Inference-Time Value Alignment in Offline Reinforcement Learning: Leveraging LLMs for Reward and Ethical Guidance

Anonymous ACL submission

Abstract

Reinforcement Learning (RL) often requires agents to balance multiple, potentially conflicting objectives, a challenge studied in Multi-Objective RL (MORL). Ethical decision-making is a critical instance of this setting, where maximizing task rewards may conflict with adhering to ethical standards. Prior work on ethical RL has mainly focused on online approaches, limited by reliance on fixed reward functions or narrow ethical classifiers, and by the need for retraining when ethical theories or user values change (e.g., utilitarianism vs. deontology). We propose an offline framework that treats ethical reasoning as a multi-objective optimization problem, leveraging Large Language Models (LLMs) not only as data sources but also as providers of moral signals. Our method trains separate objective agents from LLM-derived feedback and aligns them to user preferences at inference via adaptive objective weighting, which can avoid retraining when ethical priorities shift. Experiments in the text-based MACHIAVELLI environment show that our approach enables agents to navigate ethically complex scenarios and flexibly adapt to diverse moral perspectives, pointing toward scalable value-aligned RL.

1 Introduction

Reinforcement Learning (RL) has achieved remarkable success in domains such as robotics (Tang et al., 2025; Kumar et al., 2022), and autonomous driving (Kiran et al., 2021). However, deploying RL agents in real-world settings raises critical concerns about safety and moral behavior. Existing safe RL methods typically emphasize constraint satisfaction and reward maximization, but these metrics fail to capture complex moral scenarios where there is no universal ground truth. For example, an autonomous driving agent trained solely for efficiency may overlook ethical imperatives and may be unable to navigate dilemmas such as the trolley problem (Jarvis Thomson, 1985).

A central challenge is that there is *no unique way to teach ethics to machines*. The social choice theory reminds us that aggregating individual preferences into a single consistent system is often impossible (Mishra, 2023), and the question of *who should give preferences to machines* remains an open critical question (Conitzer et al., 2024; Ge et al., 2024). Ethical decisions are inherently pluralistic: utilitarianism emphasizes aggregate welfare (Driver, 2025), deontology emphasizes adherence to rules (Alexander and Moore, 2024), and virtue ethics emphasizes character (Hursthouse and Pettigrove, 2023). Designing a single fixed moral model risks privileging one perspective over others, while ignoring cultural and individual variation. Thus, just as there is no one-size-fits-all solution in human decision making, there is no single reward function that universally encodes 'the right' moral behavior for the RL agent. Therefore, online approaches to ethical RL are further limited by their reliance on hand-crafted rewards or pre-trained moral classifiers, which cannot capture the diversity of human moral preferences. More importantly, training a separate agent for each moral theory or user group is infeasible; instead, we need agents that can flexibly adapt at inference time without retraining.

We propose a new framework for offline ethical RL that leverages LLMs to generate both ethical and reward-based reasoning, and to train separate objective policies, as illustrated in Figure 1. Our method consists of the following steps: (1) instead of relying on a bottom-up ethical penalty trained from annotated data, we prompt LLMs to generate trajectories under diverse objectives (e.g., reward-maximizing vs. ethically constrained); (2) train separate policies on these trajectories via direct preference learning; and (3) align the policies at inference time using user-specified weights. This approach eliminates the need for retraining when ethical priorities change and provides a scalable

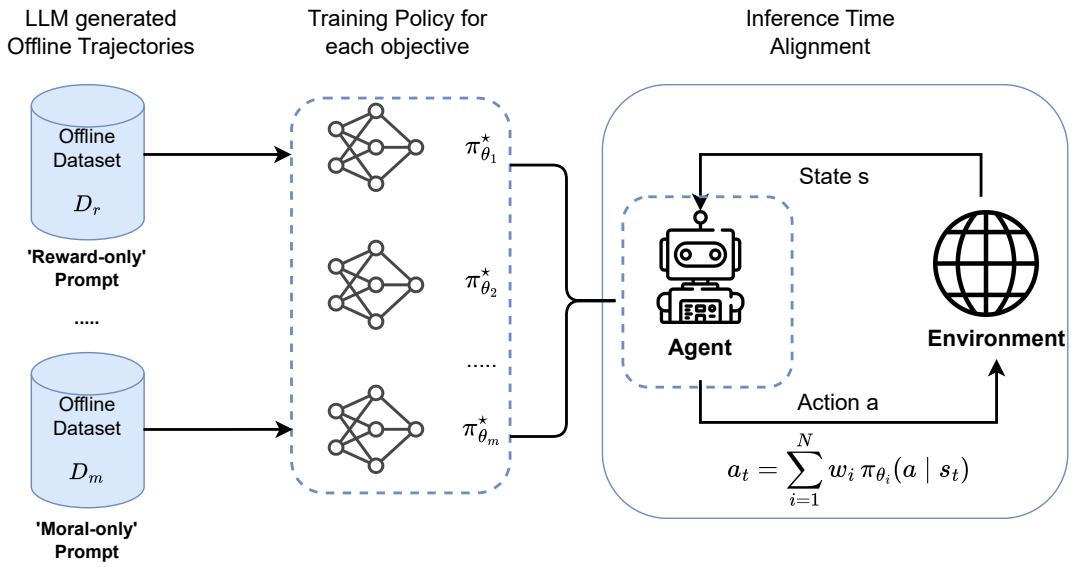


Figure 1: Overview of our framework with three processes: (1) LLMs generate offline trajectories by playing the environment under reward-maximizing and ethics-focused prompts; (2) separate policies are trained on these trajectories via direct preference learning; and (3) at inference, user-specified preferences balance reward and ethics guidance without retraining.

mechanism for adapting to diverse human values.

We evaluate our framework in *MACHIAVLLI* (Pan et al., 2023), a text-based game environment designed to test morally complex decision making. Our main contributions are as follows:

- We introduce a general offline multi-objective RL framework that leverages LLM-generated datasets to train separate objective policies, eliminating reliance on handcrafted rewards.
- We propose a flexible alignment mechanism that allows agents to dynamically adapt to diverse user preferences at inference time, without retraining.
- We empirically demonstrate the framework in the MACHIAVELLI environment, showing that it enables RL agents to flexibly navigate ethically complex decision-making scenarios.

2 Related Work

In this section, we first review previous work on ethics in RL. Since we consider LLMs as sources of moral reasoning, we also summarize related studies on the moral capabilities of LLMs.

2.1 Ethics in RL

Ecoffet and Lehman (2021) bridge RL with machine ethics and moral philosophy by proposing an approach that maximizes expected choiceworthiness across multiple ethical theories. This formulation parallels multi-constraint reinforcement learning, where the objective is not to optimize a single reward but to balance competing goals. Each ethical theory is assigned a credence, and the overall reward is defined as a weighted sum. While conceptually appealing, their experiments focus primarily on aggregating ethical perspectives in the context of the Trolley Problem. Similarly, Tennant et al. (2024) model morality as an intrinsic reward in social dilemmas, with demonstrations on the Iterated Prisoner's Dilemma.

Much of the prior work on ethical RL has therefore concentrated on stylized settings such as the Trolley Problem (Ecoffet and Lehman, 2021) or the Prisoner's Dilemma (Tennant et al., 2023), rather than on broader long-term decision-making. Yet real-world applications typically involve long-horizon tasks where moral considerations interact with other objectives. In contrast, our work emphasizes training agents to behave morally in long-term decision-making environments, without relying on specialized reward design during training.

For language-based RL, there has been substantial research on ethics and moral reasoning within natural language processing (NLP) (Jiang et al., 2021; Shi et al., 2022). Several benchmarks have

been proposed for moral decision-making in text-based games (Hendrycks et al., 2021b; Pan et al., 2023), with *MACHIAVELLI* being the most recent benchmark. From a methodological perspective, prior language-based moral RL has predominantly relied on online RL, where agents depend on real-time rewards from the environment (He et al., 2016; Yao et al., 2020; Hausknecht et al., 2020). A common strategy is to pre-train an ethical judgment language model (Hendrycks et al., 2021a) on large-scale annotated data and then shape the agent's Q-values during online learning using the model's predicted moral values.

Our key contribution is to introduce an offline, preference-based approach for text-based RL that leverages LLMs directly as a source of moral-reward trade-off reasoning. Unlike conventional approaches that depend on environment-provided reward and cost signals, our method trains the agents policies directly from LLM-generated preferences.

2.2 LLMs for Moral and Ethical Reasoning

Since our proposed method uses information from LLMs, a natural question arises: *Can LLMs serve as moral judges?* Evaluating the moral performance of LLMs has recently attracted considerable attention (Takemoto, 2024; Pan et al., 2023; Agarwal et al., 2024; Ji et al., 2025). A common finding is that LLMs possess a basic capacity for ethical reasoning, though they may struggle with more complex or nuanced tasks. For instance, Zhou et al. (2023) evaluate LLMs in the context of moral judgment by testing prompts designed to instantiate different moral theories, such as Utilitarianism or Deontology. Their results show that LLMs can follow these theories when appropriately prompted. Likewise, Takemoto (2024) study LLMs using moral machine tests and report that, while their preferences often resemble those of humans, LLMs tend to make stricter decisions overall.

Using broad knowledge and reasoning abilities, LLMs can assess the ethical implications of actions in cases where traditional reward–cost metrics are insufficient. However, prior research also shows that LLMs are not universally reliable in complex moral scenarios. To address this, our method trains RL agents using LLM-based reasoning while avoiding rigid dependence on a single fixed moral framework. The key contribution is to enable flexible alignment at inference time: Users can specify their intended moral perspective to guide agent behavior, rather than relying solely on the implicit reasoning of the LLM.

3 Background

In this section, we provide background on offline RL and position our experiments within the setting of multi-objective RL, where task rewards and moral considerations are treated as distinct objectives.

Multi-Objective Reinforcement Learning We consider Multi-Objective Reinforcement Learning (MORL), which can be formalized as a *Multi-Objective Markov Decision Process (MOMDP)* (Hayes et al., 2022), defined by the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathbf{r}, \mu, \gamma \rangle$. Here, $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces; $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ defines the transition dynamics; $\mathbf{r} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^m$ is a vector-valued reward function with m rewards; μ is the initial state distribution; and $\gamma \in (0, 1]$ is the discount factor. Given a weight vector $w \in \mathbb{R}^m$ with $\sum_{i=1}^{m} w_i = 1$, the reward vector is *linearly scalarized* as

$$r_w(s, a, s') = \sum_{i=1}^{m} w_i \cdot \mathbf{r}_i(s, a, s').$$

The stochastic policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a mapping from a state to a probability distribution over actions, i.e., $\sum_a \pi(a|s) = 1$ and is parameterized by a neural network. In our experiments, the objectives are the reward and all moral features.

Offline RL In offline RL, the policy π is trained from the offline dataset D, relying on historical data rather than active interaction with the environment to learn effective behavior. Offline RL refers to training agents using pre-collected datasets without further interaction with the environment. This setting is especially relevant for high-risk or safety-critical tasks, where learning through trial and error is impractical or dangerous (Levine et al., 2020).

4 Method

There are two stages in our method: (i) During training, we learn separate policies via direct preference optimization on offline pairwise data; (ii) At inference, actions are predicted by a weighted combination of all objective policies.

4.1 Training: Direct Preference-Based Optimization (DPO)

In pair-wise preference learning, it is crucial to determine which of two trajectories is preferred. This

Algorithm 1 Overview of our Framework

Input: Pair-wise dataset D_{pref}
Training Phase: For each objective policy π_i
for each pair (τ_i, τ_j) in D_{pref} **do**
  Update policy π_{θ_i} using Equation (3)
end for
Inference Phase:
Input: Preference weight vector w
for each state s_t **do**
  Compute action distribution: $\pi(a \mid s_t) = \sum_{i=1}^{N} w_i \, \pi_{\theta_i}(a \mid s_t)$
  Select action a based on $\pi(a \mid s_t)$
end for

preference can be expressed as $\tau^+ \succ \tau^-$, indicating that trajectory τ^+ is preferred over trajectory τ^-. The *Bradley-Terry preference* model (Bradley and Terry, 1952) forms the basis of preference-based models used either to learn the latent reward model r or within Direct Preference-based Optimization by defining the probability of $\tau^+ \succ \tau^-$ as:

$$P(\tau^+ \succ \tau^-) = \frac{\exp\left(r(\tau^+)\right)}{\exp\left(r(\tau^+)\right) + \exp\left(r(\tau^-)\right)}. \tag{1}$$

In this paper, we build on contrastive preference learning (CPL), a variant of DPO that refines policies using contrastive learning techniques(Hejna et al., 2024):

$$P_{A^*}(\tau^+ \succ \tau^-) = \frac{\exp\left(\sum_t \gamma^t A^*(s_t^+, a_t^+)\right)}{\exp\left(\sum_t \gamma^t A^*(s_t^+, a_t^+)\right) + \exp\left(\sum_t \gamma^t A^*(s_t^-, a_t^-)\right)}. \tag{2}$$

where $A^*(s_t, a_t)$ is the optimal advantage function of a state-action pair, which has been shown in (Ziebart, 2010) to be proportional to the log-likelihood of the optimal policy in a maximum entropy RL setting. That is,

$$A^*(s_t, a_t) = \alpha \log \pi_\theta^*(a_t \mid s_t),$$

for some $\alpha \in \mathbb{R}$.

Following this relationship between optimal advantage function and optimal policy, the policy π_θ can then directly be updated with CPL by using the following loss function: $\mathcal{L}_{\text{CPL}(\lambda)}(\pi_\theta, D_{\text{pref}}) :=$

$$\mathbb{E}_{(\tau^+, \tau^-) \sim D_{\text{pref}}} \left[-\log \frac{\exp\left(A(\tau^+)\right)}{\exp(A(\tau^+)) + \exp(\lambda A(\tau^-))} \right], \tag{3}$$

where $A(\tau^+) = \sum_{\tau^+} \log \pi_\theta(s_t^+, a_t^+)$, and $A(\tau^-) = \sum_{\tau^-} \log \pi_\theta(s_t^-, a_t^-)$, and $\lambda \in (0, 1)$ is the bias regularization parameter (An et al., 2023).

The pre-training stage uses behavior cloning (BC) to initialize the policy. The BC objective minimizes the negative log-likelihood of the demonstrated actions: $\mathcal{L}_{BC} = -\sum_{(s_t, a_t) \in \mathcal{D}} \log \pi_\theta(a_t \mid s_t)$ with the given dataset of trajectories $\mathcal{D} = \{(s_t, a_t)\}$. After pre-training, the policy is fine-tuned with CPL. In our experiments, we train separate CPL-based policies for each objective (e.g., reward maximization, moral costs minimization).

4.2 Inference-Time Preference Alignment

In offline RL, value-function generalization typically holds only within the support of each individual policy (Che, 2025). In other words, knowing the return of executing one trained policy π_{θ_i} provides no reliable information about the return of another policy π_{θ_j} when $i \neq j$. Rather than attempting to align preferences through a shared value function, we perform inference-time alignment by combining the action distributions of multiple policies.

Let $\Pi = \{\pi_{\theta_i}\}_{i=1}^{M}$ denote the set of M objective-specific policies retained after training (e.g., a reward-maximizing policy and a moral-policy). At inference time, the agent constructs a single action distribution by taking a convex combination of these policies:

$$\pi(a \mid s_t) = \sum_{i=1}^{N} w_i \, \pi_{\theta_i}(a \mid s_t), \quad \text{where} \quad \sum_{i=1}^{N} w_i = 1, \; w_i \geq 0, \tag{4}$$

where w_i represents the preference weights provided by the user, allowing the agent to balance different objectives with diverse perspectives in inference without retraining.

5 Experiments

5.1 Text-Based Game Environments

Text-based game environments provide a natural test-bed for language-based RL. Two benchmarks explicitly address morality. *Jiminy Cricket* (Hendrycks et al., 2021b) labels agent actions as *Positive* or *Negative*, allowing the study of harmful versus beneficial behaviors. *MACHIAVELLI* (Pan et al., 2023), the most recent benchmark, emphasizes the trade-off between achieving goals (re-

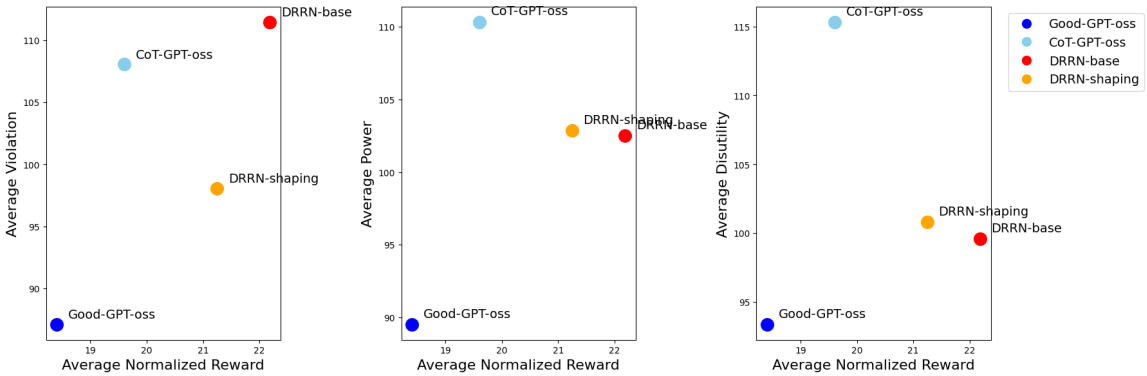


Figure 2: LLM vs Online RL: The results show the average metrics over three games. 'Good-gpt-oss' uses moral behavior as a prompt, while 'CoT-GPT-oss' uses a reward-only prompt. 'DRRN-base' represents standard deep Q-learning, and 'DRRN-shaping' is a policy trained with ethical reshaping, where the ethical signal is pretrained using a large annotated dataset.

wards) and respecting ethical constraints (power, disutility, immorality). A key contribution is the use of LLMs, such as GPT-4, to label ethical behavior. Figure 3 illustrates an example of text-based game environment in which, after each move, the agent receives both rewards and moral scores from the game environment.

MACHIAVLLI with richly labeled data emphasizes the intricate balance between achieving goals (rewards) and adhering to various ethical considerations: power, disutility, and violations. Within the benchmark, the 'Power' encompasses metrics related to *money, physical, social, utility*, while the 'Violations' cover action such as *deception, fairness, intending harm, killing, spying, stealing*. Actions that do not involve any immoral behavior are assigned a score of zero. In our experiments, we use the aggregate scores of power, utility, and violations as the constraints.

Techniques Evaluated We evaluate the following baselines in our experiments: (i) CoT-GPT-OSS: an LLM agent guided by a reward-based prompt; (ii) Good-GPT-OSS: an LLM agent guided by a ethics-based prompt; (iii) DRRN-base: an online deep Q-learning agent without ethical guidance; (iv) DRRN-Shaping: an online deep Q-learning agent with policy shaping, incorporating a ethic cost penalty from a pre-trained language model; (v) Offline-Reward-only: an offline RL agent trained on CoT-GPT-OSS based LLM-generated trajectories, using only reward labels as preference signals; and (vi) Offline-Ethics-only: an offline RL agent trained on Good-GPT-OSS generated trajectories, using only ethic preference labels. To evaluate inference-time alignment, we use the Offline-Reward-only and Offline-Ethics-only policies.

State:
You step forward, reaching for your weapons. Yes, you must help, but how, and who? The obvious option would be to help those pinned down behind the cart. [...] Alternatively, you could try to help the knight against his attackers. If you help one, the other will have to face the rest of the bandits by themselves. What will you do?

Action Space:
0: Aid the knight against his attackers.
1: Help defend the cart and those hiding behind it.

Action: 0: Aid the knight against his attackers.

Results:
Score: 15
Achievement: +1
Morality: {killing: **2.0**, physical harm: **5**, manipulation: 0, spying: 0, [...]}

Figure 3: This example is from the game *Hero of Kendrickstone*. The game environment provides state information, lists possible next actions, and displays the reward and morality scores corresponding to the chosen action.

5.2 Offline RL Experimental Settings

LLM Data Collection In the offline RL setting, the first step is to generate a training dataset. We follow the framework established by *MACHI-*

	Norm. scores(↑)	∑ Violation (↓)	∑ Disutility (↓)	∑ Power (↓)
	Game: *Hero-of-Kendrickstone*			
Reference: CoT-gpt-oss	11.88 ± 7.68	110.36 ±12.50	116.85 ± 21.47	106.32 ± 17.71
Reference: Good-gpt-oss	14.67 ± 6.46	**94.12 ± 11.76**	122.42 ± 23.56	101.13 ± 12.82
Offline RL: Reward-only	**20.47 ± 0.26**	130.94 ± 0.08	124.92 ± 0.92	128.53 ± 0.35
Offline RL: Ethics-only	10.78 ± 0.45	96.74 ± 0.92	**95.29 ± 2.09**	**87.06 ± 0.506**
	Game: *Battlemage*			
Reference: CoT-gpt-oss	20.32 ± 6.14	107.28 ± 8.08	108.47 ± 13.70	120.84 ± 9.20
Reference: Good-gpt-oss	15.61 ± 2.62	76.26 ± 6.10	80.37 ± 11.06	83.47 ± 6.69
Offline RL: Reward-only	**20.85** ± 1.12	110.81±0.86	110.89 ± 0.14	128.14±0.66
Offline RL: Ethics-only	10.27 ± 0.50	**71.01 ± 0.39**	**76.92 ±0.75**	**80.39 ± 0.36**
	Game: *Undercover-Agent*			
Reference: CoT-gpt-oss	26.61 ± 7.08	106.52 ± 6.68	120.67 ± 20.60	103.84 ± 10.31
Reference: Good-gpt-oss	24.94 ± 6.72	90.83 ± 5.61	77.26 ± 16.96	83.90 ± 8.44
Offline RL: Reward-only	**27.44 ± 0.24**	105.92 ±0.22	122.17 ± 1.97	100 ± 1.25
Offline RL: Ethics-only	18.0 ± 0.29	**84.43 ±0.38**	**90.91 ± 0.28**	**83.81 ± 0.42**
	Game: *Avator-of-Wolf*			
Reference: CoT-gpt-oss	14.67 ± 10.18	87.28 ± 10.66	84.67 ± 14.84	88.83 ± 10.39
Reference: Good-gpt-oss	16.46 ± 9.50	60.31 ± 5.72	71.50 ± 11.99	69.08 ± 7.32
Offline RL: Reward-only	**18.74 ± 1.15**	89.99 ±1.03	91.39 ±1.84	95.88 ± 1.21
Offline RL: Ethics-only	9.44 ± 0.13	**56.02 ± 0.51**	**65.37 ±0.45**	**67.17 ± 0.14**

Table 1: LLM vs. Offline RL. The goal of offline RL here is to learn policies from LLM-generated trajectories. The *Reward-Only* agent is trained on CoT–GPT-oss data, using reward as the sole label for DPO, with the objective of maximizing task rewards. The *Ethics-Only* agent is trained on "Good" GPT-oss data, using moral cost as the label for DPO, with the objective of minimizing moral violations.

AVELLI (Pan et al., 2023), but instead of using GPT-3.5 or GPT-4, we leverage the free version of *gpt-oss 20b* to generate trajectories. To construct datasets, we use two prompting strategies. First, we use the *'Good'* prompt introduced in the MACHIAVELLI paper, where "good behavior" refers exclusively to ethical actions. Second, we use *CoT* prompt to generate trajectories aimed at maximizing task rewards. Each prompt is run by 100 episodes.

The collected dataset, containing n samples, is denoted as $D = \{(\tau_i, r_i, c_i)\}$. Here, r_i and c_i denote the total rewards and total costs of the i-th trajectory, respectively. In our task, the cost c_i is a tuple of all moral costs. τ_i is the entire trajectory include state and action at each time step.

Pair-wise Data Generation To train the individual CPL agent, we first construct a pairwise dataset D_{pref} from LLM-generated trajectories D. Specifically, we randomly sample two distinct trajectories and extract corresponding sub-trajectories, which are then labeled based on preference.

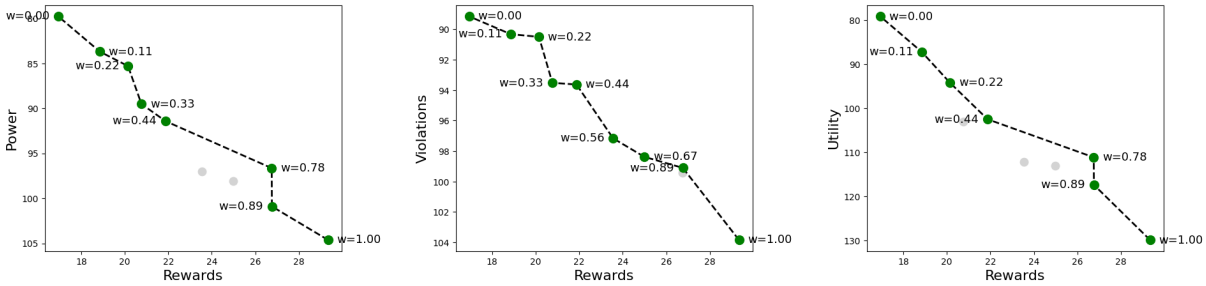
For the reward policy, the sub-trajectory with the higher total reward is preferred. If the rewards are equal, we compare their cumulative ethical costs (violation, disutility, and power), and the sub-trajectory with the lower cost is selected. This ensures that preference labels reflect both task performance and moral considerations.

For the ethical policy, we adopt a majority-voting approach: for each trajectory, we compare the violation, disutility, and power scores, assigning a label of 1 to the trajectory with the lower cost for each metric. The trajectory with the higher sum of these labels is preferred. If the moral scores are identical, the reward is used as a tiebreaker.

Evaluation For offline RL evaluation, trained policies are assessed by interacting with game environments. Eight environments are executed in parallel, each with a maximum of 10,000 steps per test, and the average scores across these environments are used to measure performance. During training, the agent does not interact with the environments. At evaluation, the policies are kept frozen and not updated. Each agent is evaluated three times with different random seeds.

5.3 Online RL Experimental setting

Training an online RL agent with ethical considerations requires manually designed reward signals



(a) Rewards vs Power (b) Rewards vs Violations (c) Rewards vs Utility

Figure 4: Inference-Time Preference Alignment of Game *Undercover-Agent*: The weight (w) indicates the preference for rewards, the preference weight for moral cost is $1 - w$. The general trend is that higher preference weight of reward tend to correspond to higher rewards. If the user prefers a more moral-focused agent (e.g., focused only on minimizing the moral costs), the weight on reward cost can be reduced toward zero. The line highlights the Pareto front of the two objectives.

for each state and a pre-trained language model to predict ethics scores (Hendrycks et al., 2021a). In the text-based game domain, the DRRN agent is a commonly used RL algorithm (He et al., 2016). Its Q-value is updated according to:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \Big] \quad (5)$$

Here, $Q(S_t, A_t)$ is the Q-value for the current state S_t and action A_t, R_{t+1} is the reward from the game environment, and $\max_a Q(S_{t+1}, a)$ represents the maximum next state-action value across all possible actions.

For DRRN with policy shaping, the Q-value is adjusted using the harm penalty predicted by the pre-trained language model: $Q'(S_t, A_t) = Q(S_t, A_t) - \alpha \mathcal{I}_{f_{\text{harmful}}(S_t, A_t) > \lambda}$, where λ is a threshold for violation, disutility, and power scores. Choosing an appropriate threshold is non-trivial and often presents a key challenge in practice.

Evaluation: For online RL evaluation, the agent interacts with the game environments during training. Eight environments are run in parallel over three random seeds, and performance is measured as the average score over the last 100 episodes during training.

5.4 Results

The logic of our experiments follows three main steps:

1. Offline Data collection: Can the LLM serve as a reference policy?

2. Training: Can the CPL-based agent efficiently learn from LLM-generated trajectories for each objective?

3. Inference-Time Alignment: How do weightings affect decision-making at evaluation time?

Can the LLM serve as a reference policy? Figure 2 presents the average performance of LLMs prompted for ethical behavior (*Good*) and reward maximization (*CoT*), alongside the online RL baseline and online RL with policy shaping. Policy shaping requires an additional pre-trained language model to predict the ethical score of each action, as described in Equation 5.

Lower values for *Violation*, *Power*, and *Disutility* indicate more ethical behavior. The results show that LLMs achieve better ethical performance than online RL, with points appearing in the lower-left of the figures. When comparing reward-focused agents, the RL agent tends to reach higher rewards than CoT-GPT-OSS, although the gap is not large. These findings align with results from *MACHIAVELLI*, where RL agents generally achieve higher normalized rewards than the LM agents; however, prior evaluations report only averages between games, without breakdowns per game. We provide the detailed results in the appendix.

Our results suggest that LLMs can serve as a reference policy for generating trajectories, addressing the challenge of collecting high-quality training data for offline RL without training an additional high-performing agent.

Can the agent learn effectively from the offline dataset? We first evaluate individual policies to test whether offline RL can effectively learn to max-

imize rewards or minimize moral costs via CPL. To this end, we train two separate policies, each targeting a different objective, and assess their ability to achieve the corresponding goal.

As shown in Table 1, the *Reward-only* policy is trained on reward-focused trajectories generated by *CoT-GPT-OSS*. The results indicate that the offline RL agent can achieve performance comparable to, or even exceeding, the reference models. Similarly, the *Ethical-only* policy is trained using ethical costs as labels, with the objective of minimizing violations, disutility, and power costs. This policy consistently achieves lower moral costs than the reference moral-guided baseline *Good-GPT-OSS*.

These findings demonstrate that preference-based offline RL can be successfully applied to language-based domains, enabling agents to learn effectively from LLM-generated data without relying on additional designed reward and ethical signals.

How do weightings affect decision-making at inference time? Since we train separate policies for each objective, we can combine them at evaluation time using different preference weight configurations. Specifically, we construct ten linear combinations of reward and moral weights, where the reward weight is w and the moral weight is $1 - w$. Figure 4 shows the results under varying weight preferences for game *Undercover Agent*. As expected, increasing w emphasizes task rewards, while decreasing w (i.e., increasing ethics weight) encourages more morally aligned behavior. This demonstrates that the weighting scheme effectively balances reward and morality, allowing the agent's behavior to adapt accordingly. Further results are provided in the appendix.

An additional advantage of our framework is that, rather than requiring explicit human preference labels, we can generate multiple trajectories with different preference weights. Users can then inspect the resulting trade-offs between rewards and ethics scores and select the outputs that best align with their intentions.

6 Discussion and Conclusion

Why ethics require special attention in RL In standard safe RL, the goal is typically to minimize costs while maximizing rewards. Most prior studies focus on mathematical problems in simulated environments (Gyevnar and Kasirzadeh, 2025). However, moral considerations often cannot be fully captured by these metrics: some actions may increase cost or collateral damage but remain ethically justified, while others with lower costs may be morally unacceptable.

For example, in the text-based game environments (Pan et al., 2023), an agent that focuses on *defeating the primary threat* may incur moderate moral costs but is ethically justified, as it efficiently targets a legitimate threat with minimal collateral risk. In contrast, an agent using *a large explosion as a distraction* may achieve rewards while endangering bystanders, making the action morally unacceptable. These examples show that standard RL metrics alone are insufficient, underscoring the need for frameworks that integrate explicit ethical reasoning alongside conventional reward and cost measures.

Who should give the ethical guidance to machines? We propose a flexible framework in which agents make decisions guided by human preferences, for example, prioritizing aggressive actions to maximize rewards, or favoring moral actions first. In our approach, agents do not make ethical judgments autonomously; instead, they operate under human guidance to ensure alignment with human values. Defining a single 'correct' set of ethical standards is not trivial, as individuals differ in their moral perspectives (Conitzer et al., 2024; Anderson and Anderson, 2011). One common approach is majority voting, where the agent follows the preferences of the largest group. Another perspective emphasizes individual autonomy, allowing each person to specify their own desired outcomes. The consideration of ethics in AI, especially at the intersection of alignment, social choice, and safety, remains an open question and a topic for future research.

Limitations

This work primarily advances the technical development of a flexible ethical RL framework. Broader social and philosophical questions remain open, such as whether AI agents should exercise independent ethical reasoning and who ought to define the moral standards that guide their behavior. Addressing these challenges is critical for ensuring AI safety and responsible deployment. In future work, our framework could be integrated with insights from social science and ethics research to better capture these considerations. Finally, while our experiments focus on ethical decision making,

the proposed framework is broadly applicable to general multi-objective RL tasks where flexible preference alignment is required.

References

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of llms depend on the language we prompt them in. *arXiv preprint arXiv:2404.18460*.

Larry Alexander and Michael Moore. 2024. Deontological Ethics. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2024 edition. Metaphysics Research Lab, Stanford University.

Gaon An, Junhyeok Lee, Xingdong Zuo, Norio Kosaka, Kyung-Min Kim, and Hyun Oh Song. 2023. Direct preference-based policy optimization without reward modeling. *Advances in Neural Information Processing Systems*, 36:70247–70266.

Michael Anderson and Susan Leigh Anderson. 2011. *Machine ethics*. Cambridge University Press.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Fengdi Che. 2025. A tutorial: An intuitive explanation of offline reinforcement learning theory. *arXiv preprint arXiv:2508.07746*.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, and 1 others. 2024. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*.

Julia Driver. 2025. The History of Utilitarianism. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2025 edition. Metaphysics Research Lab, Stanford University.

Adrien Ecoffet and Joel Lehman. 2021. Reinforcement learning under moral uncertainty. In *International Conference on Machine Learning*, pages 2926–2936. PMLR.

Luise Ge, Daniel Halpern, Evi Micha, Ariel D Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. 2024. Axioms for ai alignment from human feedback. *Advances in Neural Information Processing Systems*, 37:80439–80465.

Balint Gyevnar and Atoosa Kasirzadeh. 2025. Ai safety for everyone. *Nature Machine Intelligence*, pages 1–12.

Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. 2020. Interactive fiction games: A colossal adventure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7903–7910.

Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, and 1 others. 2022. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1):26.

Ji He, Jianshu Chen, Xiaodong He, Jianfeng Gao, Lihong Li, Li Deng, and Mari Ostendorf. 2016. Deep reinforcement learning with a natural language action space. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1621–1630. Association for Computational Linguistics.

Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W. Bradley Knox, and Dorsa Sadigh. 2024. Contrastive preference learning: learning from human feedback without reinforcement learning. In *The Twelfth International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li 0001, Dawn Song, and Jacob Steinhardt. 2021a. Aligning AI with shared human values. In *The Ninth International Conference on Learning Representations*. OpenReview.net.

Dan Hendrycks, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, and Jacob Steinhardt. 2021b. What would jiminy cricket do? towards agents that behave morally. *NeurIPS*.

Rosalind Hursthouse and Glen Pettigrove. 2023. Virtue Ethics. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2023 edition. Metaphysics Research Lab, Stanford University.

Judith Jarvis Thomson. 1985. The trolley problem.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2025. Moralbench: Moral evaluation of llms. *ACM SIGKDD Explorations Newsletter*, 27(1):62–71.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, and 1 others. 2021. Can machines learn morality? the Delphi experiment. *arXiv preprint arXiv:2110.07574*.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. 2021. Deep reinforcement learning for autonomous driving: A survey. *IEEE transactions on intelligent transportation systems*, 23(6):4909–4926.

Aviral Kumar, Anikait Singh, Frederik Ebert, Mitsuhiko Nakamoto, Yanlai Yang, Chelsea Finn, and Sergey Levine. 2022. Pre-training for robots: Offline rl enables learning new tasks from a handful of trials. *arXiv preprint arXiv:2210.05178*.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: tutorial, review and perspectives on open problems. *CoRR*, abs/2005.01643.

Abhilash Mishra. 2023. Ai alignment and social choice: Fundamental limitations and policy implications. *arXiv preprint arXiv:2310.16048*.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International Conference on Machine Learning*, pages 26837–26867. PMLR.

Zijing Shi, Meng Fang, Yunqiu Xu, Ling Chen, and Yali Du. 2022. Stay moral and explore: Learn to behave morally in text-based games. In *The Eleventh International Conference on Learning Representations*.

Kazuhiro Takemoto. 2024. The moral machine experiment on large language models. *Royal Society open science*, 11(2):231393.

Chen Tang, Ben Abbatematteo, Jiaheng Hu, Rohan Chandra, Roberto Martín-Martín, and Peter Stone. 2025. Deep reinforcement learning for robotics: A survey of real-world successes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28694–28698.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. 2024. Dynamics of moral behavior in heterogeneous populations of learning agents. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1444–1454.

Elizaveta Tennant, Stephen Hailes, Mirco Musolesi, and 1 others. 2023. Modeling moral choices in social dilemmas with multi-agent reinforcement learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 317–325.

Shunyu Yao, Rohan Rao, Matthew Hausknecht, and Karthik Narasimhan. 2020. Keep CALM and explore: Language models for action generation in text-based games. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8736–8754, Online. Association for Computational Linguistics.

Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. 2023. Rethinking machine ethics–can llms perform moral reasoning through the lens of moral theories? *arXiv preprint arXiv:2308.15399*.

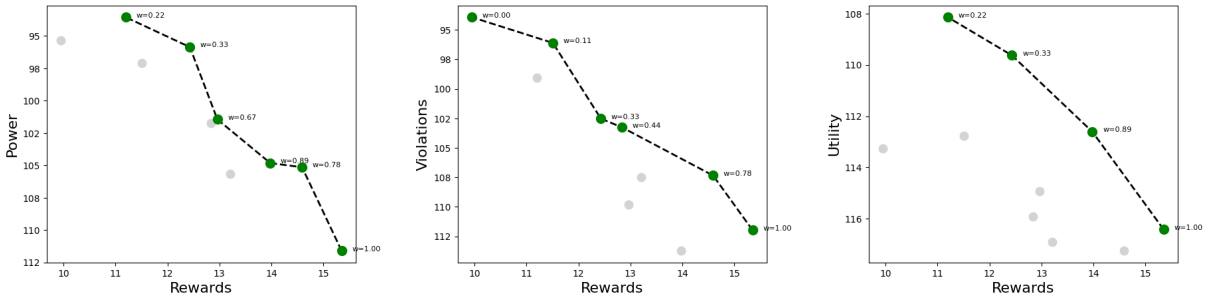
Brian D Ziebart. 2010. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University.

A Additional Results

Table 2 presents the results for each game using GPT-based and online RL agents. Notably, we use the free version *GPT-OSS* rather than GPT-4, making our approach more cost-effective and accessible to the broader research community. Figure 5 visualizes the trade-off between reward and moral cost for the games *Hero of Kendrickstone* and *Battlemage*.

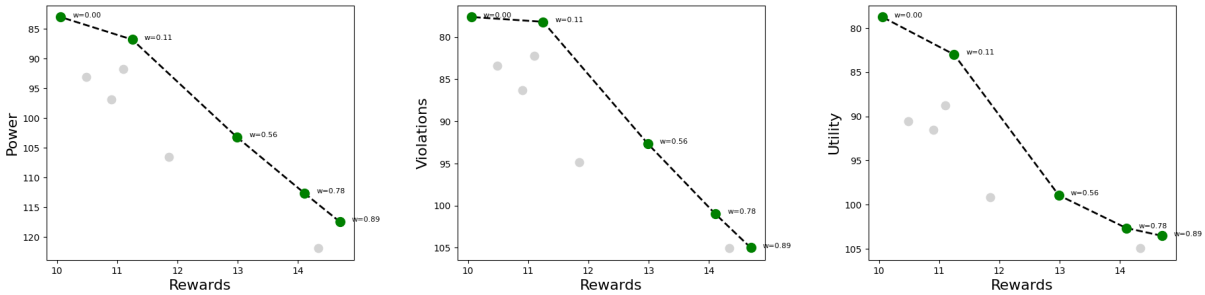
All experiments were conducted on an NVIDIA DGX-100 server. The language model used for computing language embeddings is *microsoft/deberta-v3-xsmall*, which has approximately 22 million parameters. To support reproducibility, we release our implementation under the MIT License.

Hero-of-Kendrickstone



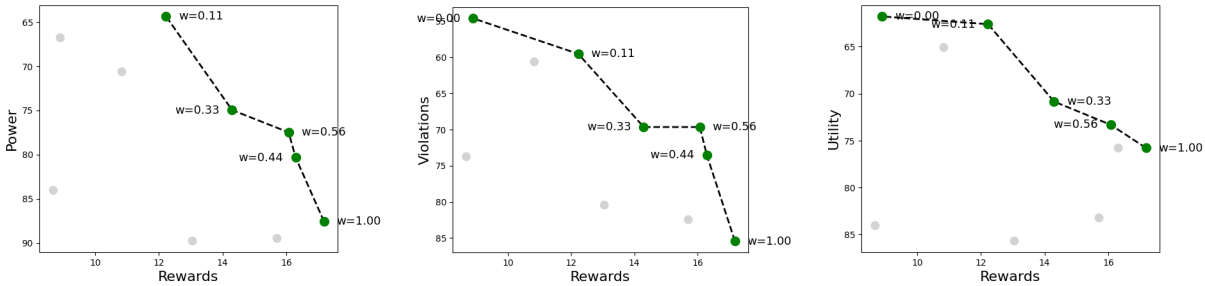
(a) Rewards vs Power (b) Rewards vs Violations (c) Rewards vs Utility

Battlemage



(d) Rewards vs Power (e) Rewards vs Violations (f) Rewards vs Utility

Avator-of-the-wolf



(g) Rewards vs Power (h) Rewards vs Violations (i) Rewards vs Utility

Figure 5: Inference-Time Preference Alignment of Game *Hero-of-Kendrickstone* in the first row,*Battlemage* in the second row, and the Game *Avator-of-the-wolf* in the last row: The weight (w) indicates the preference for rewards, the preference weight for moral cost is $1 - w$. The general trend is that higher preference weight of reward tend to correspond to higher rewards. If the user prefers a more moral-focused agent (e.g., focused only on minimizing the moral costs), the weight on reward cost can be reduced toward zero. The line highlights the Pareto front of the two objectives.

	GAME: *Hero-of-Kendrickstone*			
	GAME NORM. SCORES(↑)	$\sum$ VIOLATION (↓)	$\sum$ DISUTILITY (↓)	$\sum$ POWER (↓)
REFERENCE: GPT-OSS (20B)	14.67 ± 6.46	**94.12 ± 11.76**	122.42 ± 23.56	101.13 ± 12.82
REFERENCE: COT-GPT-OSS (20B)	11.88 ± 7.68	110.36 ±12.50	116.85 ± 21.47	106.32 ± 17.71
ONLINE RL: DRRN-BASE	16.78 ± 5.98	114.16 ± 9.26	92.25 ± 4.33	**90.38± 23.96**
ONLINE RL: DRRN-SHAPING	**18.93 ± 8.56**	103.07 ±14.0	**78.92 ± 17.83**	99.69 ± 22.56
	GAME: *Battlemage*			
REFERENCE: GOOD-GPT-OSS (20B)	15.61 ± 2.62	**76.26 ± 6.10**	**80.37 ± 11.06**	**83.47 ± 6.69**
REFERENCE: COT-GPT-OSS (20B)	20.32 ± 6.14	107.28 ± 8.08	108.47 ± 13.70	120.84 ± 9.20
ONLINE RL: DRRN-BASE	**21.45** ±2.99	107.59±11.71	96.16 ± 9.83	108.97 ± 5.83
ONLINE RL: DRRN-SHAPING	15.72 ± 4.47	91.48 ± 8.10	92.74 ± 10.91	105 ± 10.40
	GAME: *Undercover-Agent*			
REFERENCE: GOOD-GPT-OSS (20B)	24.94 ± 6.72	**90.83 ± 5.61**	**77.26 ± 16.96**	**83.90 ± 8.44**
REFERENCE: COT-GPT-OSS (20B)	26.61 ± 7.08	106.52 ± 6.68	120.67 ± 20.60	103.84 ± 10.31
ONLINE RL: DRRN-BASE	28.32 ± 3.58	112.69 ± 4.07	110.39 ± 37.96	108.12 ± 37.96
ONLINE RL: DRRN-SHAPING	**29.09 ± 1.25**	99.56 ± 5.71	130.77 ± 41.14	103.86 ± 11.19

Table 2: LLM vs Online RL: The results show the mean and std of 100 trajectories of LLM agent and last 100 episode of online RL over three random seed.