Raw Parsed

{
  "query": "alignment",
  "num_results": 3,
  "results": [
    {
      "chunk": "agents that behave morally. arXiv preprint arXiv:2110.13136. Hu, B.; Ray, B.; Leung, A.; Summerville, A.; Joy, D.; Funk, C.; and Basharat, A. 2024. Language Models are Alignable Decision-Makers: Dataset and Application to the Medical Triage Domain. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Com- putational Linguistics: Human Language Technologies (Vol- ume 6: Industry Track), 213–227. Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; et al. 2023. Ai alignment: A comprehensive survey. arXiv preprint arXiv:2310.19852. Jiang, L.; Hwang, J. D.; Bhagavatula, C.; Bras, R. L.; Liang, J. T.; Levine, S.; Dodge, J.; Sakaguchi, K.; Forbes, M.; Hes- sel, J.; et al. 2025. Investigating machine moral judgement through the Delphi experiment. Nature Machine Intelli- gence, 7(1): 145–160. Kiely, M.; Ahiskali, M.; Borde, E.; Bowman, B.; Bowman, D.; Van Bruggen, D.; Cowan, K.; Dasgupta, P.; Devendorf, E.; Edwards, B.; et al. 2025. Exploring the efficacy of multi- agent reinforcement learning for autonomous cyber defence: A cage challenge 4 perspective. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 28907–28913. Lee, H.; Phatale, S.; Mansoor, H.; Lu, K. R.; Mesnard, T.; Ferret, J.; Bishop, C.; Hall, E.; Carbune, V.; and Rastogi, A. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. Leike, J.; Krueger, D.; Everitt, T.; Martic, M.; Maini, V.; and Legg, S. 2018. Scalable agent alignment via reward model- ing: a research direction. arXiv preprint arXiv:1811.07871. Liu, O.; Fu, D.; Yogatama, D.; and Neiswanger, W. 2024. Dellma: Decision making under uncertainty with large lan- guage models. arXiv preprint arXiv:2402.02392. Liu, X.; Yoneda, T.; Stevens, R. L.; Walter, M. R.; and Chen, Y. 2023. Blending imitation and reinforcement learning for robust policy improvement. arXiv preprint arXiv:2310.01737. Meng, X.; Yan, X.; Zhang, K.; Liu, D.; Cui, X.; Yang, Y.; Zhang, M.; Cao, C.; Wang, J.; Wang, X.; et al. 2024. The application of large language models in medicine: A scoping review. Iscience, 27(5). 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 pro- cessing systems, 35: 27730–27744. Pan, A.; Chan, J. S.; Zou, A.; Li, N.; Basart, S.; Wood- side, T.; Zhang, H.; Emmons, S.; and Hendrycks, D. 2023. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the Machiavelli Benchmark. In Krause, A.; Brunskill, E.; Cho, K.; Engel- hardt, B.; Sabato, S.; and Scarlett, J., eds., Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, 26837–26867. PMLR. Perez, E.; Ringer, S.; Lukosiute, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; et al. 2023. Discovering language model behaviors with model-written evaluations. In Findings of the association for computational linguistics: ACL 2023, 13387–13434. Rigley, E.; Chapman, A.; Evers, C.; and Mcneill, W. 2025. ME: Modelling Ethical Values for Value Alignment. In Pro- ceedings of the",
      "metadata": {
        "paper_id": "2511.11551v1",
        "chunk_index": 12,
        "total_chunks": 24
      },
      "distance": 1.2848505973815918,
      "rank": 1

```
},
{
  "chunk": "verse training-time alignment, in cases where the original objectives may be undesirable. We also analyze positive and negative correlations between attributes, which can inform the selection of alignment targets. • A comparison of our method with prior environment- specific alignment methods, including training-time pol- icy shaping and LLM agents, provides empirical evi- dence of superior alignment by our approach. 2 Related Work 2.1 LLM Agent Alignment Research on the alignment of LLM agents has gained mo- mentum due to their increasing use in decision-making set- tings. For LLMs, reward modeling from human preferences has reduced harmful behaviors (Ouyang et al. 2022), and multi-objective methods can adapt LLMs to multiple pref- erences (Gupta et al. 2025). Recent work also includes con- stitutional AI, where models utilize predefined ethical prin- ciples to critique and guide their outputs, and RL from AI Feedback (RLAIF) (Lee et al. 2023) that scales alignment by replacing human feedback with model-based feedback. Sim- ilarly, test-time techniques, such as zero-shot prompts (Hu et al. 2024), chain-of-thought reasoning (Liu et al. 2024), and structured reasoning frameworks (Chen et al. 2025), have been used to support ethical decision-making. 2.2 RL Agent Alignment: Reward and Policy Compared to LLM agents, RL agents optimize behavior through interaction and reward, enabling stronger perfor- mance in tasks requiring long-term planning and real-time feedback, such as games (Pan et al. 2023), robotics (Wang et al. 2024), and cybersecurity (Kiely et al. 2025). Align- ing these agents with human intent typically involves hu- man feedback, either through reward modeling and prefer- ence learning (Christiano et al. 2017; Leike et al. 2018) or reward shaping (Goyal, Niekum, and Mooney 2019). An alternative approach is policy shaping, which directly modifies an RL agent’s policy using human feedback, ad- dressing issues like reward hacking and ambiguity in reward signals (Griffith et al. 2013; Rigley et al. 2025). Our ap- proach is similar to (Pan et al. 2023; Hendrycks et al. 2021) in applying policy shaping with external classifiers to guide RL agents. However, these are training-time methods and re- quire agent retraining, which limits flexibility and scalabil- ity. In contrast, our test-time approach enables fine-grained, scalable control over alignment attributes and adjustment of the trade-off between reward and ethical behavior. 2.3 Safe RL and Moral Value Alignment Value alignment in AI systems is a nuanced challenge, as human values and intentions can vary widely, necessitating flexible and diverse alignment constraints (Sorensen et al. 2024). Prior work in RL has shown that misaligned agents can develop power-seeking behavior (Turner et al. 2019; Pan et al. 2023; Perez et al. 2023; Ji et al. 2023). However, it has also been shown that AI models can recognize moral judg- ments (Jiang et al. 2025), supporting the development of eth- ical decision-making. Pan et al. (2023) and Hendrycks et al. (2021) are closest to our work, and characterize ethical be- haviors using broad attributes such as power, disutility, and ethical violations. In contrast, we introduce a fine-grained framework for specifying individual moral and ethical val- ues and examine the relationships between these attributes in agents",
  "metadata": {
    "paper_id": "2511.11551v1",
    "chunk_index": 2,
    "total_chunks": 24
  },
  "distance": 1.3305654525756836,
  "rank": 2
},
{
```

"chunk": "often rely on a rigid, predefined set of ethical norms. In reality, values for alignment can vary widely across cultures, communities, and application contexts (Sorensen et al. 2024), making the adaptability of alignment a challenging problem. The limited generaliz- ability of alignment attributes across domains further com- pounds this problem, e.g., when relying on domain-specific preferences (Ji et al. 2023). Although task-specific agents excel within their domains, maintaining ethical consistency and performance across environments is not scalable, as it often requires retraining (Zhou et al. 2022). To address these challenges, we propose a novel test-time approach for aligning text-based RL agents (Fig. 1). Using lightweight classifiers, pre-trained agents are steered through model-guided policy shaping, a method in which external feedback adjusts the agent’s policy or action selec- tion probabilities (Griffith et al. 2013). This approach con- trasts with alignment methods that rely heavily on training- time interventions or post hoc fine-tuning (Pan et al. 2023; Hendrycks et al. 2021), and instead enables guidance with- out retraining, improving adaptability across environments and reward functions. This adaptability is crucial for align- ing agents across diverse tasks, as ethical priorities of- ten vary by application (Gabriel 2020; Awad et al. 2018). By steering behavior along specific alignment dimensions rather than broad categories, our method also enables more interpretable and context-sensitive control. Overall, the main contributions of our paper are: • A novel test-time, model-driven, policy-shaping ap- proach for aligning text-based agents trained to maxi- mize reward, that also supports generalization across en- vironments despite the agents being trained in specific environments. • A thorough evaluation on the MACHIAVELLI bench- mark (Pan et al. 2023), covering a diverse set of agents arXiv:2511.11551v1 [cs.AI] 14 Nov 2025 Figure 1: Overview of our proposed alignment approach using test-time policy shaping. Given a scenario, ethical attribute classifiers predict the likelihood of different attributes for each available action. These predictions are then used to adjust an agent’s policy during inference to discourage actions misaligned with ethical target attributes, e.g. avoiding killing. trained in multiple text-based game environments. The agents are assessed on Machiavellian behaviors, includ- ing 10 morality, four power-seeking, and the disutility at- tributes. We have also contributed a new interactive deci- sion trajectory viewer (Fig. 3) that clearly illustrates the decisions and their alignment to ethical behavior made by an agent across game scenarios. • A study of the trade-off between reward and ethical be- havior in pre-trained agents, exploring different align- ment tensions, such as the effects of varying the weights between reward and different moral or power-seeking at- tributes. Our approach enables fine-grained steering of agent behavior along the Pareto front of ethical alignment with agent reward. In such cases, we also demonstrate the ability to steer an agent in any direction and to re- verse training-time alignment, in cases where the original objectives may be undesirable. We also analyze positive and negative correlations between attributes, which can inform the selection of alignment targets. • A comparison of our method with prior environment- specific alignment methods, including training-time pol- icy shaping and LLM agents, provides",
"metadata": {
"paper_id": "2511.11551v1",
"chunk_index": 1,
"total_chunks": 24
},
"distance": 1.389431118965149,
"rank": 3
}
]

}