AI Alignment

Weichen Li · RPTU Kaiserslautern-Landau · weichen@cs.uni-kl.de

1 Vision: Train Once, Align at Inference Time

Human values are diverse: people often disagree about what makes a decision good, and in many situations, there is no single right answer. As AI systems increasingly support decisions in healthcare, education, public services, and everyday life, they must consider different goals, values, and preferences that vary across people and situations. These differences are not simply noise in human feedback; they often reflect valid but conflicting priorities. However, many AI alignment methods combine human preferences into a single reward or preference signal, which can overlook important differences and favour an "average" user [7, 8]. This problem becomes more challenging after deployment, when different users, changing situations, or new regulations require different decisions. Although some methods allow models to adjust their behaviour after training, adapting reliably to changing preferences and safety requirements remains difficult. Retraining a model for every new situation is also costly and impractical. My research aims to develop inference-time alignment methods that allow AI systems to adapt to diverse human values while respecting essential safety constraints.

One central question is: *How can AI systems make good decisions when people from different backgrounds disagree about what "good" means?* My approach is to separate learning from decision-making (Figure 1). During training, an AI agent learns about different objectives and safety constraints from offline data, without fixing how these objectives should be balanced. At inference time, the agent can adapt its decisions to the preferences of different users and the needs of specific situations, without retraining. However, not everything should be a matter of preference. Some objectives, such as effectiveness and cost, can be balanced according to individual priorities, while safety and ethical requirements must be respected.

My work is organised around three questions:

Q1 **Flexible trade-offs:** How can an agent adapt to new preferences after training, without retraining?

Q2 **Safety and ethical constraints:** How can safety and ethical constraints remain satisfied while preferences change?

Q3 **Human–AI decision-making:** How should human preferences guide AI decisions, and how can humans and AI make better decisions together? *(Future research)*

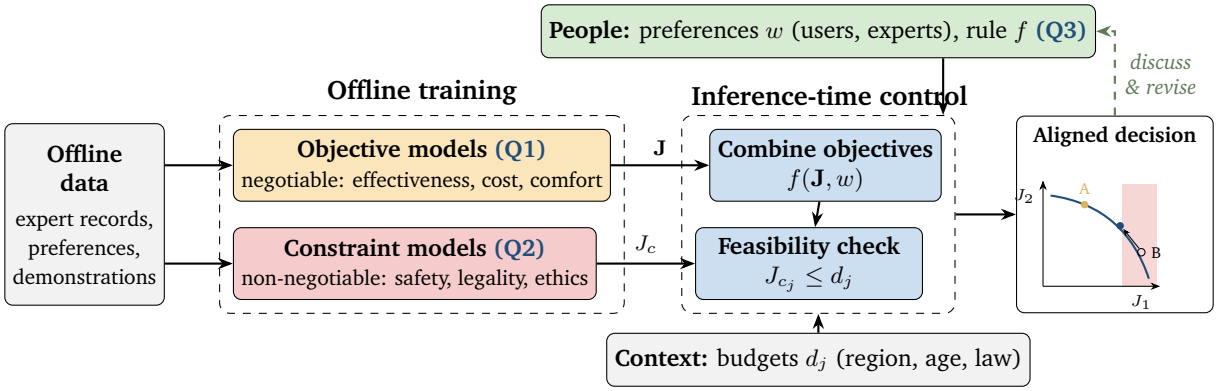


Figure 1: Overview of my research. Training produces reusable models of negotiable objectives (Q1; TORA, LGPO) and non-negotiable constraints (Q2; Ethics in Action). At inference time, people choose the trade-off and the deployment context sets the constraint budgets, without retraining. User A's preference is realised directly; user B's request violates a constraint (red region) and is moved to the nearest feasible decision. My future work (Q3, green) studies who chooses the trade-off and how humans and AI decide together.

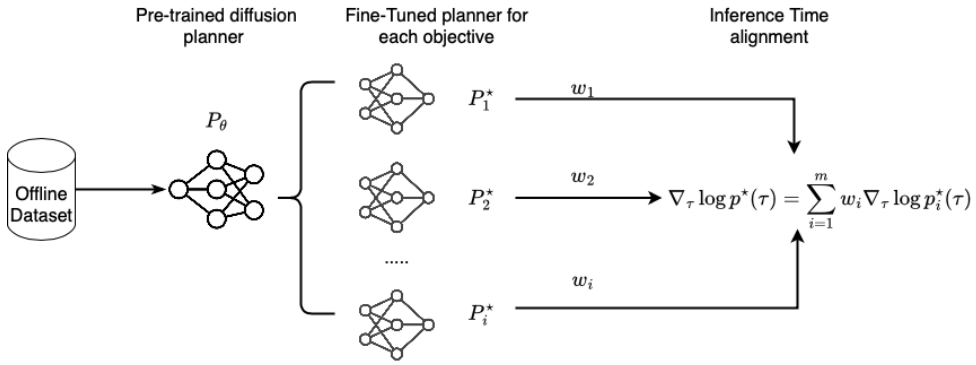


Figure 2: Overview of TORA framework: We first pre-train a diffusion planner model, and then fine-tune each objective P_i independently using Direct Preference Optimization (DPO),without requiring preference weights during training. At inference time, preference weights are provided for each objective to sample the planned trajectory τ. Since each objective is trained separately, new objectives can be added at any time.

2 Research Contributions

2.1 Flexible trade-offs after training (Q1)

In *multi-objective reinforcement learning* (MORL), an agent receives a vector of rewards $\mathbf{r} = (r_1, \ldots, r_m)$ instead of a single number, and a user expresses how much each objective matters through a preference vector $w = (w_1, \ldots, w_m)$. Because objectives conflict, improving one can worsen another, so there is a *set* of good policies rather than one. Existing offline MORL methods condition the policy on preferences *during training*; they are reliable only for preferences seen in the data, and adding a new preference or objective typically means retraining.

TORA: Train Once, Realign Anytime [2] (AAAI 2026) addresses this directly. We proposed TORA Framework which is trained *once* on an offline dataset and can then be *realigned* to new user preferences at deployment time, without retraining the policy. This makes preference changes cheap and turns the trade-off into something chosen at inference time rather than baked in at training time. The framework is shown in Figure 2 In **Landmark-Guided Policy Optimization** [3] (ICML 2026), we carried this multi-objective perspective over to language models, treating the selection of a language model as a decision that must balance several competing criteria.

2.2 Non-negotiable constraints and ethical values (Q2)

Flexibility alone is not enough: an agent must also respect constraints that no user preference can override. I first studied this in *language-based agents* that act through text, a setting that naturally contains morally relevant choices. After developing agents for text-based adventure games with maximum-entropy RL [6], **Ethics in Action** [4] (AISTATS 2024) formulated moral decision-making as a *constrained Markov decision process*: the agent maximises task reward while keeping a separate *moral cost* below a budget. We solved it with Lagrangian relaxation together with meta-gradient adaptation of the constraint threshold. This gave me hands-on experience with how reward and constraint objectives interact during learning.

I then moved these constraints to *inference time*. In [5] (WORDPLAY @ EMNLP 2025), large language models provide reward and ethical guidance to offline RL agents at inference time, so that value judgements expressed in natural language can steer an already-trained agent.

2.3 A unified framework for pluralistic alignment (Q1 + Q2)

These lines of work come together in my NeurIPS 2026 position paper [1], which argues that *inference-time multi-objective control should be a central goal of pluralistic AI alignment*. We show that offline MORL, offline safe RL, and multi-objective LLM alignment, usually studied separately, are instances of one optimisation problem:

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

where $\mathbf{J}(\pi)$ collects the m negotiable objectives, f combines them according to preferences w, each J_{c_j} is a non-negotiable constraint with budget d_j, and $\Pi_{\mathcal{D}}$ keeps the policy close to the offline data. Each field supplies one ingredient: MORL contributes trade-off representations, safe RL contributes constraint formalism, and LLM alignment shows that control must happen after training. Our position is that w, f, and d_j should all be set *at deployment*, by users and deployers respectively, so that inference-time alignment means *controlled adaptation within a safety-feasible region*, not unconstrained user control.

3 Future Research Agenda

Who selects the trade-off? Normative and philosophical guidance. Current multi-objective methods can identify a set of possible trade-offs along a Pareto front, but they do not tell us which one should be chosen. This is not only a technical question, but also a philosophical one. Whose preferences should matter, and how should conflicting values be balanced? These questions are especially important for moral decisions, where values may differ across individuals and cultures, but the consequences can extend beyond individual preferences. Addressing these challenges requires insights from social choice theory, ethics, and philosophy. I aim to study different ways of balancing competing objectives, including Nash social welfare and max–min fairness, and investigate what assumptions they make about fairness and whose interests they prioritize.

Human–AI collaborative decision-making. In many settings, the goal is not for AI to decide instead of people, but to help people make better-informed decisions. Achieving this requires understanding how humans make decisions: how they balance competing values, reason under uncertainty, and reflect on their own judgments. An important part of this is metacognition: the ability to reflect on and evaluate one's own thinking and decision-making. Drawing on insights from psychology, cognitive science, neuroscience, and the social sciences, I hope develop AI systems that support this reflective process and act as partners in human decision-making.

For such collaboration to be effective, AI systems must also communicate their reasoning, limitations, and uncertainty in ways that people can understand and use. Because the framework in Figure 1 is domain-general, this agenda can be pursued in healthcare, psychology, education, or any field where people with different values make decisions together with AI, in close collaboration with researchers from that field.

Summary. My long-term goal is AI that is *flexible* enough to reflect diverse human values and *constrained* enough to remain safe. I look forward to pursuing this agenda with collaborators in machine learning, the human sciences, and application domains such as healthcare and psychology.

References

[1] W. Li, M. Stojanović, D. Neider, M. Kloft, S. Fellenz. Pluralistic AI Alignment Requires Inference-Time Multi-Objective Control. *NeurIPS 2026 (Position Paper)*, to appear.

[2] W. Li, W. Mustafa, M. Monteiro, P. Wang, M. Kloft, S. Fellenz. TORA: Train Once, Realign Anytime for Offline Multi-Objective RL. *AAAI 2026*.

[3] M. Monteiro, W. Li, P. Wang, M. Kloft, S. Fellenz. Landmark-Guided Policy Optimization for Multi-Objective Language Model Selection. *ICML 2026*.

[4] W. Li, R. Devidze, W. Mustafa, S. Fellenz. Ethics in Action: Training RL Agents for Moral Decision-Making in Text-Based Adventure Games. *AISTATS 2024*.

[5] W. Li, W. Mustafa, R. Devidze, M. Kloft, S. Fellenz. Inference-Time Value Alignment in Offline RL: Leveraging LLMs for Reward and Ethical Guidance. *WORDPLAY Workshop @ EMNLP 2025*.

[6] W. Li, R. Devidze, S. Fellenz. Learning to Play Text-Based Adventure Games with Maximum Entropy RL. *ECML-PKDD 2023*.

[7] T. Sorensen et al. Position: A Roadmap to Pluralistic Alignment. *ICML 2024*.

[8] C. Garbacea. Position: Large Language Models Should Learn Personalized Rather Than Aggregated Human Preferences. *ICML 2026*.