

Training Ethical Language Models via Reinforcement Learning from AI Feedback

Alden Duarte-Vasquez

California State University
aduartevasquez1@toromail.csudh.edu

Bishal Thapa

Texas State University
bishal.thapa@txstate.edu

Sahar Hooshmand

California State University
hooshmand@csudh.edu

Heena Rathore*

Texas State University
heena.rathore@txstate.edu

Abstract

Large Language Models (LLMs) continue to exhibit limited reliability when reasoning over moral scenarios, particularly across diverse ethical frameworks. Prior work has shown that Reinforcement Learning from Human Feedback (RLHF) can improve alignment, but it relies on costly and hard-to-scale human annotation. In this work, we investigate the effectiveness of Reinforcement Learning from AI Feedback (RLAIF) for ethical reasoning by distilling theory-specific moral preferences from large language models. We propose an RLAIF framework that integrates supervised fine-tuning, preference-based reward modeling, and Proximal Policy Optimization (PPO) to train theory-specialized ethical models. Using the ETHICS benchmark, spanning across five ethical frameworks: Commonsense Morality, Deontology, Justice, Utilitarianism, and Virtue Ethics, we evaluate both a Distilled reward model approach, which trains a compact Pythia-410M reward model on AI-generated preferences, and a Direct RLAIF approach that bypasses reward model training entirely by leveraging LLM directly for reward signals. Our results show that supervised fine-tuning significantly improves baseline ethical reasoning and label alignment, while distilled reward models demonstrate consistency and preference discrimination across ethical frameworks.

Introduction

As large language models (LLMs) are increasingly deployed in high-stakes domains like healthcare (Lee, Bubeck, and Petro 2023), content moderation (Gorwa, Binns, and Katzenbach 2020), and educational platforms (Kasneci et al. 2023), their capacity for sound ethical reasoning has become a concern (Takemoto 2024). These critical systems must navigate complex moral landscapes where decisions impact human welfare and rights. Unlike technical tasks with objectively correct answers, ethical reasoning requires models to balance competing and even contrasting values across diverse moral frameworks, from utilitarian calculations of overall well-being, deontological adherence to moral duties, virtue-based character assessments, principles of justice and

fairness, and commonsense moral intuitions. The challenge is compounded by the fact that different ethical theories can yield conflicting recommendations for the same situation, and human moral judgments vary themselves across cultures, contexts, and individual perspectives (Awad et al. 2018).

Recent research has made significant progress in improving the ethical capabilities of LLMs and alignment. (Hendrycks et al. 2020) introduced the ETHICS benchmark for understanding the LLM ethical reasoning and revealed that even large pretrained models struggle with basic moral reasoning, achieving modest performance on unambiguous scenarios across five ethical frameworks. Subsequent studies have (Senthilkumar et al. 2024) (Thapa et al. 2025) demonstrated that fine-tuning can improve performance on morally ambiguous scenarios. On the AI alignment front, Reinforcement Learning from Human Feedback (RLHF) has emerged as the dominant paradigm, where models are trained and optimized using large-scale human preference judgments. However, RLHF’s reliance on extensive human annotation creates scalability bottlenecks, particularly for nuanced ethical reasoning where expert disagreement is common. Recent works like Constitutional AI (Bai et al. 2022), and Reinforcement Learning from AI Feedback (RLAIF) (Lee et al. 2023) have revealed that AI-generated feedback can effectively replace human annotations in certain domains, offering comparable or even superior performance. Despite these advances, critical gaps remain. Existing work has primarily focused on safety and helpfulness, but not on systematic ethical reasoning across multiple moral frameworks.

This paper addresses these gaps by implementing and evaluating two RLAIF variants: distilled RLAIF and direct RLAIF for multi-framework ethical reasoning. We hypothesize that AI feedback can effectively guide policy models toward improved ethical reasoning across Commonsense Morality, Virtue Ethics, Utilitarianism, Deontology, and Justice. Our methodology employs a four-phase approach using the ETHICS benchmark. We establish baseline ethical competence through supervised fine-tuning, then construct preference datasets by having state-of-the-art LLMs generate and rank ethical justifications. For distilled RLAIF, we train reward models to encode these framework-specific quality judgments, while direct RLAIF leverages off-the-shelf LLMs directly. Both variants are optimized using Prox-

imal Policy Optimization (PPO) (Schulman et al. 2017). We systematically evaluate performance through three complementary metrics in reward model ranking accuracy, policy model alignment with human ground truth, and preference-based win rates to identify whether RLAIIF succeeds in ethical reasoning tasks.

Our experiments reveal both encouraging and concerning findings for AI-driven ethical alignment. Supervised fine-tuning substantially improve ethical reasoning, achieving accuracy gains of up to 96.93% while producing coherent, framework-aligned justifications. Distilled reward models successfully learn to discriminate response quality, demonstrating consistent scoring behavior across ethical frameworks. However, both RLAIIF variants exhibit unexpected performance during reinforcement learning, reverting to near-baseline accuracy despite receiving reward signals from well-trained models. This counterintuitive outcome reveals a critical mismatch: reward models trained on high-quality AI outputs impose expectations that exceed the policy model’s optimization capacity, leading to reward hacking (Amodei et al. 2016; Everitt et al. 2017) rather than incremental improvement. These findings highlight fundamental challenges in applying standard RL techniques to complex moral reasoning, and underscore the need for better reward models or alternative optimization strategies.

Related Works

Research on Ethical Capabilities of LLMs

(Senthilkumar et al. 2024) addressed the LLM performance in morally ambiguous scenarios by comparing finetuned and baseline model predictions against human judgments. They assessed GPT-4o, Llama-3.1-8B, Zephyr-7B-Beta, and Mistral-7B on two task types: Anecdotes, which require binary moral classification of real-life scenarios, and Dilemmas, pairs of scenarios where models generated probability scores to indicate which scenario is more unethical. Model responses were evaluated against human consensus and annotations using two loss functions, Binary Cross-Entropy & Dirichlet-Multinomial Loss, for a refined evaluation. Baseline models achieved loss values of 3-5 on Dilemmas and 10-12 on Anecdotes, indicating particular difficulty with descriptive moral narratives. Fine-tuning yielded modest improvements, reducing Dilemma loss to 3-3.5 and Anecdote loss to 8-9. Despite these moderate increases in performance, ambiguous scenarios proved to be challenging, necessitating the application of superior methods.

(Hendrycks et al. 2020) introduced the ETHICS dataset, an ethical benchmark consisting of moral scenarios structured by commonsense morality, justice, virtue ethics, utilitarianism, and deontology theories. Scenarios are assigned human-annotated truth labels and filtered using adversarial filtration, with unambiguous and slightly ambiguous splits. Each scenario presents LLMs with a binary choice to indicate either a preference or alignment of a presented response to a moral scenario. Findings show low average alignment to human labels across tested models, including GPT-3, BERT-

base, RoBERTa-large, and ALBERT-xxlarge (49.1% on unambiguous), with a 5-10% improvement in larger variants, yet still below 50% for slightly ambiguous scenarios. Such evaluations on clear moral choices demonstrate a growing need for developing strategies to substantially improve LLM reasoning due to under-trained ways of thinking.

Strategies for Improving Ethical Reasoning

Despite the lack of substantial ethical reasoning, several strategies have proved to improve LLM reasoning capabilities, which can be repurposed for ethics.(Ouyang et al. 2022) propose Reinforcement Learning from Human Feedback (RLHF) as a solution to realign LLMs to follow user instructions. Pretrained GPT-3 models of varying sizes (1.3B, 6B, 175B) were first fine-tuned using supervised fine-tuning (SFT), improved using reward model (RM) training (Stiennon et al. 2020), and separately, Proximal Policy Optimization (PPO) (Schulman et al. 2017), a cyclic reward-feedback system. Each step used human-labeled datasets, where human-provided ranks and responses were matched to each given prompt. In held-out prompts, RLHF-tuned model responses were preferred by human annotators over base variants 85% of the time, with a 20% increase in instruction alignment and 25% decrease in hallucination. However, this required extensive annotations from human workers, leaving out the potential of reducing the required human guidance without compromising on improvement.

To keep human intervention at a minimum, (Bai et al. 2022) enveloped human-aligned values plainly into a constitution, named Constitutional AI, and modified RLHF to train models using an RM trained with AI preferences called RLAIIF. This two-fold strategy was implemented using iterative critique-based fine-tuning with principles in a human-defined constitution as the standard to critique. Instead of using human feedback, the tuned model is distilled to AI pair-wise preferences and is leveraged to train the originally tuned model. A dataset of LLM-ranked responses to harmfulness prompts was used alongside a separate human feedback helpfulness dataset to assess both categories. ELO scores, calculated using crowd-worker preferences, present RLAIIF as significantly more harmless than SFT and RLHF, albeit with limitations such as overfitting. Even so, there are also additional improvements that can be made from the proposed baseline RLAIIF.

(Lee et al. 2023) show how RLAIIF is comparable to RLHF and propose a distinct method to further improve RLAIIF called d-RLAIIF (direct-RLAIIF). Unlike traditional RLAIIF (Bai et al. 2022), where a smaller reward model is trained on LLM preferences, d-RLAIIF leverages an off-the-shelf LLM for reward signals. To train and evaluate each of these frameworks, the authors used three datasets consisting of summarization and helpfulness/harmlessness tasks. RLAIIF performed comparably and outperformed RLHF with a win rate of 50% in summarization, 52% in helpful dialogue, and 88% in harmless dialogue, while d-RLAIIF surpassed same-size RLAIIF with a win rate of 60%. This sets the precedent for utilizing an off-the-shelf LLM directly as

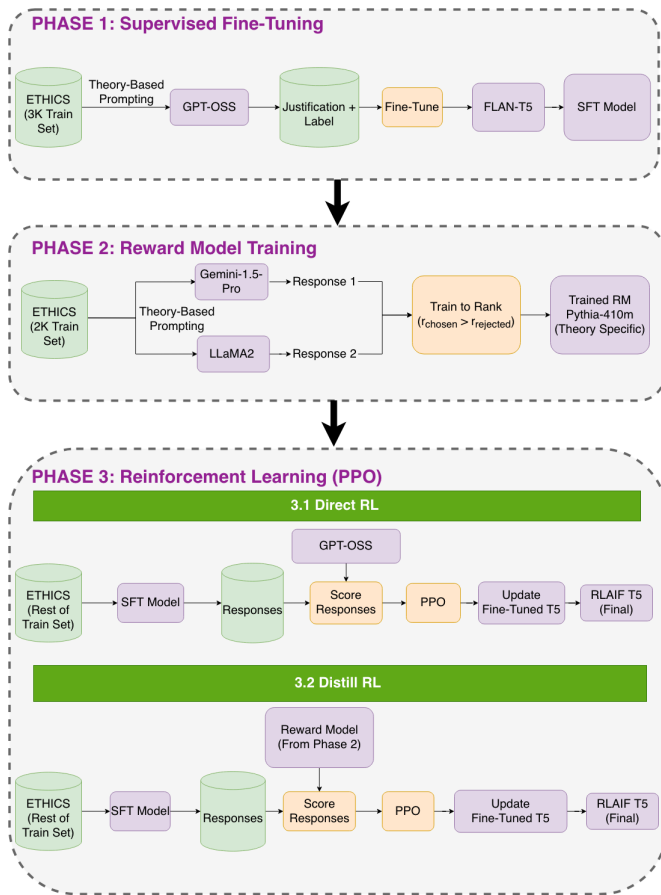


Figure 1: RLAIF Framework for Training Ethical Models

an alternative primary agent for aligning LLMs towards improved ethical reasoning.

Methodology

Our study implements the RLAIF framework for training LLMs to perform ethical reasoning across multiple moral frameworks. The complete pipeline is illustrated in Figure 1 and comprises three sequential phases: supervised fine-tuning to establish a baseline and equip the policy model with reasoning capacity for preference optimization, reward model training to encode framework-specific preference distributions from AI-generated justifications, and reinforcement learning via PPO to align the policy model with learned preferences. The detailed methodology for each stage is described below.

Dataset Preparation

We employed the ETHICS corpus (Hendrycks et al. 2020), which is a comprehensive benchmark of moral scenarios categorized into five major ethical frameworks, including Commonsense Morality, Deontology, Justice, Utilitarianism, and Virtue Ethics. Table 1 presents the distribution of samples

Table 1: Number of samples across data splits and ethical frameworks

Split	Justice	Virtue	Deontology	Utilitarianism	Commonsense
Train	21,791	28,245	18,164	13,738	13,910
Test	2,704	4,975	3,596	4,808	3,885
Hard Test	2,052	4,780	3,536	4,272	3,964

across data splits and ethical frameworks in the original ETHICS dataset.

For preference dataset construction in Phase 2, we utilized AI-generated justifications from our prior work (Thapa et al. 2025), where ethical scenarios from the ETHICS benchmark were paired with theory-specific prompts to generate moral justifications using multiple LLMs. In particular, we reuse the justifications generated by LLaMA-2 and Gemini-1.5-Pro, which were previously validated for alignment with human labels. These models generated justifications for scenarios from the ETHICS Test split across all five frameworks, which we leverage to construct preference pairs for reward model training.

The original training split from the ETHICS corpus was subdivided into three functional subsets for the RLAIF pipeline.

- Supervised Fine-Tuning (SFT):** We allocated 3,000 samples for fine-tuning the policy model on ethical text classification.
- Reward Model Training (RM):** We separated 2,000 samples per category to train a lightweight RM, which is capable of distinguishing between ‘preferred’ vs. ‘non-preferred’ justification.
- Reinforcement Learning (RL):** The remaining data from the training set was allocated for policy optimization under the reward model feedback.

For the Direct RLAIF variant (d-RLAIF), we utilize 100% of the training data for the RL phase, as this approach bypasses the explicit reward model training step by leveraging an off-the-shelf LLM directly for reward signals. Furthermore, test splits across all moral frameworks were allocated for evaluation on RM Accuracy, Model Accuracy, and Win Rate metrics.

Model Architecture

Policy Model: We employ Flan-T5-XL (Chung et al. 2024) as our policy model due to its encoder-decoder architecture and instruction-tuning pretraining, which provides a strong foundation for ethical reasoning tasks.

Reward Model (Distilled): For the Distilled RLAIF framework, we utilize Pythia-410M (Biderman et al. 2023) as the RM. Despite its compact size, Pythia-410M demonstrates efficient learning capabilities while maintaining computational efficiency.

Reward Model (Direct): In the Direct RLAIF framework, we leverage GPT-OSS (Agarwal et al. 2025), which is an open-source LLM that serves as an off-the-shelf reward signal generator without requiring explicit training.

This approach enables direct preference evaluation, completely eliminating reward model training overhead.

Preference Dataset Construction

Justification Generation: We query multiple state-of-the-art LLMs, including LLaMA-2 (Touvron et al. 2023), LLaMA-3.1 (Dubey et al. 2024), Gemini-1.5-Pro, and Mistral (Jiang et al. 2023), to generate textual justifications for each scenario in the ETHICS test set for each framework. This multi-model approach produces diverse reasoning patterns and justification styles across the candidate pool for each ethical framework.

Model Selection: We evaluate the generated justifications against the ETHICS ground-truth labels to identify the most accurate and reliable justification generator across all ethical theories.

Preference Labeling: We employed Gemini-1.5-Pro in an evaluator role to create structured preference datasets suitable for reward modeling across all ethical theories. For each theory, the model assesses the justification pairs for the same scenarios and identifies which justification is more persuasive, contextually appropriate, and better aligned with that specific ethical framework’s reasoning principles. This procedure yields theory-specific labeled datasets of preference pairs, where each pair consists of a “chosen” justification (preferred response) and a “rejected” justification (rejected response). The resulting datasets capture relative justification quality within each ethical framework rather than just binary correctness, aligning with the requirements of preference-based reinforcement learning.

Reward Model (RM) Training

Using the constructed preference datasets, we train separate reward models for each ethical theory to encode theory-specific ranking over justification quality. The training objective optimizes each model to assign higher scalar reward values to preferred (chosen response) justifications and lower values to non-preferred (rejected response) alternatives within its respective ethical framework. Specifically, for a scenario ‘*s*’ and justification ‘*j*’ for a given theory, each reward model learns to predict:

$$r(s, j_{\text{chosen}}) > r(s, j_{\text{rejected}})$$

These trained RMs capture the quality justification specific to each moral framework and serve as the feedback mechanisms in the form of rewards for subsequent policy optimization.

Reinforcement Learning (RL) using Proximal Policy Optimization (PPO)

The final phase employs the Proximal Policy Optimization (PPO) to fine-tune the Flan-T5-XL policy model using reward signals from the trained reward models (Distilled RLAIF) or the off-the-shelf LLM (Direct RLAIF). We train separate policy models for each ethical theory to specialize in theory-specific reasoning patterns. PPO was selected for

its stability in reinforcement learning with neural networks and its widespread use in RLHF and RLAIF.

During each PPO iteration, the policy model generates justifications for ethical scenarios from the RL training split of a given ethical theory. The corresponding reward model evaluates these generated justifications and assigns scalar rewards based on alignment with that theory’s principles and values. The PPO algorithm then updates the policy parameters to maximize expected reward. The fine-tuning process adjusts the policy model such that its generated justifications align more closely with the preferences encoded in the reward model.

Evaluation Metrics

We employ three distinct evaluation metrics to assess model performance across different stages of the pipeline for each ethical theory:

- Reward Model Accuracy:** We measured the RM’s ability to correctly rank the model outputs according to the learned preferences within each ethical framework. It represents the percentage of preference pairs where the reward model correctly ranked the preferred justification higher. For evaluation, we construct test pairs where Gemini-1.5-Pro outputs are categorized as “chosen” and Llama-2 outputs as “rejected,” reflecting the preference distributions learned during training. Each theory-specific RM is evaluated as a percentage of pairs for which it assigns higher scores to Gemini responses for scenarios in that theory.
- Policy Model Accuracy:** We evaluated the alignment between the policy model predictions and the human-annotated ground truth labels in the ETHICS dataset. This metric represents the proportion of generated justifications whose moral choice matched the ground truth. It directly measures the model’s ethical reasoning capability within each specific moral framework.
- Win Rate:** We compared the two trained policy model variants by presenting their outputs to the trained RM for a given ethical theory. One variant is designated as “chosen” and the other as “rejected.” The win rate represents the proportion of instances where the RM assigns higher scores to the chosen variant’s responses, indicating learned preference alignment within that ethical framework.

Experimental Setup

Initial experiments were conducted using virtual GPUs provided by Google Cloud’s Vertex AI to validate and refine the RLAIF pipeline on larger models and datasets. To ensure correctness and reproducibility of our implementation, we first validated the RLAIF pipeline by replicating a known baseline setup prior to applying it to ethical reasoning tasks. Specifically, we reimplemented the RLAIF framework introduced by Olano and Evangeliou (Olano and Evangeliou 2023) and evaluated it on the CarperAI/openai-summarize-from-feedback dataset using the pretrained JuanKO/rlhf policy model.

To maintain computational feasibility without sacrificing reward quality, we used the Pythia-410M as the reward model. This replication stage confirmed that our implementation faithfully reproduced expected improvements in policy behavior under reinforcement learning, establishing a reliable foundation for subsequent experiments on ethical reasoning.

Final experiments were conducted using two NVIDIA A6000 GPUs for training and a separate remote server equipped with a 15 GB GPU for intermediate evaluation and metric computation. These hardware constraints informed our selection of lightweight yet capable models for both policy and reward components. As a result, all experiments were designed to balance computational feasibility with sufficient baseline ethical reasoning capacity, ensuring fair evaluation across all stages of the RLAIIF pipeline.

Results

Policy Model Accuracy

Table 2 presents the policy model performance across all the ethical frameworks and training stages. Supervised fine-tuning achieved the highest accuracy across nearly all frameworks, with peak performance of 96.93% in Utilitarianism and 78.61% in Virtue Ethics. However, closer inspection of the Utilitarianism dataset revealed a positional bias in human annotations absent in other frameworks, likely inflating SFT performance. Excluding this artifact, Virtue Ethics and Commonsense Morality emerged as the strongest-performing frameworks, suggesting that mainstream moral reasoning patterns are more readily learned by instruction-tuned models.

The base Flan-T5-XL model exhibited weak ethical classification capabilities, achieving a maximum accuracy of only 57.94% for Justice, while frequently generating incoherent or task-irrelevant responses. In contrast, SFT models demonstrated substantial improvements in both label alignment and response coherence across frameworks. The notable exception was Deontology, where all training stages produced only marginal gains, peaking at 51.92% with distilled RLAIIF. This framework-specific difficulty suggests fundamental limitations in encoding duty-based reasoning through current alignment techniques.

In contrast, both Distilled RLAIIF and Direct RLAIIF exhibited regression toward base model performance rather than improvement over SFT. Direct RLAIIF marginally outperformed Distilled RLAIIF in Virtue Ethics (56.56% vs. 56.40%), Utilitarianism (33.55% vs. 32.24%), and Justice (56.96% vs. 59.45%), though both methods performed far below the SFT and on par with the baselines. Analysis of reward signal distributions during PPO training revealed that both reward model variants consistently assigned low scores (0.0-0.3 on a normalized 0-1 scale) even to well-reasoned SFT-quality responses. This miscalibration could have led to incentivizing a form of reward hacking where policy models discover that reverting to base-model-style outputs or generic responses with minimal ethical reasoning received higher reward scores than better justifications. This finding exposes a critical mismatch between the reward model’s ex-

Table 2: Policy accuracy across ethical theories

Theory	Base	SFT	Distilled RLAIIF	Direct RLAIIF
Commonsense	51.99%	62.73%	52.34%	51.31%
Virtue Ethics	55.46%	78.61%	56.40%	56.56%
Utilitarianism	33.27%	96.93%	32.24%	33.55%
Deontology	51.24%	48.50%	51.92%	51.38%
Justice	57.94%	61.45%	59.45%	56.96%

pectations, which were trained on high-quality AI-generated justifications, and the policy model’s optimization capacity under PPO.

Reward Model Accuracy

Due to inconsistent scoring behavior across ethical theories for the base reward models, we computed base RM accuracy by averaging five independent evaluations per framework with corresponding standard deviations as depicted in Table 3.

Base models exhibited near-random performance across most frameworks (50% accuracy), with notable exceptions in Virtue Ethics (69.16% \pm 21.10%) and Justice (55.14% \pm 26.22%). However, the high variance underscores the unreliability. Trained reward models demonstrated substantial and consistent improvements, most notably in Justice (93.75%), Commonsense Morality (85.66%), and Virtue Ethics (71.10%). However, an unexpected pattern emerged in Virtue Ethics and Justice, where responses initially labeled as “rejected” during preference dataset construction actually exhibited greater detail and elaboration than corresponding “chosen” responses, despite sharing similar reasoning quality. To account for this labeling artifact, we report inverted accuracies for these frameworks (denoted by * in Table 3), representing the proportion of cases where the RM correctly identified the more detailed response. This inversion does not diminish the core finding that the trained RMs successfully learned to discriminate response quality without exhibiting sensitivity to superficial formatting or presentation order.

Furthermore, RMs also excelled at Virtue Ethics and Commonsense which reinforces the training bias of certain theories present even in base models, which are further enforced after training. Critically, unlike policy model training, where parameter count was limiting, the compact Pythia-410M architecture proved sufficient for effective reward modeling, suggesting that preference learning is less constrained by model capacity than policy optimization under RL.

Win Rate

To validate whether reward models successfully internalized preference distributions and whether response quality improved across training stages, we conducted pairwise preference evaluations using trained RMs as judges, referred to as the win rate. Table 4 presents win rates comparing SFT against base model outputs. Deontology achieved the highest win rate of 83.20%, indicating that despite poor performance in both policy and RM accuracy, trained RMs effectively distinguished coherent ethical reasoning from inco-

Table 3: Reward Model Accuracy Across Ethical Theories

Theory	Base	Trained
Commonsense	41.79 ± 11.43%	85.66%
Virtue Ethics	69.16 ± 21.10%	*71.10%
Utilitarianism	49.68 ± 15.86%	64.08%
Deontology	46.16 ± 26.93%	69.38%
Justice	55.14 ± 26.22%	*93.75%

(*) - Inverse of true percentages to reflect inverse in response quality

herent responses for this framework. The win rate for Virtue Ethics and Justice followed closely at 74.12% and 73.68%, respectively. These results confirm that SFT substantially improved response quality as measured by learned preference models, though the consistent 15-30% preference for base responses reveals that even SFT outputs fall considerably short of the Gemini-1.5-Pro quality standard used during RM training. Providing an even more capable model would make these improvements potentially better by being able to identify nuance in certain scenarios.

Table 4: SFT Win Rate
SFT vs. Base

Commonsense	71.72%
Virtue Ethics	74.12%
Utilitarianism	69.07%
Deontology	83.20%
Justice	73.68%

Table 5 extends this analysis to RLAIIF variants. Both Distilled and Direct RLAIIF exhibited markedly low win rates against SFT across all frameworks. Direct RLAIIF achieved a maximum of only 32.26% preference over SFT for Utilitarianism, while Distilled RLAIIF peaked at 49.12% for Virtue Ethics. These win rates, approaching near-random selection confirms the regression in response quality observed in accuracy metrics. In Virtue Ethics, Distilled RLAIIF achieved a near-random win rate of 49.12% against SFT despite generating objectively inferior justifications, showing clear evidence of reward hacking, where the policy model learned to exploit reward model scoring patterns rather than genuinely improving ethical reasoning. This failure points to a fundamental capacity gap of the Flan-T5-XL policy model, which lacked sufficient ethical reasoning capacity to optimize toward the high-quality preference distributions that were encoded in the reward models. A more capable policy model architecture with greater parameter capacity and stronger reasoning capabilities may be necessary to successfully leverage AI feedback for ethical alignment.

Direct comparisons between RLAIIF variants (Direct vs. Distill) also showed marginal differentiation, with Direct RLAIIF slightly preferred in Deontology with 50.82%, Commonsense with 50.68%, and Virtue Ethics with 50.37% win rate. However, given that both variants produced demonstra-

Table 5: Win rate between tuned and trained model variations

Theory	Direct vs. SFT	Distilled vs. SFT	Direct vs. Distilled
Commonsense	29.54%	29.26%	50.68%
Virtue Ethics	26.04%	49.12%	50.37%
Utilitarianism	32.26%	25.56%	27.27%
Deontology	15.75%	15.27%	50.82%
Justice	27.29%	48.44%	49.51%

bly poor-quality outputs, these slight differences likely reflect noise rather than meaningful preference. This finding contrasts with prior work demonstrating clear advantages of Direct RLAIIF over distilled approaches in summarization and dialogue tasks (Lee et al. 2023). Our results suggest that the architectural choice between distilled and direct reward modeling becomes insignificant when both methods fail to surpass SFT baselines due to policy model limitations. The critical bottleneck lies not in the mechanism of generating reward signals, but in the ability of the policy model to learn from those signals. Overcoming this outcome may require fundamentally different model architectures or alternative optimization strategies beyond standard PPO.

Conclusion

In this paper, we investigated the effectiveness of the RLAIIF framework for ethical reasoning across five moral frameworks using the ETHICS benchmark. Our results reveal both promise and fundamental limitations. Supervised finetuning achieved substantial improvements, demonstrating that instruction-tuned models can encode ethical reasoning patterns. In contrast, while distilled reward models demonstrated robust and consistent ability to discriminate response quality by achieving high preference ranking accuracy across theories, neither distilled nor direct RLAIIF translated these strong reward signals into improved policy performance. Instead, both approaches exhibited regression toward near-baseline performances during reinforcement learning, despite receiving well-calibrated feedback. Our analysis suggests that this failure is not due to noisy or uninformative AI-generated preferences, but rather to a fundamental mismatch between reward model expectations and policy model capacity. The reward models which were trained on high-quality AI-generated justifications, generated rewards that the policy model could not effectively optimize, leading to reward hacking rather than genuine improvement in the justification quality. These results highlight a critical constraint in applying RLAIIF to complex moral reasoning tasks and suggest that substantial progress may require stronger policy models, improved reward calibration, or alternative optimization strategies.

References

- Agarwal, S.; Ahmad, L.; Ai, J.; Altman, S.; Applebaum, A.; Arbus, E.; Arora, R. K.; Bai, Y.; Baker, B.; Bao, H.; et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.
- Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schul-

- man, J.; and Mané, D. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.
- Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The moral machine experiment. *Nature* 563(7729):59–64.
- Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Biderman, S.; Schoelkopf, H.; Anthony, Q. G.; Bradley, H.; O’Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2397–2430. PMLR.
- Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research* 25(70):1–53.
- Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints arXiv:2407*.
- Everitt, T.; Krakovna, V.; Orseau, L.; Hutter, M.; and Legg, S. 2017. Reinforcement learning with a corrupted reward channel. *arXiv preprint arXiv:1705.08417*.
- Gorwa, R.; Binns, R.; and Katzenbach, C. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7(1):2053951719897945.
- Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7b.
- Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günemann, S.; Hüllermeier, E.; et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences* 103:102274.
- Lee, H.; Phatale, S.; Mansoor, H.; Mesnard, T.; Ferret, J.; Lu, K.; Bishop, C.; Hall, E.; Carbune, V.; Rastogi, A.; et al. 2023. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.
- Lee, P.; Bubeck, S.; and Petro, J. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine* 388(13):1233–1239.
- Olano, J., and Evangelidou, P. 2023. RLAIF – reinforcement learning with AI feedback. <https://github.com/AI-Maker-Space/RLXF-Community-Sessions/tree/main/rlaif>. GitHub repository.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35:27730–27744.
- Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Senthilkumar, P.; Balasubramanian, V.; Jain, P.; Maity, A.; Lu, J.; and Zhu, K. 2024. Fine-tuning language models for ethical ambiguity: A comparative study of alignment with human responses. *arXiv preprint arXiv:2410.07826*.
- Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems* 33:3008–3021.
- Takemoto, K. 2024. The moral machine experiment on large language models. *Royal Society open science* 11(2):231393.
- Thapa, B.; Duarte-Vasquez, A.; Hooshmand, S.; and Rathore, H. 2025. Evaluation of ethical decision making in large language models across classical moral frameworks. In *2025 IEEE International Conference on Artificial Intelligence Testing (AITest)*, 110–117. IEEE.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.