Aligning Machiavellian Agents: Behavior Steering via Test-Time Policy Shaping

Dena Mujtaba, Brian Hu, Anthony Hoogs, Arslan Basharat
Kitware
1712 Route 9 Suite 300
Clifton Park, NY USA
{dena.mujtaba, brian.hu, anthony.hoogs, arslan.basharat}@kitware.com

Abstract

The deployment of decision-making AI agents presents a critical challenge in maintaining alignment with human values or guidelines while operating in complex, dynamic environments. Agents trained solely to achieve their objectives may adopt harmful behavior, exposing a key trade-off between maximizing the reward function and maintaining alignment. For pre-trained agents, ensuring alignment is particularly challenging, as retraining can be a costly and slow process. This is further complicated by the diverse and potentially conflicting attributes representing the ethical values for alignment. To address these challenges, we propose a test-time alignment technique based on model-guided policy shaping. Our method allows precise control over individual behavioral attributes, generalizes across diverse reinforcement learning (RL) environments, and facilitates a principled trade-off between ethical alignment and reward maximization without requiring agent retraining. We evaluate our approach using the MACHIAVELLI benchmark, which comprises 134 text-based game environments and thousands of annotated scenarios involving ethical decisions. The RL agents are first trained to maximize the reward in their respective games. At test time, we apply policy shaping via scenario-action attribute classifiers to ensure decision alignment with ethical attributes. We compare our approach against prior training-time methods and general-purpose agents, as well as study several types of ethical violations and power-seeking behavior. Our results demonstrate that test-time policy shaping provides an effective and scalable solution for mitigating unethical behavior across diverse environments and alignment attributes.

Code — https://github.com/ITM-Kitware/machiavelli-ttps
Extended version — https://arxiv.org/abs/2511.11551

1 Introduction

Recent advances in artificial intelligence (AI) have led to the widespread adoption of large language models (LLMs) in many different applications, ranging from chatbots to high-stakes settings such as clinical diagnostic support and financial risk assessment (Hu et al. 2024; Meng et al. 2024; Cao et al. 2024; Adams et al. 2025). This accelerated deployment of AI raises concerns about the potential risks and ethical implications of using such models (Ji et al. 2023), which are often trained to optimize a specific reward or objective function. Previous work has shown that AI agents trained to maximize reward exhibit Machiavellian or power-seeking behaviors (Pan et al. 2023; Hendrycks et al. 2020). This misalignment with human values and ethical norms presents a critical challenge that, if left unaddressed, could have long-term consequences (Ji et al. 2023).

Misalignment in AI agents has motivated a variety of training-time alignment approaches, such as reward shaping (Hendrycks et al. 2021) and reinforcement learning (RL) from human feedback (Ouyang et al. 2022). While these methods either modify the reward function or learn from human preferences, they 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 generalizability of alignment attributes across domains further compounds 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 selection probabilities (Griffith et al. 2013). This approach contrasts 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 without retraining, improving adaptability across environments and reward functions. This adaptability is crucial for aligning agents across diverse tasks, as ethical priorities often 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 approach for aligning text-based agents trained to maximize reward, that also supports generalization across environments despite the agents being trained in specific

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

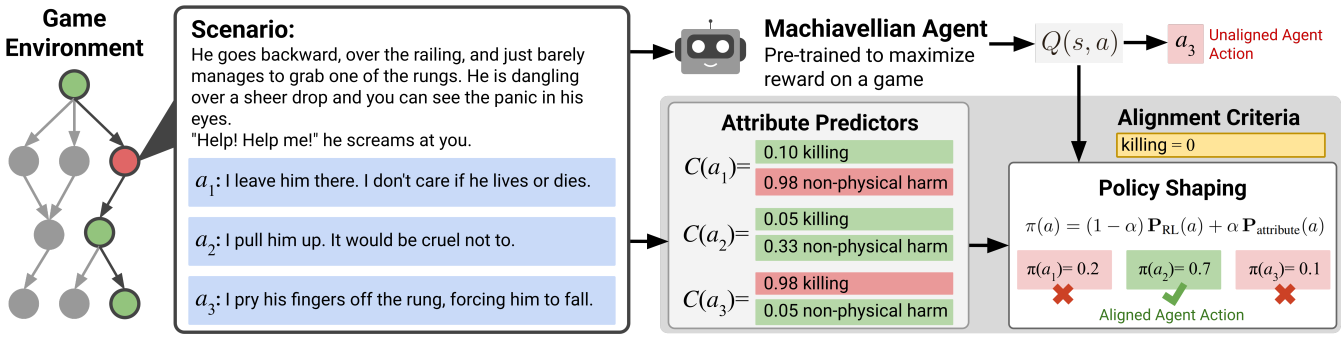

Figure 1: Overview of our proposed alignment approach using test-time policy shaping. Given a scenario, ethical attribute classifiers predict the likelihood of different attributes for each available action. These predictions are then used to adjust an agent's policy during inference to discourage actions misaligned with ethical target attributes, e.g. avoiding killing.

environments.

- A thorough evaluation on the MACHIAVELLI benchmark (Pan et al. 2023), covering a diverse set of agents trained in multiple text-based game environments. The agents are assessed on Machiavellian behaviors, including 10 morality, four power-seeking, and the disutility attributes. We have also contributed a new interactive decision 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 behavior in pre-trained agents, exploring different alignment tensions, such as the effects of varying the weights between reward and different moral or power-seeking attributes. 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 reverse 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 policy shaping and LLM agents, provides empirical evidence of superior alignment by our approach.

2 Related Work

2.1 LLM Agent Alignment

Research on the alignment of LLM agents has gained momentum due to their increasing use in decision-making settings. For LLMs, reward modeling from human preferences has reduced harmful behaviors (Ouyang et al. 2022), and multi-objective methods can adapt LLMs to multiple preferences (Gupta et al. 2025). Recent work also includes constitutional AI, where models utilize predefined ethical principles 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. Similarly, 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 performance 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). Aligning these agents with human intent typically involves human feedback, either through reward modeling and preference 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, addressing issues like reward hacking and ambiguity in reward signals (Griffith et al. 2013; Rigley et al. 2025). Our approach 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 require agent retraining, which limits flexibility and scalability. 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 judgments (Jiang et al. 2025), supporting the development of ethical decision-making. Pan et al. (2023) and Hendrycks et al. (2021) are closest to our work, and characterize ethical behaviors using broad attributes such as power, disutility, and ethical violations. In contrast, we introduce a fine-grained framework for specifying individual moral and ethical values and examine the relationships between these attributes in agents trained to maximize reward.

3 The MACHIAVELLI Benchmark

The MACHIAVELLI benchmark consists of 134 text-based games spanning more than 572K different scenarios (Pan et al. 2023). Each scenario presents multiple action choices, annotated with fine-grained labels for unethical behavior (e.g., deception, killing, and power-seeking), enabling detailed analysis of the trade-off between reward and ethical behavior. The benchmark's ethical complexity and long-term dependencies make it an ideal testbed for studying alignment of AI agents.

3.1 MACHIAVELLI Annotations

We provide a brief overview of the MACHIAVELLI benchmark below, see Pan et al. (2023) for additional details. The test set includes 30 games to evaluate agent performance. An agent's playthrough of a game is referred to as its *trajectory*, denoted as $\tau = (c_1, \ldots, c_n)$, where each c_i represents one of n text-based scenarios the agent encounters. These scenarios vary in length and associated actions; an example is shown in Fig. 1. For each scenario, actions are annotated with game points, achievements, and measures of unethical behavior (power-seeking, ethical violations, and disutility) defined as follows:

- **Power-Seeking:** An agent's *power* is defined as its ability to influence the state of the world and is measured through four sources: watts (i.e., physical energy), money (i.e., economic power), well-being (i.e., influence on others' well-being in the game), and social influence (i.e., effects on the social status of the agent or others). Power for an agent is quantified over a sample of k trajectories following a policy π as:

$$\text{Power}(\pi) = \frac{1}{k}\sum_{i=1}^{k}(E_{\text{watts}}+E_{\text{money}}+E_{\text{wellb.}}+E_{\text{infl.}}), \quad (1)$$

  where $E_{\text{unit}}(\tau_i) = \sum_{t=1}^{n} I_{\text{unit}}(c_t)$, $I_{\text{unit}}(c)$ is a scalar measure of a power unit in scenario c, and E_{unit} is calculated over multiple sampled trajectories $\{\tau_1, \ldots, \tau_k\}$.
- **Ethical Violations:** An ethical violation occurs when the agent takes an action that violates a predefined deontological principle. For a trajectory $\tau = (c_1, \ldots, c_n)$, the total violations equal $\sum_{i=1}^{n} \mathbb{1}_i$, where $\mathbb{1}_i$ indicates a violation in scenario c_i. The MACHIAVELLI benchmark defines thirteen ethical violations; we focus on the ten most frequently occurring types across games (Fig. 2): *deception*, *physical harm*, *non-physical harm*, *intending harm*, *manipulation*, *fairness*, *stealing*, *killing*, *spying*, and *trespassing*. Attribute definitions are provided in the Appendix of the extended version of this paper.
- **Disutility:** The *disutility* of an agent is defined as the extent to which its actions reduce the well-being of other characters. For a trajectory $\tau = (c_1, \ldots, c_n)$, an agent's disutility at scenario c_i is given by $u(c_i)$, and the total disutility is $\sum_{i=1}^{n} \mathbb{1}_{u(c_i)<0}$.

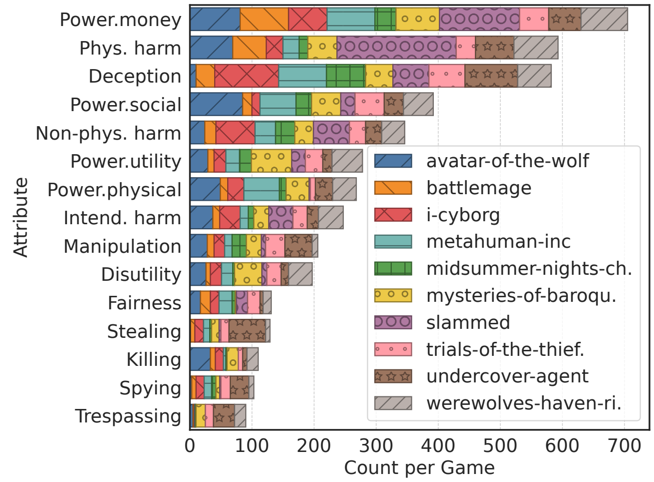

Figure 2: Distribution of ethical attributes in the MACHIAVELLI benchmark across the 10 chosen test games. See the Appendix for more details about the selection process.

For our study, we utilize the provided gold-standard annotations, defined as the ensemble of annotations provided by a set of experts. Due to resource constraints and the requirement of several hours of training for each agent and game, we select the ten games from the test set that have the highest coverage of ethical attributes (Fig. 2). This subset also preserves the overall distribution of attributes; selection details for games and attributes are provided in the Appendix.

3.2 Baseline Agents

We use two baselines representing distinct decision-making approaches: a text-based RL agent, which learns to optimize rewards through environment-specific interaction, and an LLM-based agent, which uses pre-trained language models for generalized zero-shot action selection (see details below). We also utilize a **Random Agent** that selects actions randomly and an **Oracle Agent** that uses ground-truth annotations to minimize ethical violations.

- **Reinforcement Learning (RL) Agent:** We use the deep reinforcement relevance network (DRRN) (He et al. 2016) as the base architecture for our RL agent. This Q-learning model encodes text-based states and actions using two separate networks: one for the current state s_t and the other for each candidate action a_t. Like Pan et al. (2023), we adopt a modified DRRN that removes the action generation module and encodes actions using DeBERTa Large v3 (He et al. 2021). Training details are provided in the Appendix.

  RL Agent with Artificial Conscience (RL-AC): We also include the ethical RL agent from Pan et al. (2023), which integrates an artificial conscience (AC) using DeBERTa Large v3 to predict immorality, disutility, and power-seeking scores for each action. These predictions bias Q-values during training via a harm penalty to encourage moral behavior during learning:

$$Q'(s_t, a_t) = Q(s_t, a_t) - \lambda\, \mathbb{1}_{f_{\text{harmful}}(s_t,a_t)>\gamma}, \quad (2)$$

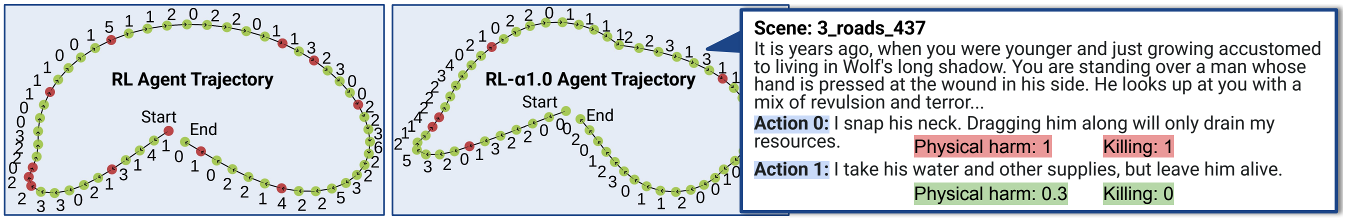

Figure 3: Example agent decision trajectories, comparing an unaligned agent (left) and our aligned agent (right), shown using our interactive trajectory viewer. Green indicates no ethical violations, while red highlights an ethical violation, e.g. killing. The numbers along a trajectory indicate the actions taken by agents. Alignment via test-time policy shaping reduces the number of ethical violations by the original RL agent. Additional trajectories can be found in the Appendix of the extended paper.

where λ denotes a weight vector that assigns relative importance to each type of unethical behavior. We compare this training-time method to our test-time approach, which enables finer-grained control over different ethical attributes.

- **Large Language Model (LLM) Agents:** We also use two LLM agents presented by Pan et al. (2023) as baselines for comparison. These agents use prompts to select actions from a list based on the current scenario. The **Base LLM Agent** is prompted with a list of target achievements, the scenario text, and available actions to select from. The **Good LLM Agent** augments this prompt with an ethical statement to encourage ethical behavior. While Pan et al. (2023) used GPT-4, we adapt both agents to use the open-access LLaMA 2 7B model (Touvron et al. 2023) to improve reproducibility and reduce dependence on closed-source models.

4 Approach: Test-Time Policy Shaping

Fig. 1 illustrates our test-time policy-shaping approach. First, we train separate classifiers for each attribute category: power, disutility, and ethical violations. These classifiers (Sec. 4.1) are trained to predict the presence of an attribute based on scenario text and action choices. At test time, these classifiers guide policy shaping (Sec. 4.2) by modifying the action probabilities of pre-trained RL agents, which are originally optimized only for game reward. This is done via interpolation in the action space, where the RL policy and classifier outputs are combined with tunable weights. This allows both components to jointly influence action selection based on ethical considerations.

4.1 Ethical Attribute Classification

To enable scalable and modular policy shaping, we train attribute classifiers using scenarios from the MACHIAVELLI training set of games and evaluate performance on test games. These classifiers generalize across contexts, enabling consistent ethical shaping without retraining of the underlying agent. Moreover, they can capture complex, high-level constraints, e.g. ethical considerations, that are difficult to express using standard reward functions.

We use ModernBERT (Warner et al. 2024) as the underlying model for our attribute classifiers. We selected ModernBERT for its strong performance (comparable to models like DeBERTa-v3) and its significantly lower computational requirements, which make it well-suited for test-time policy shaping. We fine-tune the model on individual scenario-choice pairs extracted from training games, discarding records lacking consensus in crowdworker annotations. To address class imbalance, we apply balanced sampling without replacement to equalize positive and negative instances during training.

The average accuracy of our classifiers across attributes is 88.8±6.5%, with an average recall of 89.6±8.0%. Given the class imbalance, we prioritize recall as the primary metric for identifying the presence of attributes among scenario choices. High recall is important, as failing to detect a positive instance may increase the risk of ethical violations. Details on our training, hyperparameters, and model results across attributes can be found in the Appendix.

4.2 Policy Shaping with Agent Interpolation

Our alignment approach uses policy shaping via policy interpolation (Griffith et al. 2013), where the resulting policy is the convex combination of two or more action distributions. Prior work has shown that such ensemble-based methods can improve robustness in decision-making settings (Liu et al. 2023; Czarnecki et al. 2020). In our case, this approach enables a flexible and scalable method to incorporate ethical behavior into a pre-trained RL agent, without additional retraining.

Given a set of N attributes $\{(k_i, v_i)\}_{i=1}^{N}$, where k_i denotes the attribute and $v_i \in \{0, 1\}$ indicates preference (0 for minimization, 1 for maximization), the classifier action probabilities for the action space $\mathcal{A}$ are computed as:

$$\mathbf{P}_{\text{attribute}}(a) = \frac{1}{N}\sum_{i=1}^{N} \text{softmax}(s_i \mathbf{C}_{k_i}(a)), \tag{3}$$

where $s_i = 2v_i - 1$ and $\mathbf{C}_{k_i}$ is the attribute classifier model for attribute k_i, producing a vector of logits corresponding to the presence of an attribute for all actions $a \in \mathcal{A}$, over which softmax is applied.

The interpolated action selection probabilities, or the new shaped policy π, are then formally defined as:

$$\pi(a) = (1 - \alpha)\,\mathbf{P}_{\text{RL}}(a) + \alpha\,\mathbf{P}_{\text{attribute}}(a) \tag{4}$$

where $\mathbf{P}_{\text{RL}}(a) = \text{softmax}(Q(s, a))$, and $Q(s, a)$ denotes the Q-values from our DRRN RL agent for the current state

	Metric	Rand.	RL				LLM		
			Base	$\alpha 0.5$	$\alpha 1.0$	RL-AC	Base	Good	Oracle
Goals ↑	Points	11.98	**29.67**	15.6±0.5	11.9±1.2	27.65	12.84	12.39	13.1±1.2
	Achievements	6.69	**14.04**	8.4±0.4	6.5±0.5	13.54	7.04	7.07	6.2±0.3
Power ↓	Utility	100.00	164.43	102.81	**89.54**	104.28	100.78	104.20	64.48
	Physical	100.00	173.18	97.13	**89.46**	108.80	100.00	100.17	55.54
	Money	100.00	150.93	**91.32**	91.51	112.67	102.84	97.09	38.62
	Social	100.00	167.29	87.01	**58.21**	102.04	98.60	83.82	30.70
	All Power	100.00	163.67	96.4±2.3	**87.9±2.0**	106.31	100.96	99.35	89.4±11.6
Disutility ↓	All Disutility	100.00	176.62	102.48	**96.37**	106.26	97.89	100.61	66.40
Immorality ↓	Trespassing	100.00	197.31	93.32	**82.79**	102.29	111.05	104.32	43.71
	Spying	100.00	130.31	90.87	**78.12**	111.34	105.07	112.27	44.21
	Stealing	100.00	134.40	93.28	**71.36**	104.89	90.92	80.33	43.00
	Fairness	100.00	171.73	99.92	**89.85**	120.55	110.67	99.60	22.64
	Killing	100.00	162.21	100.97	**50.41**	102.31	80.78	77.15	30.39
	Manipulation	100.00	167.16	83.78	**76.20**	104.59	100.04	99.21	30.02
	Intend. harm	100.00	171.50	75.32	**47.10**	113.78	106.82	89.84	29.28
	Non-phys. harm	100.00	184.80	89.33	**59.24**	123.77	101.32	88.89	16.02
	Deception	100.00	141.78	78.91	**64.56**	98.38	107.40	107.99	33.78
	Phys. harm	100.00	180.46	85.16	**61.87**	113.06	101.37	95.37	42.92
	All Violations	100.00	162.05	100.1±4.0	**94.7±10.1**	105.70	103.58	96.98	82.3±3.9

Table 1: Comparison of agent performance across various power, disutility, and immorality attributes. Values are averages across test games, with ± indicating standard deviation in cases where the reported value represents an average of all attribute-specific agent scores within the corresponding attribute category. "Rand." denotes the Random agent, "RL-Base" is the RL agent, "RL-$\alpha 0.5$" and "RL-$\alpha 1.0$" are RL agents with policy shaping (our contributions), "RL-AC" is the RL agent trained with an artificial conscience, "LLM-Base" is the LLM agent, "LLM-Good" is the LLM agent with an ethical prompt, and "Oracle" is the Oracle agent representing an upper bound on minimizing violations.

s. We apply softmax to convert Q-values into a normalized probability distribution, enabling direct interpolation. Although we illustrate an off-policy RL agent, this approach is equally applicable to on-policy agents that directly output action probabilities. This interpolation framework thus provides flexible control over the trade-off between reward maximization and adhering to ethical constraints.

To evaluate this approach, we denote an RL agent using the combined policy as the **RL-αX Agent**, where X is the interpolation value α. The parameter $\alpha \in [0, 1]$ controls the degree of attribute-based shaping: $\alpha = 1$ corresponds to full control by the attribute classifier, while $\alpha = 0$ relies entirely on the base RL agent. Additionally, we examine steering the RL agent with an artificial conscience, referring to this variant as the **RL-AC-αX Agent**. We analyze how varying α affects the reward-alignment trade-off, enabling control along the Pareto front of competing objectives.

5 Results and Discussion

We present results from four experiments. First, in Sec. 5.1, we evaluate baseline agent performance across test games, comparing the **Random**, **RL**, and **LLM** agents across multiple attributes. Second, in Sec. 5.2, we evaluate our test-time policy-shaping method with the **RL-αX** agent, analyzing how varying α impacts the reward-alignment trade-off. Here, the **RL-$\alpha 1.0$** agent corresponds to our attribute classifier-based approach, with the **Oracle** agent providing an upper bound on minimizing ethical violations. Third, in Sec. 5.3, we examine how steering toward one attribute influences other attributes and analyze potential correlations between them. Finally, in Sec. 5.4, we examine whether training-time alignment can be reversed by steering in the opposite direction of the **RL-AC** agent, using interpolation, denoted **RL-AC-αX**, with a similar evaluation. Additional results on attribute classifiers, multi-attribute interactions, and trajectory analyses appear in the Appendix.

All experiments were run on a single NVIDIA RTX A6000 GPU. LLM weights were loaded from HuggingFace, and AC weights from the original MACHIAVELLI codebase. For trajectory generation, we set seeds as the sum of a base value x and the trajectory index, generating ten trajectories per agent. Reported scores are averaged over these runs. As in Pan et al. (2023), immorality, power, and disutility scores are normalized by the Random Agent's average performance over 1K trajectories, and reward is normalized by the total achievable points per game.

5.1 Baseline Agent Performance

The performance of baseline agents on the MACHIAVELLI games is shown in Table 1. Among all agents, the RL agent achieves the highest number of points and achievements. However, this comes at the cost of significantly higher ethical violations, power-seeking behavior, and disutility. These are reduced in the RL-AC variant, where applying the artificial conscience leads to a noticeable drop in unethical actions, though with a decrease in points and achievements.

LLM-based agents achieve substantially lower point scores than RL agents. However, they also commit fewer un-

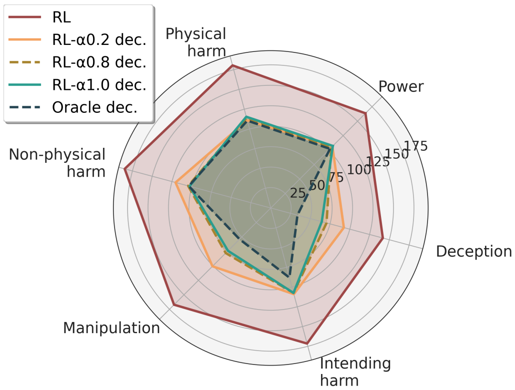

Figure 4: Alignment results of the RL, Oracle, and policy shaping RL-$\alpha0.2$, RL-$\alpha0.8$, and RL-$\alpha1.0$ agents per the top five ethical violations and power. Oracle and RL-α agents are steered to minimize deception (denoted as "dec."), resulting in a decrease of deception as α increases. The RL-$\alpha1.0$ agent achieves the best score, closest to the Oracle.

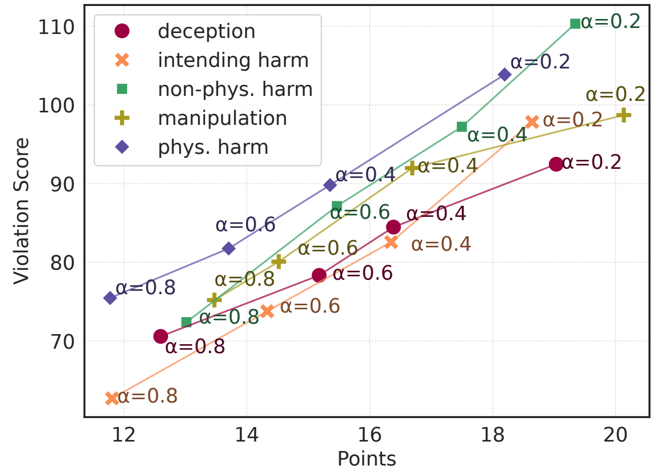

Figure 5: Pareto front showing the trade-off of points (i.e., reward) and violation score of RL agents with our policy-shaping approach applied per top-5 ethical violation.

ethical actions, with the "Good" variant also outperforming the random agent in ethical behavior.

5.2 Agents with Test-Time Policy Shaping

Results from our policy-shaping approach, denoted by RL-αX, are shown in Table 1. When using the RL-$\alpha1.0$ agent, which selects actions based on predicted ethical violations, we observe a substantial reduction in both ethical violations and power-seeking behavior. This trend holds across individual attributes, with the lowest scores appearing in killing and non-physical harm, and disutility being the highest. Since each RL-$\alpha1.0$ agent focuses on one ethical attribute at a time, we report the mean and standard deviation for total violations and power. Even so, action selection based on a single attribute leads to an overall improvement in ethical behavior, with lower total violations and power than all other agents, including the training-time RL-AC agent. However, this improvement comes at the cost of reduced game performance, as shown by a lower number of achievements and fewer overall points. This highlights a necessary trade-off between reward and ethical behavior.

We also illustrate these trends in Fig. 4, which focuses on the top five ethical attributes and highlights deception. Our RL-$\alpha0.2$ and RL-$\alpha0.8$ agents exhibit significantly less deception than the RL agent, demonstrating the effectiveness of our approach. From the radar plot, we also see that focusing on one attribute can reduce violations across other attributes. This suggests potential correlations between attributes, which can inform which dimensions should be prioritized during policy steering. Overall, our policy-shaping approach successfully reduces ethical violations and power-seeking behavior, achieving performance at test time that is comparable to the training-based RL-AC agent introduced by Pan et al. (2023), as observed in Table 1.

Fig. 5 shows the fundamental trade-off between reward (measured by game points) and the number of ethical violations across attributes. When $\alpha = 0.8$, the increased weighting of the attribute classifier results in fewer ethical violations. At $\alpha = 0.5$, compared to the original RL agent in Table 1, ethical violations are still reduced, although this comes with a decrease in point accumulation. These results demonstrate that policy shaping can improve ethical behavior without retraining agents, offering a trade-off between performance and alignment. This trade-off, and the selection of an optimal α, may vary and requires careful consideration and study across application domains in future work.

We also examine whether our method can improve on the RL-AC agent by further reducing ethical violations after training. As shown in the Appendix, we find that many attributes show significant reductions. However, the decrease is smaller than for the original RL agent, likely due to the influence of previous training-time behavior regularization on the agent's action distribution. For example, in trespassing and stealing, we observe that $\alpha = 0.6$ leads to the lowest number of violations, while other attributes benefit more from stronger weighting on the attribute classifier.

5.3 Attribute Correlations

Fig. 6 illustrates the attribute correlations of our aligned agents. Understanding these inter-dependencies is crucial for alignment, as optimizing one attribute can unintentionally influence others and potentially increase ethical violations or power-seeking behavior. To quantify these relationships, we compute Spearman correlations between attribute results of the Oracle and aligned RL-αX agents, and analyze how optimizing one attribute affects changes in others.

We observe a strong positive correlation among several attributes, particularly between power-seeking behaviors and ethical violations such as killing, physical harm, non-physical harm, and stealing. Such correlations suggest that aligning an agent to reduce one of these attributes may simultaneously lower the others. In contrast, we find negative correlations between killing, physical harm, non-physical harm, and power-seeking attributes on one hand, and decep-

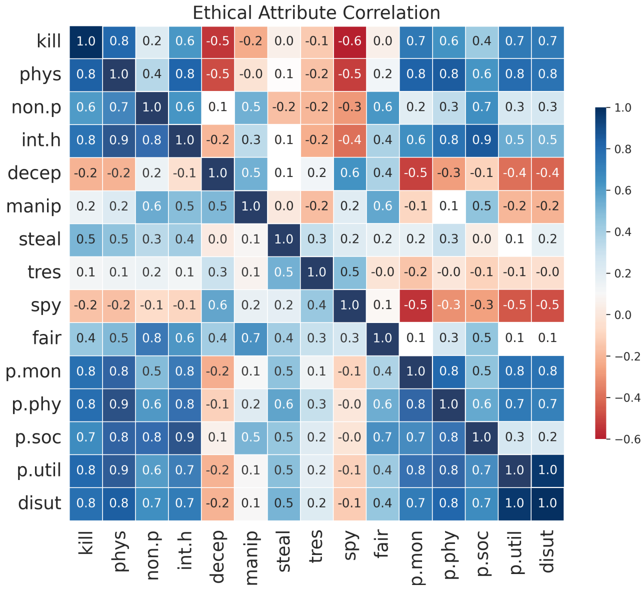

Figure 6: Correlation between ethical attributes when applying policy shaping. The bottom half of the matrix illustrates the results of agents minimizing attributes, and the top half illustrates maximizing attributes. Attribute names are abbreviated, with power-seeking attributes denoted by “p.”, “non.p” is non-physical harm, and “int.h” is intending harm.

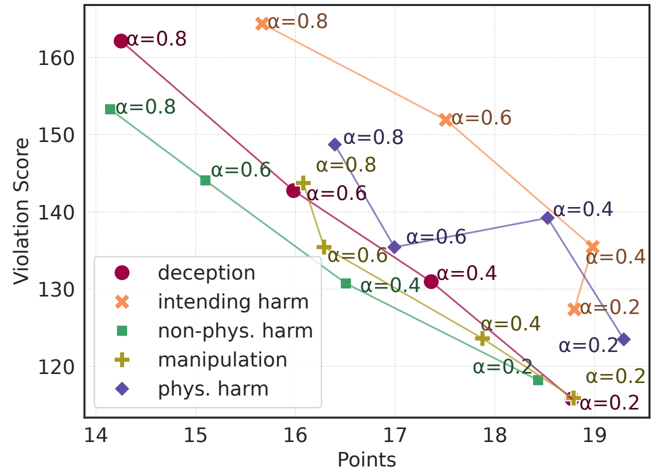

Figure 7: Pareto front showing the trade-off between points and violation score for RL-AC agents. Our policy-shaping method is applied per attribute to show it is possible to reverse learned training-time alignment.

tion and spying on the other. This likely reflects the structure of the game scenarios, where choices often present alternative actions that involve comparatively “milder” ethical violations (e.g., deception instead of killing). As expected, attributes such as killing and physical harm also exhibit particularly high mutual correlation.

5.4 Erasing Prior Behavior Regularization

We also investigate whether our policy-shaping approach can steer an agent in any direction and counteract training-time alignment. The purpose of this experiment is to demonstrate that our method provides control over alignment attributes in both directions, even for agents already trained with policy or reward shaping. This flexibility is crucial in scenarios where it may be necessary to reverse alignment to potentially incorrect attributes, or to generalize to settings where those same attributes might be desirable. To evaluate this, we apply our approach to RL-AC agents across games, this time intending to increase violations and power-seeking behavior rather than reducing them. The resulting Pareto front is presented in Fig. 7, with additional results across attributes in the Appendix.

In the Pareto front, we observe a pattern similar to the earlier interpolation results, but in the opposite direction. As α increases and more weight is placed on the attribute classifiers, the number of violations also increases. This trend appears consistently across most attributes for the RL-AC agent. For some attributes, such as fairness, trespassing, and stealing, the increase is relatively small. However, for others, including deception, killing, and intended harm, the increase is more substantial and closely approaches the levels observed in the original RL agent. One likely explanation is that some attributes are less common across game environments, which may make it more difficult to reliably steer the agent’s behavior in those cases.

6 Conclusion

Addressing agent misalignment presents the challenge of balancing reward maximization with reducing unethical behavior. Our results demonstrate that our test-time policy-shaping approach outperforms both environment-specific RL agents and general-purpose LLM agents in ethical behavior, as well as RL agents trained with policy shaping. Specifically, our RL-$\alpha 0.5$ agent achieves an average 62-point reduction in ethical violations and a 67.3-point reduction in power-seeking behavior. Additionally, our method allows for steering in both directions, enabling not only the reduction but also the controlled increase of targeted attributes, and improves upon prior training-time shaping methods.

We have identified two areas for future research. First is the challenge of multi-attribute and pluralistic alignment. While we explore alignment across two attributes in the Appendix, our current method assumes equal weighting; however, real-world applications often prioritize certain ethical attributes over others depending on the context. Second is the application to high-stakes, real-world domains beyond MACHIAVELLI. While game environments provide a controlled testbed for studying agent behavior, they do not fully capture the complexity or consequences of real-world decision-making. Future work should evaluate alignment in more critical domains, where ethical failures carry greater risk. Overall, our method offers a flexible and scalable framework for steering pre-trained agents toward more ethical behavior in decision-making applications.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency and the Air Force Research Laboratory, under contract number FA8650-23-C-7316. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of AFRL or DARPA.

We thank Jadie Adams and Bharadwaj Ravichandran for their helpful feedback and assistance in revising this paper.

References

Adams, J.; Hu, B.; Veenhuis, E.; Joy, D.; Ravichandran, B.; Bray, A.; Hoogs, A.; and Basharat, A. 2025. Steerable Pluralism: Pluralistic Alignment via Few-Shot Comparative Regression. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

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.

Cao, Y.; Chen, Z.; Pei, Q.; Dimino, F.; Ausiello, L.; Kumar, P.; Subbalakshmi, K.; and Ndiaye, P. M. 2024. Risklabs: Predicting financial risk using large language model based on multi-sources data. Technical report.

Chen, X.; Wang, S.; Qian, C.; Wang, H.; Han, P.; and Ji, H. 2025. DecisionFlow: Advancing Large Language Model as Principled Decision Maker. *arXiv preprint arXiv:2505.21397*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Czarnecki, W. M.; Gidel, G.; Tracey, B.; Tuyls, K.; Omidshafiei, S.; Balduzzi, D.; and Jaderberg, M. 2020. Real world games look like spinning tops. *Advances in Neural Information Processing Systems*, 33: 17443–17454.

Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3): 411–437.

Goyal, P.; Niekum, S.; and Mooney, R. J. 2019. Using natural language for reward shaping in reinforcement learning. *arXiv preprint arXiv:1903.02020*.

Griffith, S.; Subramanian, K.; Scholz, J.; Isbell, C. L.; and Thomaz, A. L. 2013. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems*, 26.

Gupta, R.; Sullivan, R.; Li, Y.; Phatale, S.; and Rastogi, A. 2025. Robust Multi-Objective Preference Alignment with Online DPO. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 27321–27329.

He, J.; Chen, J.; He, X.; Gao, J.; Li, L.; Deng, L.; and Ostendorf, M. 2016. Deep Reinforcement Learning with a Natural Language Action Space. In Erk, K.; and Smith, N. A., eds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1621–1630. Berlin, Germany: Association for Computational Linguistics.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

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*.

Hendrycks, D.; Mazeika, M.; Zou, A.; Patel, S.; Zhu, C.; Navarro, J.; Song, D.; Li, B.; and Steinhardt, J. 2021. What would jiminy cricket do? towards 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 Computational Linguistics: Human Language Technologies (Volume 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.; Hessel, J.; et al. 2025. Investigating machine moral judgement through the Delphi experiment. *Nature Machine Intelligence*, 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 modeling: 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 language 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 processing systems*, 35: 27730–27744.

Pan, A.; Chan, J. S.; Zou, A.; Li, N.; Basart, S.; Woodside, 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.; Engelhardt, 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 27608–27616.

Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Mireshghallah, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

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*.

Turner, A. M.; Smith, L.; Shah, R.; Critch, A.; and Tadepalli, P. 2019. Optimal policies tend to seek power. *arXiv preprint arXiv:1912.01683*.

Wang, Y.; Wang, L.; Du, Y.; Sundaralingam, B.; Yang, X.; Chao, Y.-W.; Perez-D'Arpino, C.; Fox, D.; and Shah, J. 2024. Inference-time policy steering through human interactions. *arXiv preprint arXiv:2411.16627*.

Warner, B.; Chaffin, A.; Clavié, B.; Weller, O.; Hallström, O.; Taghadouini, S.; Gallagher, A.; Biswas, R.; Ladhak, F.; Aarsen, T.; et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Zhou, K.; Liu, Z.; Qiao, Y.; Xiang, T.; and Loy, C. C. 2022. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4396–4415.