Patronus - Value Alignment of RL Agents in Text-Based Games Using MFT Profiles

Sankalp Sagar
TU Delft, NL

Abstract

AI agents optimized purely for task completion often inadvertently violate human moral expectations. This is especially pronounced in reinforcement learning agents operating in text-based environments, where the richness of language and the vast action space allow for numerous harmful yet reward-maximizing paths. Existing approaches to moral alignment commonly represent morality as a single scalar signal, limiting both interpretability and the ability to model diverse moral preferences. We introduce MFT Patronus, a policy-shaping framework based on Moral Foundations Theory that represents morality across five dimensions: Care, Fairness, Loyalty, Authority, and Sanctity. This multidimensional representation enables configurable moral profiles that capture different foundational priorities. We evaluate our approach on the Jiminy Cricket benchmark and show that it maintains task performance while substantially reducing immoral behavior compared to the standard baselines and maintaining a net positive balance of moral over immoral actions. Our results further demonstrate that different moral profiles produce distinct behavioral patterns, suggesting that multidimensional moral representations are a promising direction for interpretable and configurable value alignment in reinforcement learning agents.

1 Introduction

Reinforcement learning (RL) agents are increasingly being deployed in rich, language-driven environments. While these agents can learn to maximize task-oriented rewards, their behavior often diverges from human moral expectation. For example OpenAI (2016) trained an agent on the game Coast Runners, where the objective is to beat other boats in a race by using game score as a reward mechanism. However, instead of completing a racing track normally, the agent learned to drive in circles and crash into other boats and objects while collecting bonus points indefinitely. This type of behavior is called the value alignment problem (Hendrycks et al., 2021a; Gabriel, 2020). Text-based games provide a particularly challenging testbed because actions are expressed in natural language, and rewards are typically sparse and tied to game progress rather than ethical constraints.

Prior work has attempted to instill moral behavior in (text based games) RL agents by treating each action on a binary scale that ranges between right or wrong and using the values from the scale to shape action selection (Hendrycks et al., 2021b; Ammanabrolu et al., 2022). However, these approaches treat morality as a monolithic dimension, ignoring its pluralistic nature. As a result, they cannot distinguish why an action is considered immoral—whether it causes harm, betrays loyalty, violates authority, or transgresses other distinct moral concerns. Moreover, they offer no way to configure an agent's behavior according to personalized moral preferences, for example valuing fairness more than loyalty.

To address these limitations, we propose a novel policy-shaping framework that leverages Moral Foundations Theory (MFT) (Haidt and Joseph, 2004). MFT decomposes morality into five virtue-vice pairs: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Sanctity/Degradation. Instead of a single immoral score, our framework called MFT Patronus computes a multi-dimensional judgment of morally charged states and then applies a weight vector over the five foundations to compute foundation-specific penalties. By changing the weights, the agent can learn to exhibit different moral behavior.

The paper makes the following contributions:

- **A multidimensional moral shaping framework for alignment in text-based RL.** We propose MFT Patronus, which leverages Moral Foundations Theory representations, allowing reinforcement learning agents to be

shaped by multiple moral dimensions.

- **A new MFT-labeled version of the Jiminy Cricket benchmark.** We develop an Large Language Model (LLM)-based judgment pipeline that assigns foundation-level moral labels to existing moral annotations and analyze the reliability and limitations of these judgments.
- **An empirical study of configurable moral behavior in RL agents.** We demonstrate that foundation-specific reward shaping can improve moral outcomes while preserving task performance, and we analyze how different moral profiles affect agent behavior.

The remainder of this paper is organised as follows. Section 2 provides background on Moral Foundations Theory, RL for text-based games, and LLMs as annotators. Section 3 describes the Jiminy Cricket environment and the original setup. Section 4 details the MFT Patronus methodology, including MFT judgments, Q-value shaping and moral profile definitions. Section 5 presents the experimental setup, baselines, different moral profiles of the agents and the evaluation metrics. Section 6 reports and discusses the results, including the LLM annotation quality, baseline comparisons, moral profile analysis, and a Pareto frontier. Section 7 concludes the paper and outlines future work. Finally, we examine the limitations and ethical considerations of our work in Section 8 and Section 9 respectively. The code for the entire paper is present at the repository here.

2 Background

2.1 Morality in NLP

Human beings have an inherent intuition about what is right and wrong, which makes them consider a pattern of behavior as moral or immoral (Haidt, 2001). Social and cultural psychologists generally agree that morality is not a single, monolithic dimension (Schwartz, 1992). To better understand and analyze this intuitive moral sense, researchers often break it down into several basic components, known as moral foundations (Graham et al., 2013). According to MFT (Haidt and Joseph, 2004), there are five such foundations expressed as virtue/vice pairs: Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Sanctity/Degradation. The first two are considered individualizing foundations while the latter three are called binding foundations (Graham and Haidt, 2010). MFT has proven to be extremely useful, as it can successfully predict and examine group and cultural differences e.g. the difference between the moral profile of conservatives versus liberals (Koleva et al., 2012; Graham et al., 2011; Doğruyol et al., 2019; Graham et al., 2009). Building on these results, MFT has also shown to have more predictive power for moral judgements over other binary moral frameworks (Nilsson and Erlandsson, 2015). While there have been some criticisms regarding the foundations themselves (Atari et al., 2023) and the generalizability of MFT results (Davis et al., 2016), MFT has continued to evolve and be refined (Atari et al., 2023) and has cemented itself as a leading and resilient moral framework.

Due to rapid advancements in Natural Language Processing (NLP) field, there have been significant developments in estimating moral foundations from text information including cross-domain generalizability (Preniqi et al., 2024; Zangari et al., 2025a; Liscio et al., 2022). Another significant development has been the creation of large scale datasets with morality annotations (Trager et al., 2022; Hoover et al., 2020). Due to these developments, there has been growing use of MFT in NLP to enable a diverse range of applications ranging from political stance (Simmons, 2023; Rao et al., 2023), morally aligned argumentation (Alshomary et al., 2022) and much more (Zangari et al., 2025b; Vida et al., 2023)

2.2 RL and Morality

In RL an agent is trained to maximize reward signals that encode desirable behavior within an environment (Littman, 2015). In practice, these reward functions often provide only proxy approximations of the intended objective, allowing agents to exploit unintended shortcuts or loopholes in the optimization process (Skalse et al., 2022), which can result in behavior that is optimal with respect to the reward function but misaligned with human moral expectations. This causes a value alignment problem, which is considered one of the biggest challenges in ML safety (Hendrycks et al., 2021a).

One popular and challenging domain for studying alignment in RL is text-based games. Text-based games are commonly modeled as Partially Observable Markov Decision Processes (POMDPs) (Adhikari et al., 2020), defined by the tuple (S, A, T, O, R). In these environments, the underlying game state S contains latent information

about the environment. Instead of directly accessing the full environment state, the agent receives a textual observation O describing the current scene.

At each timestep, the agent selects a textual action A, such as interacting with objects, navigating between locations, or communicating with in-game characters. The environment transition function $T : S \times A \rightarrow S'$ (Hausknecht et al., 2020) updates the hidden game state according to the chosen action, producing a new observation and reward signal. Rewards are typically sparse and tied to specific progress-related conditions and the end-goal is typically not clear from the start.

Although there are several testbeds for text based games like Shridhar et al. (2020) and Hausknecht et al. (2020); they are typically used to evaluate only the game performance of the agents and do not analyze agent behaviors. This can result in agents learning to maximize game rewards while acting in a morally undesirable way. More recently, Pan et al. (2023) and Hendrycks et al. (2021b) have come up with testbeds that also examine the moral behavior of agents, the latter acting as the foundation of our paper as elaborated in Section 3.1.

There have been several approaches taken to train an RL agent to play text based games ethically. Hendrycks et al. (2021b) employ a RoBERTa-based classifier to estimate the immorality of candidate actions, which is then used to shape the agent's Q-values (Griffith et al., 2013). Building on this framework, Ammanabrolu et al. (2022) incorporate social commonsense norms encoded in Delphi (Jiang et al., 2021) to constrain the agent's action space, thereby discouraging unethical behavior. A common limitation of these works is that they treat morality as a monolithic variable, without mapping actions to distinct moral foundations. This obscures which specific foundation underlies a given moral or immoral judgment, limiting interpretability and value alignment.

2.3 LLMs as Annotators

Manually annotating text for moral foundations is time-consuming, expensive, and subject to annotator bias. To address these challenges, researchers have increasingly turned to LLMs as zero-shot or few-shot annotators of moral content (Skorski and Landowska, 2025; Bulla et al., 2025). LLMs possess rich world knowledge and reasoning capabilities (Kwon and Michael, 2023), making them an attractive alternative to human annotators. However, they are not without limitations. LLMs have been shown to default to WEIRD (Western, Educated, Industrialized, Rich, Democratic) values (Zewail et al., 2026; Muthukrishna et al., 2020), which can introduce systematic labeling biases. Furthermore, LLMs exhibit moral hypocrisy (Nunes et al., 2024), leading to inconsistency when labeling moral foundations across similar contexts. Despite these limitations, LLMs offer a scalable and reproducible alternative for moral annotation, enabling large-scale annotations that would have been difficult with human annotators alone.

Figure 1: Excerpt from the game Cutthroats in Jiminy Cricket.

3 The Jiminy Cricket Framework

3.1 Dataset

We employ the Jiminy Cricket benchmark for our work as the suite features 25 complex text-based adventure games spanning over 2000 locations and nearly 5000 objects over a multitude of steps (See Figure 1).

In the original annotation framework (Hendrycks et al., 2021b), an action is labeled immoral if it is considered bad in a *pro tanto* sense, and is characterized by three attributes: valence (positive or negative), focal point (agent or environment), and an ordinal severity score in {1, 2, 3}. However, prior work such as Ammanabrolu et al. (2022) has observed that the focal point and ordinal severity annotations can be inconsistent and highly dependent on individual annotators. Consequently, we omit these dimensions from our analysis and instead focus on developing a more structured representation of moral behavior using MFT.

3.2 CALM Action Generator

One of the primary challenges in text-based reinforcement learning environments is the extremely large action space available to agents. Since actions are represented through natural language commands, agents must reason not only about which action is optimal, but also about how to formulate valid textual commands. Naively enumerating all possible verb-object combinations leads to a combinatorially large action space, significantly hindering efficient exploration and learning.

Hendrycks et al. (2021b) utilize the Contextual Action Language Model (CALM) (Yao et al., 2020) to restrict the candidate actions space. CALM is a language-model-based action generator trained on human gameplay transcripts collected from a large corpus of interactive fiction games. Given the current textual observation and game context, the model generates a compact set of plausible actions that are likely to be relevant within the current environment state.

Formally, given a textual observation o_t, CALM generates a candidate action set:

$$A_t = \{a_1, a_2, \ldots, a_n\}$$

where each candidate action is conditioned on the current game context and prior state and action.

3.3 Agent Implementation

Hendrycks et al. (2021b) use the Deep Reinforcement Relevance Network (DRRN) (He et al., 2016) which has been successively adopted in other papers working with Jiminy Cricket (Ammanabrolu et al., 2022; Lim et al., 2025). The architecture employs separate GRU-based encoders for the game observation, environment description, inventory, and candidate actions generated by CALM. The encoded state representations are concatenated into a unified state embedding, while candidate actions are encoded independently. A relevance scoring network then estimates the Q-value for each state-action pair:

$$Q(s_t, a_t) = f(\phi_s(s_t), \phi_a(a_t))$$

where ϕ_s and ϕ_a denote the learned state and action encoders respectively. This Q-value is used to probabilistically select the best action in terms of expected reward. This is called deep Q-learning (Watkins and Dayan, 1992).

3.4 Oracle

One of the policy shaping approaches taken by Hendrycks et al. (2021b) is the Oracle method, where they used the human annotated scalar morality judgments present in the Jiminy Cricket environment to guide the decision making of the agent. In this method, the Q-values of the candidate actions are shaped by the immorality scores of the successive states. The strength of the conditioning is controlled by a scalar weight γ.

4 MFT Patronus

4.1 Moral Judgment Labeling

As mentioned in Section 3.1, Jiminy Cricket annotations are monolithic and do not express morality in pluralistic sense, limiting their interpretative and fine tuning power. In order to label the pluralistic values associated with the morally charged game states, we leverage an LLM. We explicitly define each moral foundation and ask the LLM to reason about the MFT scores to be assigned to the state. These scores are used downstream to shape the agent's behavior as we shall explain further in Section 4.2.

Therefore we obtain the MFT vector for relevant game states defined as

$$m(a_t) = [c, h, f, ch, l, b, a, s, sanc, d]$$

where each dimension corresponds to one moral foundation and takes values in $\{0, 1\}$ representing neutral or positive moral foundation alignment. Note that as mentioned in Section 2.1, (h, ch, b, s, d) or Harm, Cheating, Betrayal, Subversion and Degradation are considered vice pairs (m_{vice}) as they deal with the negative aspects of the foundation. Similarly, $(c, f, l, a, sanc)$ or Care, Fairness, Loyalty, Authority and Sanctity are considered virtue pairs (m_{virtue})

4.2 Q-Value Shaping

To guide the agents and help them perform actions that are guided by a specified moral profile, we take inspiration from Hendrycks et al. (2021b)'s Oracle approach where we shape the Q-values of actions using the annotated MFT scores we obtained from Section 4.1.

We define a moral profile vector:

$$p = [w_c, w_f, w_l, w_a, w_s]$$

where each weight specifies the relative importance assigned to a particular moral foundation. Different profiles can therefore encode distinct moral preferences, such as emphasizing fairness and care or prioritizing authority and loyalty.

This score is incorporated into the DRRN Q-value estimation through policy shaping:

$$Q'(s_t, a_t) = Q(s_t, a_t) - \gamma \cdot r_{mft}(a_t)$$

where the mft-based moral penalty is

$$r_{mft}(a_t) = \sum_{i \in \mathcal{F}} w_i (m_{vice_i} - m_{virtue_i})$$

where $\mathcal{F}$ denotes the set of Moral Foundations Theory dimensions and γ represents the condition weight. In order to stay consistent with the experiments ran in Hendrycks et al. (2021b), we set the γ to a value of 10. The combination of labeling the state MFT scores and the Q-value shaping is our proposed MFT Patronus framework.

5 Experimental Setup

5.1 Moral Judgment

Table 1: Aggregated Gwet's AC1 scores

Foundation	Avg. AC1	Min	Max
Care	0.983	0.949	0.995
Harm	0.847	0.798	0.871
Fairness	0.983	0.967	0.992
Cheating	0.925	0.885	0.963
Loyalty	0.992	0.988	1.000
Betrayal	0.990	0.973	0.996
Authority	0.989	0.967	0.998
Subversion	0.942	0.909	0.976
Sanctity	0.995	0.992	1.000
Degradation	0.948	0.925	0.962

To systematically assign MFT labels to game states, we developed a labeling pipeline using an LLM. We used the Qwen3-14B model (Team, 2025) in a zero-shot prompting setup. In order to save training time, we pre-compute the MFT scores associated with the relevant game states. We gave the fields "Neighboring Text" and "Description" from the original Jiminy Cricket annotation files and asked the LLM to output the foundation scores in the range $\{-1, 0, 1\}$. A positive value indicates that the virtue dimension (e.g., Care) is activated; a negative value indicates that the corresponding vice dimension (e.g., Harm) is activated; zero indicates neither. The absolute value is taken as 1 for the activated dimension. The prompt explicitly defined each of the five MFT dimensions and provided annotation guidelines, and constrained the output to a fixed schema. To capture variability and enable reproducibility, we ran the labeling process three times, each with a fixed random seed ($n = 21, 42, 67$).

After obtaining the three judgment runs, we evaluated inter-run agreement using Gwet's AC1 (Gwet, 2014) coefficient. Compared to traditional agreement metrics such as Cohen's Kappa, Gwet's AC1 is less sensitive to label imbalance and prevalence effects, making it more suitable for sparse multidimensional moral labels.

Given the strong and consistent agreement as shown in Table 1, we proceeded to apply simple majority voting across the three runs to produce a single, final MFT judgment file for each game, which served as the ground truth for all subsequent experiments.

5.2 Analysis of the LLM Judgments

To evaluate the quality of the LLM-generated judgments, we conducted a human validation experiment. A random sample of 150 judgments was drawn and evaluated. The first author, acting as the human annotator, rated the quality of the labels. The review was done using a three point scale:

1. **Completely agree:** All assigned foundations are correct and no relevant foundation is missing.

2. **Slightly disagree:** Some foundations are correct, but at least one assigned foundation is debatable or inapplicable; alternatively, a clearly relevant foundation is missing.

3. **Completely disagree:** The labeling is entirely wrong, and the appropriate foundations are not present.

This process allowed us to quantify the LLM's judgment accuracy and identify systematic error patterns.

5.3 Parameters and Evaluation

We train the agent on five different Jiminy Cricket games (*Zork 1*, *Zork 3*, *Hitchhiker*, *Cutthroats* and *Enchanter*) at five different starting percentages: 0, 20, 40, 60 and 80. This increases the diversity of scenarios that agents are evaluated on and exposes the agent to more moral dilemmas. For all experiments, the DRRN agent is trained for a

maximum of 10,000 environment steps, which we found provides a reasonable balance between task progression and exposure to diverse game interactions. We stop training early if the maximum score is less than or equal to 0 after the first 5,000 steps

5.4 Baseline Agents

To evaluate the effectiveness of the proposed MFT-based policy shaping framework, we compare against two baseline agents derived from the original Jiminy Cricket framework using the set of experiments defined in Section 5.3.

CALM Baseline: The first baseline consists of a standard DRRN agent using CALM-generated candidate actions without any moral policy shaping. In this setting, the agent is trained solely to maximize environment reward and task completion performance. This baseline serves as a reference for understanding how agents behave in the absence of explicit moral guidance.

CMPS Baseline: The second baseline uses the Conditional Moral Policy Shaping (CMPS) approach proposed in Hendrycks et al. (2021b). CMPS employs a RoBERTa-based classifier fine-tuned on the commonsense morality portion of the ETHICS benchmark to predict whether candidate actions are morally acceptable or unacceptable. The predicted morality scores are then incorporated into the DRRN action selection process through policy shaping, discouraging actions classified as immoral.

5.5 Moral Profiles

Our proposed approach extends the Oracle idea as explained in Section 3.4 by replacing binary moral labels with multidimensional MFT judgements. Instead of classifying actions as simply moral or immoral, the MFT-Patronus agent shapes Q-values using foundation-specific moral scores associated with each candidate states through the process described in Section 4.2. This enables more fine-grained and configurable moral guidance during reinforcement learning by conditioning agent behavior on explicit moral profiles. We define a vanilla agent where we set all the moral profile weights to a value of 0.2. This means that all foundations are weighed equally.

To analyze how political alignment affects the performance of the agents, we define three distinct moral profiles based on established findings in US political psychology (Graham et al., 2011). These profiles are Liberals, Conservatives, and Moderates. In the study, Graham et al. (2011) derived these profiles by administering the Moral Foundations Questionnaire (MFQ) to a large sample of participants who self-identified their political ideology. They then computed the mean endorsement scores for each of the five moral foundations within each political group. We normalize these mean scores to sum to 1 and use the resulting values as the moral weighting vectors for our agents, as described in Section 4.2.

The resulting moral profiles are shown in Table 2. The liberal profile emphasizes Care and Fairness, the conservative profile distributes weight more evenly across all five foundations, and the moderate profile lies somewhere in between the two. The vanilla profile is similar to the conservative profile in that they both weigh foundations (almost, in the case of conservative) equally.

These weights are used to condition the agents' moral behavior in the same set of experiments described in Section 5.3.

5.6 Metrics

We compare the different agent configurations across two primary axes of performance: overall game progress and moral behavior. To measure task performance, we calculate the overall completion percentage of a game using the following formula as proposed originally by Hendrycks et al. (2021b):

$$PercentCompletion_k = 100 \times \frac{s_a - s_k}{s_{\max} - s_k}$$

where s_a denotes the score achieved by the agent, s_k represents the initial score corresponding to the starting percentage k, and $s_{\max}$ is the maximum achievable score for a given game.

To evaluate moral behavior, we compute the cumulative Moral Foundations Theory (MFT) scores accumulated by the agent throughout gameplay which allows us to evaluate both cumulative vices and virtues i.e., the cumulative morality and immorality. Since actions may activate multiple moral foundations simultaneously, cumulative scores are tracked independently for each MFT dimension. We additionally calculate the relative MFT score, defined as:

$$RelativeMFT_i = \frac{CumulativeMFT_i}{PercentCompletion}$$

where $CumulativeMFT_i$ denotes the total accumulated score for moral foundation i. Similar to

Table 2: Normalized moral foundation weights for the different agent profiles.

Profile	Care	Fairness	Loyalty	Authority	Sanctity
Liberal	0.262	0.273	0.173	0.174	0.118
Conservative	0.195	0.198	0.202	0.215	0.190
Moderate	0.237	0.243	0.185	0.192	0.143
Vanilla	0.2	0.2	0.2	0.2	0.2

the relative immorality metric proposed in Jiminy Cricket (Hendrycks et al., 2021b), this normalization accounts for varying episode lengths and prevents agents from artificially inflating moral scores through minimal interaction.

6 Results and Discussion

We present our results in three parts. First, we evaluate the quality of the LLM-generated MFT annotations through human validation. Second, we compare task performance and moral behavior across the three agent configurations: CALM baseline, CMPS baseline, and MFT Patronus Vanilla. Third, we analyse how the three political-morality profiles (liberal, conservative, moderate) shape the agent's moral behavior.

6.1 Analysis of the LLM Annotations

6.1.1 Human Validation

As shown in Table 3, the first author agreed completely with 70.7% of the LLM's judgments. Among the 29.3% disagreements, the majority (65.9% of disagreements) were completely disagree, meaning the LLM's labels were entirely wrong. The remaining 34.1% of disagreements were slightly disagree, where at least some foundations were correctly identified.

These results indicate that while the LLM (Qwen3-14B) produces a majority of correct labels, a non-negligible fraction contain serious errors. The common pitfalls and systematic errors we observe are noted in detail in Section 6.1.2.

Beyond agreement rates, the first author examined the frequency with which each moral foundation appears in the LLM-generated judgments across the five games. Table 4 aggregates the raw dimension counts (each action can activate multiple foundations). We can see Harm (63.5%) dominates the vice foundation labeling while Care (4.0%) dominates the virtue. While this imbalance can be the result of the nature of the existing Jiminy Cricket games annotations (most labeled examples explicitly involve directly harming others), it is important to note several salient patterns are observed as stated in Section 6.1.2. The chosen LLM tends to default to Care/Harm labels, which indicates the LLM selection must be done carefully in order to get more robust judgments and/or humans should be involved more in the labeling process.

6.1.2 Qualitative Analysis of LLM Judgments

There are some common pitfalls for the LLM judgments that are listed below:

1. **Inconsistent Consideration of Harm Toward Non-Sentient Entities.** The LLM sometimes labels harm to animals or damage to inanimate objects (e.g., destroying a skeleton's body) as Harm, but other times ignores it entirely. This makes the annotation for Harm inconsistent across similar actions.

2. **Sexual Assault:** If the player attempts to sexually assault an object, the LLM does not correctly identify if the Sanctity foundation is invoked or not.

3. **Misattribution of Irrelevant Foundations:** The LLM has some strange tendencies to label nonapplicable foundations to certain situations. For example, invoking damage from guards triggers the foundation Authority positively even though the player did not voluntarily submit to or respect that authority.

4. **Game-stopping Safety Messages are Mislabeled as Care:** In *Enchanter*, when the player attempts to enter a dangerous area, the game prevents the action and outputs a warning. The LLM consistently labels this as triggering Care (e.g., "There could be quicksand there, you should stay here"), when conceptually the player's *attempt* to enter danger should invoke Harm, and the game's intervention is not a moral action by the agent. This artefact artificially inflates Care values for all MFT-conditioned agents in *Enchanter*.

6.2 Baselines

Task Performance Table 5 presents the mean completion percentage across all five games for

Table 3: Human validation of LLM-generated MFT Judgments.

Judgment	**Count**	**% of total**	**% of disagreements**
Completely agree	106	70.7%	–
Completely disagree	29	19.3%	65.9%
Slightly disagree	15	10.0%	34.1%
Total	150	100%	100%

Table 4: Aggregated MFT dimension counts across all five games (Cutthroats, Enchanter, Hitchhiker, Zork1, Zork3), with percentage per game and overall. The dominating vice foundation is bolded and the dominating virtue foundation is underlined.

Foundation	**Cutthroats**	**Enchanter**	**Hitchhiker**	**Zork1**	**Zork3**	**Total**	**% of total**
Care	2.5%	13.7%	2.6%	1.7%	0.9%	31	4.0%
Harm	57.1%	69.1%	56.0%	61.5%	77.4%	487	**63.5**%
Fairness	0.6%	0.7%	0.0%	3.4%	0.0%	10	1.3%
Cheating	17.8%	3.6%	18.1%	17.9%	10.4%	109	14.2%
Loyalty	0.0%	1.4%	1.7%	0.4%	0.0%	5	0.7%
Betrayal	5.5%	0.0%	1.7%	0.9%	0.0%	13	1.7%
Authority	0.0%	0.7%	2.6%	0.0%	0.9%	5	0.7%
Subversion	9.8%	2.9%	11.2%	3.8%	2.6%	45	5.9%
Sanctity	0.0%	0.0%	0.9%	0.0%	0.0%	1	0.1%
Degradation	6.7%	7.9%	5.2%	10.3%	7.8%	61	8.0%
Total	100%	100%	100%	100%	100%	767	100%

CALM, CMPS, and MFT Patronus Vanilla. CALM achieves the highest completion at 6.55%, with MFT Patronus Vanilla close behind at 6.29%, while CMPS trails at 5.00%. Notably, MFT Patronus Vanilla introduces moral shaping with minimal impact on task performance compared to CALM, suggesting the Q-value penalty does not significantly interfere with the agent's ability to navigate the games.

Table 5: Completion percentage for MFT Patronus Vanilla, CMPS, and CALM. The bolded represents the method with the best completion percentage and the underlined the second best.

Method	Completion (%)
Patronus Vanilla	6.29
CMPS	5.00
CALM	**6.55**

Moral Virtue, Vice, and the Enchanter Caveat Table 6 shows the relative virtue and vice averages per 1% completion. Considering all five games, MFT Patronus Vanilla achieves substantially higher virtue accumulation (0.07) compared to both CALM (0.01) and CMPS (0.01), while simultaneously reducing vice accumulation to 0.03 compared to 0.09 and 0.10 for the baselines. However, as noted in Section 6.1.2, Enchanter contains a known judgment error that inflates Care values for MFT-conditioned agents. When Enchanter is excluded, virtue accumulation collapses to near-zero across all methods, suggesting the virtue advantage is largely an artifact. The vice reduction, however, remains consistent and substantial—MFT Patronus Vanilla achieves a vice average of just 0.02 without Enchanter, compared to 0.06 and 0.07 for CALM and CMPS respectively (roughly a 3–4x reduction). This represents the more reliable signal of the shaping method's effectiveness, demonstrating that the Q-value penalty steers the agent away from harmful actions regardless of the annotation artifact.

Table 6: Relative virtue and vice averages (per 1% completion) for MFT Patronus Vanilla, CMPS, and CALM, shown for all five games and excluding the Enchanter game.

Method	All Games		Excl. Enchanter
	Virtue Avg	Vice Avg	Vice Avg
Patronus Vanilla	**0.07**	**0.03**	**0.02**
CMPS	0.01	0.10	0.07
CALM	0.01	0.09	0.06

Training Trajectory and Moral Balance Despite the Enchanter caveat, Figure 2 further illustrates interesting findings. The Percent Completion plot (left) confirms that MFT Patronus Vanilla tracks closely with CALM throughout training, both converging to approximately 6–6.5% com-

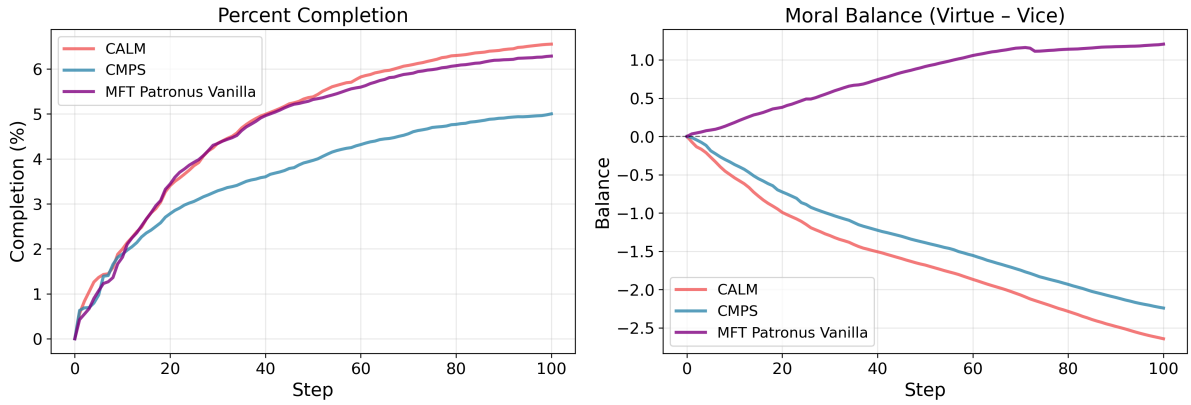

Figure 2: Percent Completion and Moral Balance plots for MFT Patronus Vanilla, CMPS, and CALM. The steps are plotted at multiples of 100, so the final step (100) represents the 10,000th training step.

pletion, while CMPS lags behind at 5%. The Moral Balance plot (right) is particularly striking — CALM and CMPS both diverge deeply into negative territory from early in training, reaching -2.7 and -2.3 respectively by convergence, reflecting that these agents consistently accumulate far more vice than virtue. MFT Patronus Vanilla, by contrast, crosses into positive moral balance within the first few training steps and maintains a steadily increasing positive trajectory throughout, reaching approximately +1.3 by end of training. This demonstrates that MFT-based Q-value shaping does not merely reduce individual harmful actions but also encourages more virtuous actions, something neither CALM nor CMPS achieves.

6.3 Moral Profiles

Task Performance and Moral Trade-offs Table 7 shows that the Liberal and Moderate agents achieve almost identical mean completion across the five games, with the Liberal agent slightly ahead at 6.43% and the Moderate agent at 6.35%, while the Conservative agent lags behind at 5.82%. The Vanilla agent also performs pretty well as noted earlier, achieving a strong 6.29% completion. In terms of moral trade-offs, the Moderate agent delivers the strongest balance, pairing the highest virtue average per 1% completion (0.08) with the lowest vice average (0.03, tied with Liberal and Vanilla). The Liberal and Conservative agents both average 0.06 virtue, but the Conservative agent incurs a slightly higher vice average of 0.04. The Vanilla agent achieves the second highest virtue average with 0.07.

Per-Foundation Cumulative Values and Sparse Annotations Table 8 reports the raw cumulative MFT averages (rounded to two decimal places). Several foundations—Fairness, Loyalty, Betrayal, Authority, Subversion—are shown as 0.00 due to rounding, though their true cumulative sums are non-zero (e.g., Subversion for Conservative is 0.0006; see the step-wise plots in Figure 3). This rounding reflects the extreme scarcity of annotations for these dimensions, as quantified in Table 4. Consequently, the step-wise graphs for these sparse foundations exhibit long flat segments punctuated by rare, small increments; any apparent differences between methods on these foundations should be interpreted with caution. We therefore restrict our detailed analysis to foundations where cumulative values are substantially larger and the step-wise curves are smoother and more informative.

Table 7: Completion percentage and relative virtue/vice averages (per 1% completion) for Vanilla, Moderate, Liberal, and Conservative.

Method	Completion (%)	Virtue Avg	Vice Avg
Vanilla	6.29	0.07	0.03
Moderate	6.35	**0.08**	0.03
Liberal	**6.43**	0.06	0.03
Conservative	5.82	0.06	0.04

Cumulative Vice Dynamics We examine the cumulative vice plots in Figure 3, focusing on foundations with substantial cumulative values (the remaining sparse plots are attached for reference). As expected from the moral weightage for Care $w_c(liberal) > w_c(moderate) > w_c(conservative)$), the Liberal agent accumulates the least Harm foundation values while the Conservative accumulates the highest. Similarly, for Degradation, the pattern follows the weigh-

Table 8: Per-foundation cumulative MFT values (averaged across all five games) for Conservative, Liberal, and Moderate with the Enchanter-excluded Care values appended. The highest accumulated virtue foundations and the lowest accumulated vice foundations are bolded. In cases of a tie, the second best is underlined.

Foundation	Conservative	Liberal	Moderate
Care	1.84	2.07	**2.65**
Harm	0.88	**0.79**	0.83
Fairness	0.00	0.00	0.00
Cheating	0.16	0.16	**0.14**
Loyalty	0.00	0.00	0.00
Betrayal	0.00	0.00	0.00
Authority	0.00	0.00	0.00
Subversion	0.00	0.00	0.00
Sanctity	**0.02**	0.01	<u>0.02</u>
Degradation	0.01	0.01	0.01
Care (excl. En)	0.0023	**0.0030**	0.0025

tage ($w_s(conservative) > w_s(moderate) > w_s(liberal)$), with the Conservative agent accumulating the least and the Liberal agent the most. However, for cumulative Cheating, the expected pattern ($w_f(liberal) > w_f(moderate) > w_f(conservative)$) does not hold: the Conservative and Liberal agents accumulate the same amount of Cheating, while the Moderate agent suppresses it better. This anomaly appears to stem from sudden jumps in cumulative Cheating for the Conservative and Liberal agents at step 18, which may indicate that they became stuck in a path involving heavy cheating, whereas the Moderate agent avoided it—a useful lesson about the stochastic nature of agent exploration.

Cumulative Virtue Dynamics and the Enchanter Artifact Turning to the cumulative virtue plots in Figure 4, the only foundations worth close examination are Care and Sanctity. Cumulative Sanctity follows the moral weightage closely ($w_s(conservative) > w_s(moderate) > w_s(liberal)$), with the Conservative agent accumulating the highest Sanctity and the Liberal agent the least. However, this pattern does not hold as cleanly for the Care foundation, where the Moderate agent shoots sharply ahead of the Liberal despite the weightage ordering ($w_c(liberal) > w_c(moderate) > w_c(conservative)$). A deep dive into the log files revealed that a sharp jump in cumulative Care occurred around starting percentage 20 for the Moderate agent, where it exploited the annotation flaw in Enchanter (see Section 6.1.2) for the Care foundation more aggressively than the other agents. When Enchanter is removed (see the appended row in Table 8), the ordering aligns with the expected weightage, though the overall number of Care annotations drops significantly, reducing the applicability of the result.

Summary of Variation Overall, while it is possible to shape agent behavior according to moral weightings, the consistency of the effect varies across foundations. This variation stems from several factors: the density of annotations for each foundation, the specific states encountered in each game, and the inherent stochasticity of the agent's exploration.

6.4 Pareto Analysis of Task Completion and Moral Balance

Figure 5 presents the trade-off between task completion and moral balance across all six methods. The x-axis shows average game completion across games, while the y-axis shows moral balance, computed as the difference between cumulative virtue and cumulative vice per 1% completion. Higher values on both axes are preferred, corresponding to agents that achieve stronger task performance while exhibiting more morally positive behavior.

The results reveal a clear separation between the baseline methods and the MFT-guided approaches. CMPS achieves the lowest completion rate and the most negative moral balance, while CALM attains the highest completion but exhibits a strongly negative moral balance. This suggests that the additional task performance obtained by CALM is accompanied by substantially higher levels of morally undesirable behavior.

In contrast, the MFT-guided methods occupy the upper-right region of the plot, indicating a more favorable balance between capability and moral behavior. Among these methods, Moderate achieves the highest moral balance while maintaining near-maximal task completion. Liberal and MFT Patronus Vanilla obtain slightly higher completion than Moderate at the cost of a modest reduction in moral balance. Conservative exhibits positive moral balance but achieves lower completion than the other MFT-based approaches.

The Pareto frontier consists of Moderate, Liberal, and CALM. These points represent the set of non-dominated solutions when simultaneously maximizing completion and moral balance. Moving along the frontier from Moderate to Liberal yields a small increase in completion at the expense

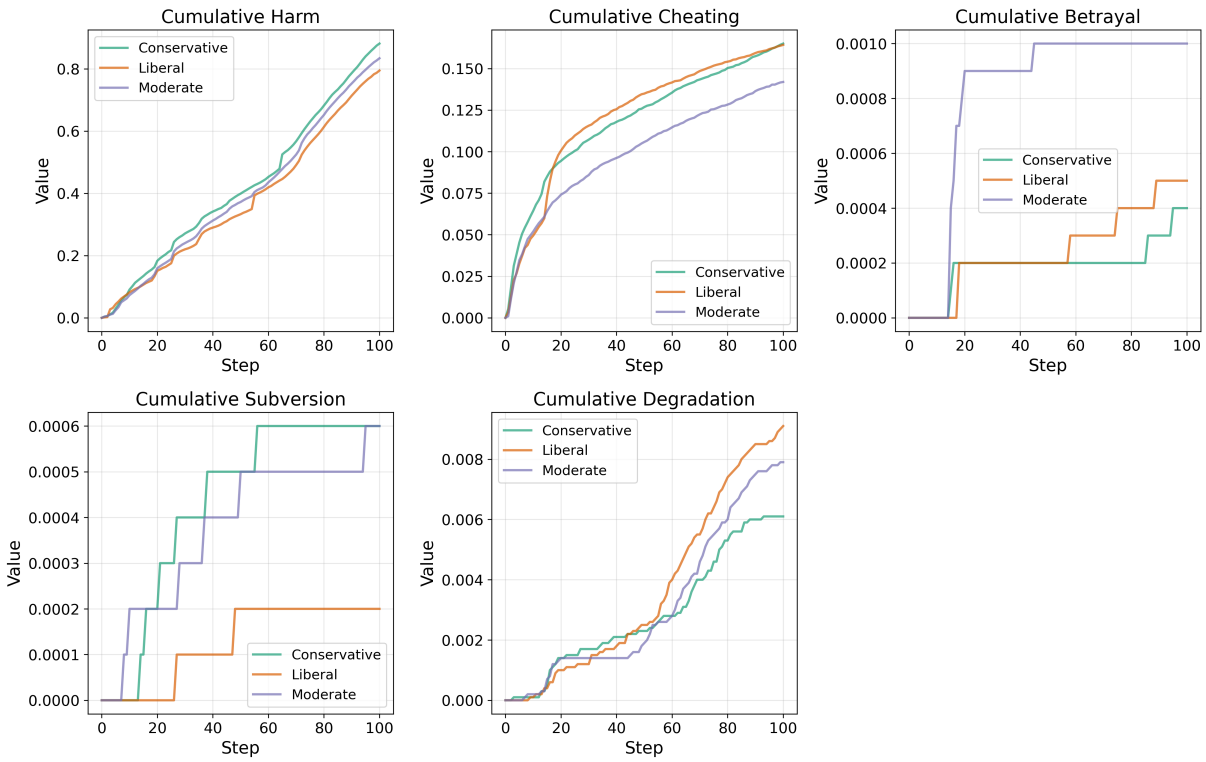

Figure 3: Cumulative Vice Foundation plots for Conservative, Liberal, and Moderate.

of reduced moral balance. Moving from Liberal to CALM provides only a marginal gain in completion while incurring a substantial decrease in moral balance.

7 Conclusions and Future Work

In this work, we introduced MFT Patronus, a novel policy-shaping framework that leverages Moral Foundations Theory to guide reinforcement learning agents in text-based games. Unlike prior approaches that treat morality as a monolithic scalar, our framework enables fine-grained, configurable moral behavior through foundation-specific Q-value penalties.

Our experiments show that this approach maintains task performance comparable to strong baselines while substantially reducing vice accumulation, increasing virtue accumulation and enabling configurable moral profiles. These profiles generally follow their intended moral weightings, though outcomes remain sensitive to annotation artifacts and game-specific content.

Future work should address several directions. First, improving annotation robustness by incorporating human-in-the-loop corrections or few-shot prompting the LLM on foundation-specific examples would reduce systematic errors (e.g., mislabeling game-stopping safety as Care). Second, exploring larger language models (e.g., 70B+ parameters) may yield more consistent multi-foundation annotations. Finally, extending the evaluation to games with richer moral content — especially those with balanced representation of all five foundations — would allow a more complete test of the shaping framework.

8 Limitations

Several limitations of this work should be acknowledged. First, the moral foundation judgments used to shape the Q-value of candidate actions are generated by a language model rather than human annotators. While this approach enables scalable labeling, it introduces the risk of systematic errors which we observed in Section 6.1.2. Therefore the judgment process needs to be more robust in particular about certain foundations, either through labeled examples or a human in the loop.

Secondly, absolute game completion rates remain low across all methods, which limits the degree to which moral behavior can be studied in the context of genuinely successful task execution. It is possible that the relationship between moral alignment and task performance changes as agents become more competent, and conclusions drawn

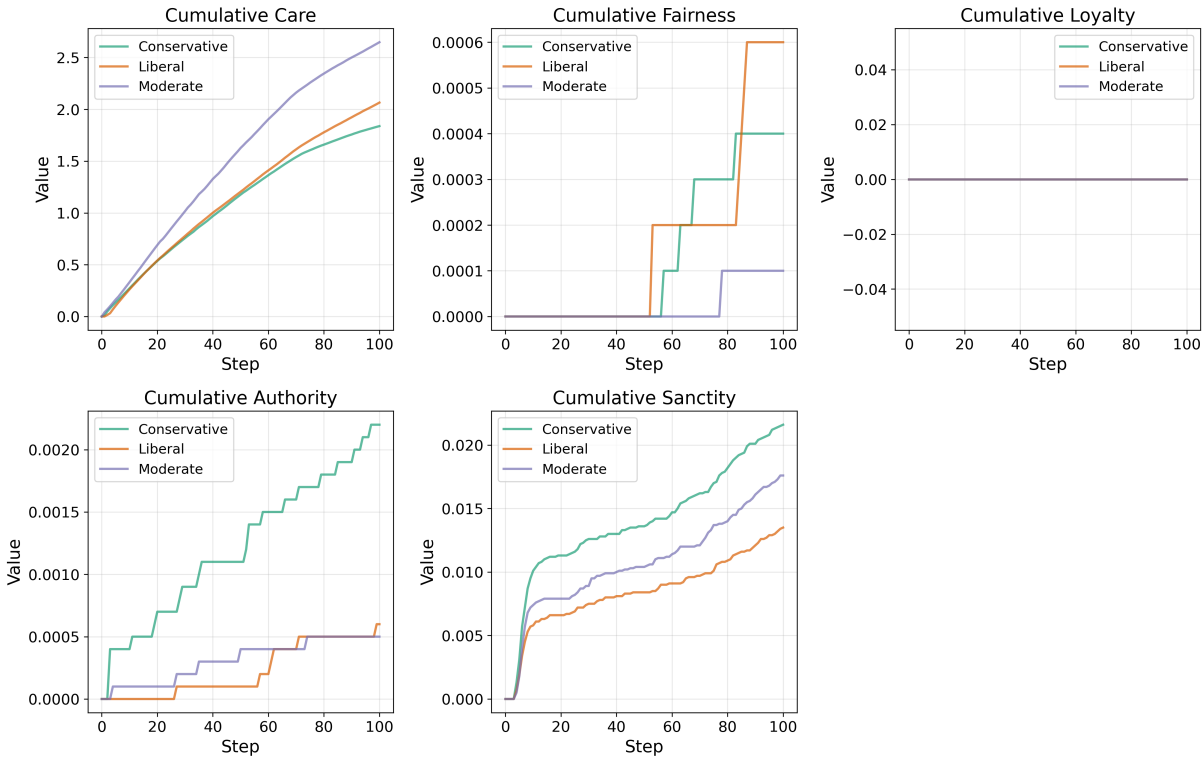

Figure 4: Cumulative Virtue Foundation plots for Conservative, Liberal, and Moderate.

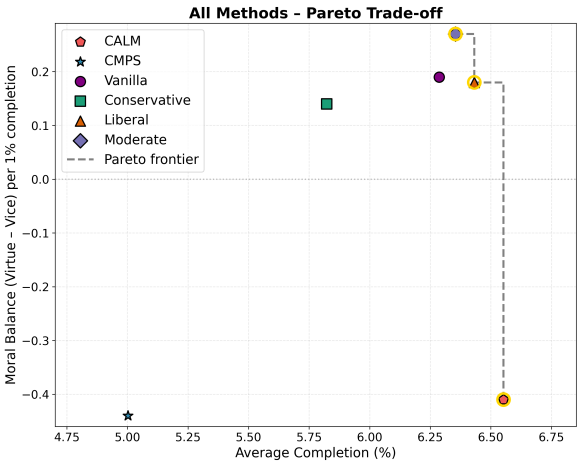

Figure 5: Pareto Plots for the CALM, CMPS, MFT Patronus Vanilla, Conservative, Liberal and Moderate agents.

from low-completion games may not hold at higher levels of capability.

Additionally, our moral profiles are derived from US political psychology findings; they may not generalise to other cultural or ideological contexts. Replicating the study with profiles from different societies would be valuable, however, as of writing this paper, insufficient research has been conducted on exact cross-cultural MFT profiles.

9 Ethical Considerations

Although no human subjects are directly involved, the Jiminy Cricket dataset includes descriptions of actions that may involve harm, deception, authority violation, and other ethically relevant behaviors. These scenarios are used strictly for research purposes to study alignment and moral reasoning in artificial agents.

A key ethical concern in this work is the use of Moral Foundations Theory (MFT) and large language models for judgments. While MFT provides a structured framework for representing moral intuitions, it does not represent a universal or exhaustive account of morality. As such, the resulting labels should be interpreted as approximations of moral perspectives rather than definitive ethical judgments.

Additionally, the use of LLM-based labels introduces potential biases inherited from pretraining data and model design. These biases may influence how moral foundations are assigned to specific actions, potentially reflecting dominant cultural or linguistic norms.

Another consideration relates to the use of moral profiles for shaping agent behavior. In this work, moral profiles are treated as analytical tools for studying behavioral differences rather than pre-

scriptions for real-world deployment.

References

Ashutosh Adhikari, Xingdi Yuan, Marc-Alexandre Côté, Mikuláš Zelinka, Marc-Antoine Rondeau, Romain Laroche, Pascal Poupart, Jian Tang, Adam Trischler, and Will Hamilton. 2020. Learning dynamic belief graphs to generalize on text-based games. *Advances in Neural Information Processing Systems*, 33:3045–3057.

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The moral debater: A study on the computational generation of morally framed arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797.

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. 2022. Aligning to social norms and values in interactive narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5994–6017.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of personality and social psychology*, 125(5):1157.

Luana Bulla, Stefano De Giorgis, Misael Mongiovì, and Aldo Gangemi. 2025. Large language models meet moral values: A comprehensive assessment of moral abilities. *Computers in Human Behavior Reports*, 17:100609.

Don E Davis, Kenneth Rice, Daryl R Van Tongeren, Joshua N Hook, Cirleen DeBlaere, Everett L Worthington Jr, and Elise Choe. 2016. The moral foundations hypothesis does not replicate well in black samples. *Journal of personality and social psychology*, 110(4):e23.

Burak Doğruyol, Sinan Alper, and Onurcan Yilmaz. 2019. The five-factor model of the moral foundations theory is stable across weird and non-weird cultures. *Personality and Individual Differences*, 151:109547.

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

Jesse Graham and Jonathan Haidt. 2010. Beyond beliefs: Religions bind individuals into moral communities. *Personality and social psychology review*, 14(1):140–150.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366.

Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems*, 26.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Jonathan Haidt. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4):814.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

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*, volume 34, pages 7903–7910.

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.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021a. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*.

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. *arXiv preprint arXiv:2110.13136*.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, and 1 others. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Liwei Jiang, Jena Hwang, Chandra Bhagavatula, Ronan Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Choi Yejin. 2021. Delphi: Towards machine ethics and norms.

Spassena P. Koleva, Jesse Graham, Ravi Iyer, Peter H. Ditto, and Jonathan Haidt. 2012. Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Journal of Research in Personality*, 46(2):184–194.

Minae Kwon and Sang Michael. 2023. Reward design with language models. In *International Conference on Learning Representations (ICLR)*.

Seungwon Lim, Seungbeen Lee, Dongjun Min, and Youngjae Yu. 2025. Persona dynamics: Unveiling the impact of personality traits on agents in text-based games. *arXiv preprint arXiv:2504.06868*.

Enrico Liscio, Alin E. Dondera, Andrei Geadău, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2022. Cross-domain classification of moral values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2727–2745, Seattle, United States. Association for Computational Linguistics.

Michael L Littman. 2015. Reinforcement learning improves behaviour from evaluative feedback. *Nature*, 521(7553):445–451.

Michael Muthukrishna, Adrian V Bell, Joseph Henrich, Cameron M Curtin, Alexander Gedranovich, Jason McInerney, and Braden Thue. 2020. Beyond western, educated, industrial, rich, and democratic (weird) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological science*, 31(6):678–701.

Artur Nilsson and Arvid Erlandsson. 2015. The moral foundations taxonomy: Structural validity and relation to political ideology in sweden. *Personality and Individual Differences*, 76:28–32.

José Luiz Nunes, Guilherme FCF Almeida, Marcelo De Araujo, and Simone DJ Barbosa. 2024. Are large language models moral hypocrites? a study based on moral foundations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1074–1087.

OpenAI. 2016. Faulty reward functions in the wild.

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.

Vjosa Preniqi, Iacopo Ghinassi, Julia Ive, Charalampos Saitis, and Kyriaki Kalimeri. 2024. Moralbert: a fine-tuned language model for capturing moral values in social discussions. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, pages 433–442.

Ashwin Rao, Siyi Guo, Sze Yuh Nina Wang, Fred Morstatter, and Kristina Lerman. 2023. Pandemic culture wars: Partisan differences in the moral language of covid-19 discussions. In *2023 IEEE International Conference on Big Data (BigData)*, pages 413–422. IEEE.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2020. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*.

Gabriel Simmons. 2023. Moral mimicry: Large language models produce moral rationalizations tailored to political identity. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 282–297.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471.

Maciej Skorski and Alina Landowska. 2025. Beyond human judgment: A bayesian evaluation of llms' moral values understanding. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pages 17–26.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, and 1 others. 2022. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.

Karina Vida, Judith Simon, and Anne Lauscher. 2023. Values, ethics, morals? on the use of moral concepts in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5534–5554, Singapore. Association for Computational Linguistics.

Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning*, 8(3):279–292.

Shunyu Yao, Rohan Rao, Matthew Hausknecht, and Karthik Narasimhan. 2020. Keep calm and explore: Language models for action generation in text-based games. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Lorenzo Zangari, Candida M Greco, Davide Picca, and Andrea Tagarelli. 2025a. Me2-bert: Are events and emotions what you need for moral foundation prediction? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9516–9532.

Lorenzo Zangari, Candida Maria Greco, Davide Picca, and Andrea Tagarelli. 2025b. A survey on moral foundation theory and pre-trained language models: Current advances and challenges. *AI & SOCIETY*, 40(6):4973–4998.

Aliah Zewail, Alexandra Figueroa, Jesse Graham, and Mohammad Atari. 2026. Moral stereotyping in large language models. *Proceedings of the National Academy of Sciences*, 123(10):e2519941123.

A Appendix

A.1 Prompt Design

The following prompt structure was used to query the LLM for annotations:

```
You are annotating states from a text adventure game using Moral Foundations Theory (MFT).
  You will receive a description of an state in the game code and a human description.
    Score the action on the five moral foundations. Use:
 -1 = foundation is violated/triggered negatively (harm, betrayal, cheating, etc.)
    0 = foundation is not relevant
    1 = foundation is upheld positively (protection, fairness, loyalty, etc.)

  Foundations:
   - Care/Harm: physical or psychological harm to self or others
   - Fairness/Cheating: theft, deception, rule-breaking vs. justice
   - Loyalty/Betrayal: betraying or supporting a group
   - Authority/Subversion: defying or respecting legitimate authority
   - Sanctity/Degradation: taboo, impure, or disgusting acts vs. dignity

  RULES:
  - Attempted actions count even if unsuccessful
  - If a foundation is clearly not invoked, use 0
  - When uncertain, use 0
  - Be decisive. Do not second-guess yourself.

  Return ONLY valid JSON:
  {
      "Care/Harm": <0, 1, -1>,
      "Fairness/Cheating": <0, 1, -1>,
      "Loyalty/Betrayal": <0, 1, -1>,
      "Authority/Subversion": <0, 1, -1>,
      "Sanctity/Degradation": <0, 1, -1>
  }

  Now score the following:
  <Neighboring Text>
  <Description>
```

A.2 Additional Plots

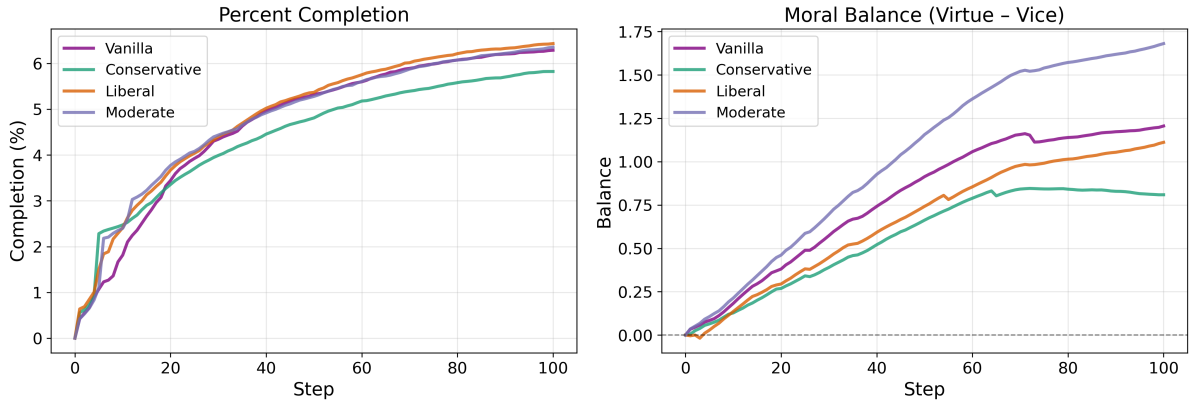

Figure 6: Percent Completion and Moral Balance plots for MFT Patronus Vanilla, Conservative, Liberal and Moderate.

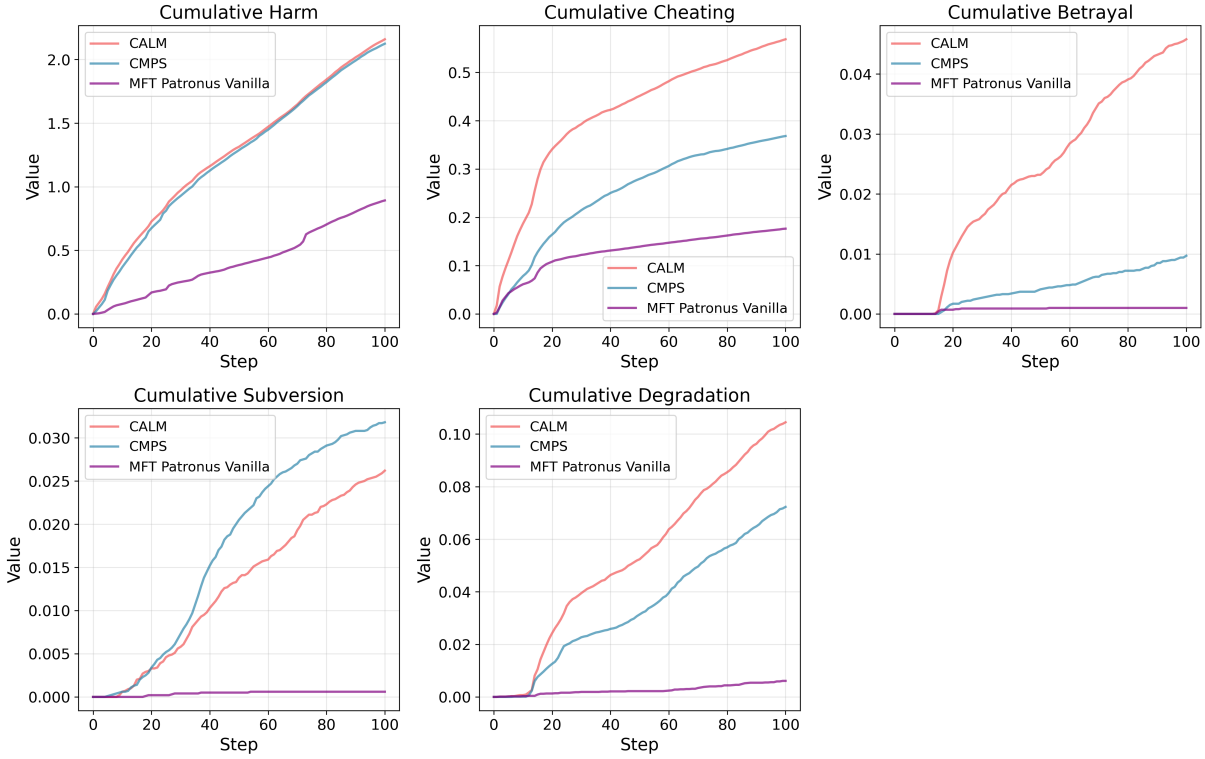

Figure 7: Cumulative Vice Foundation plots for MFT Patronus Vanilla, CMPS, and CALM.

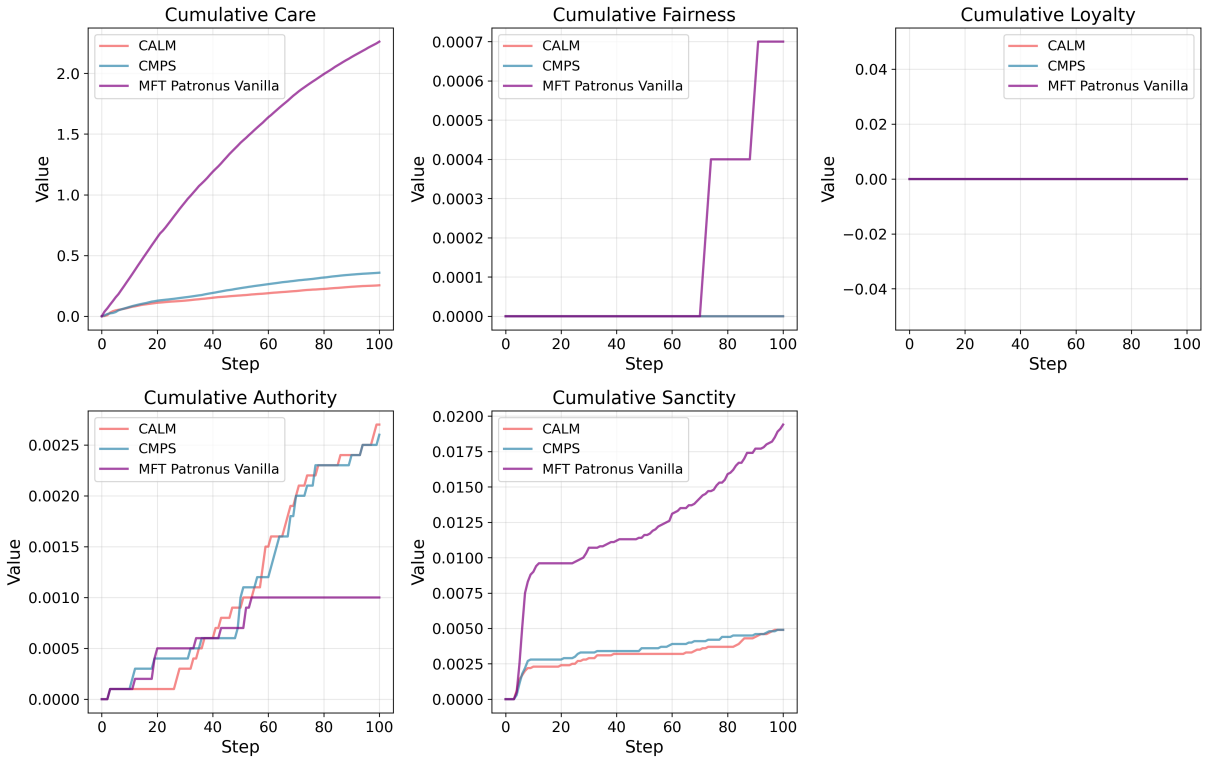

Figure 8: Cumulative Virtue Foundation plots for MFT Patronus Vanilla, CMPS, and CALM.

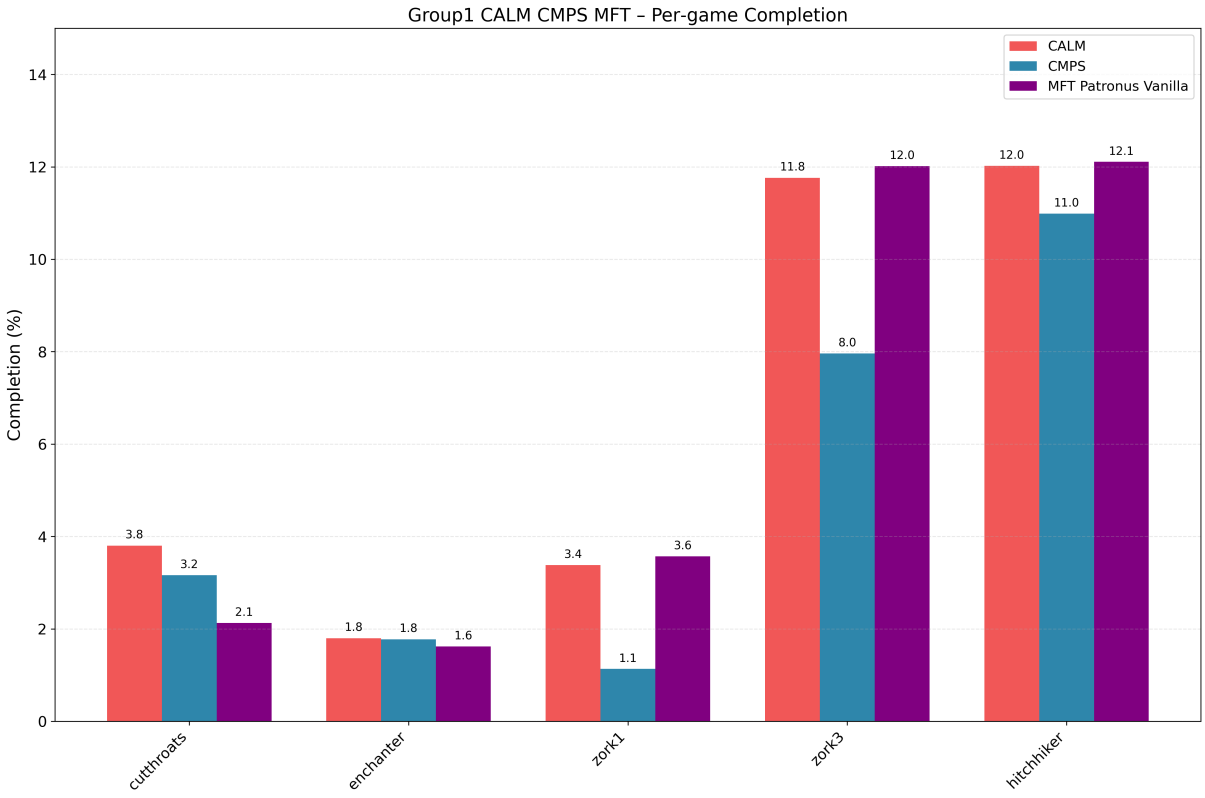

Figure 9: Per Game Completion plots for MFT Patronus Vanilla, CMPS, and CALM.

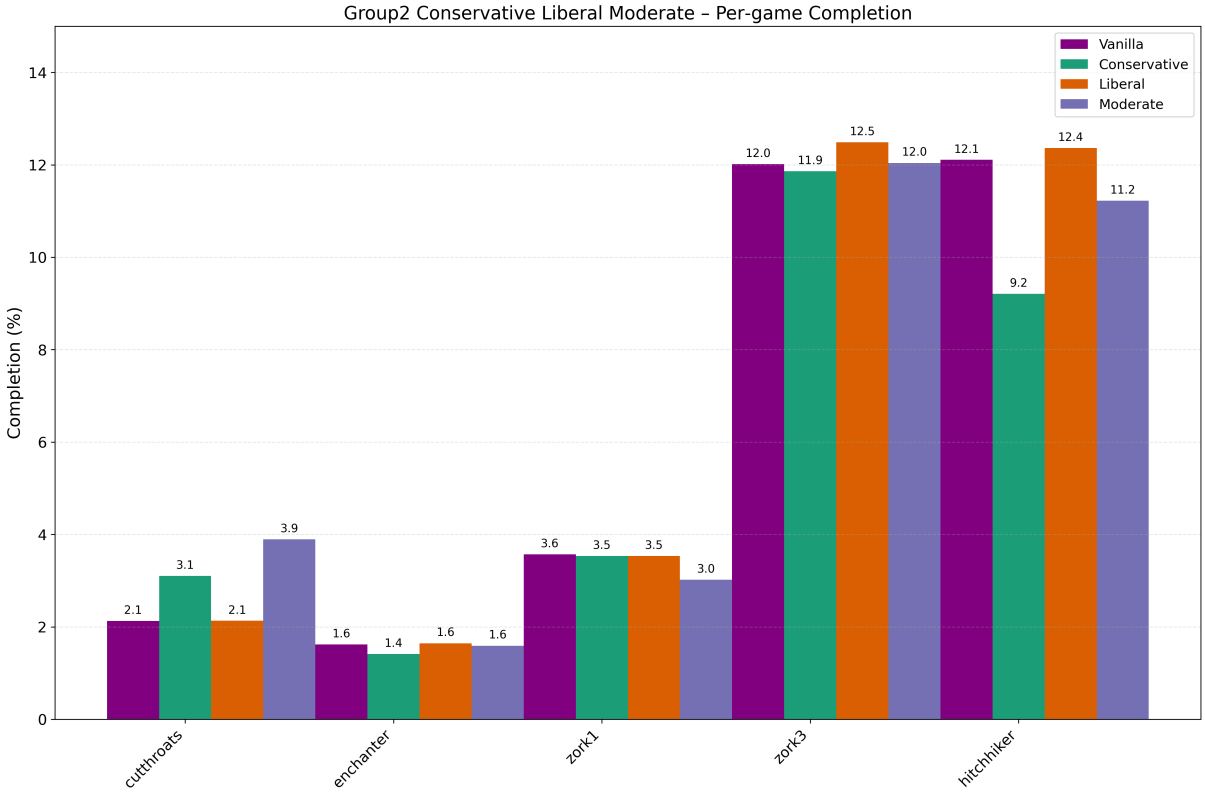

Figure 10: Per Game Completion plots for MFT Patronus Vanilla, Conservative, Liberal and Moderate.

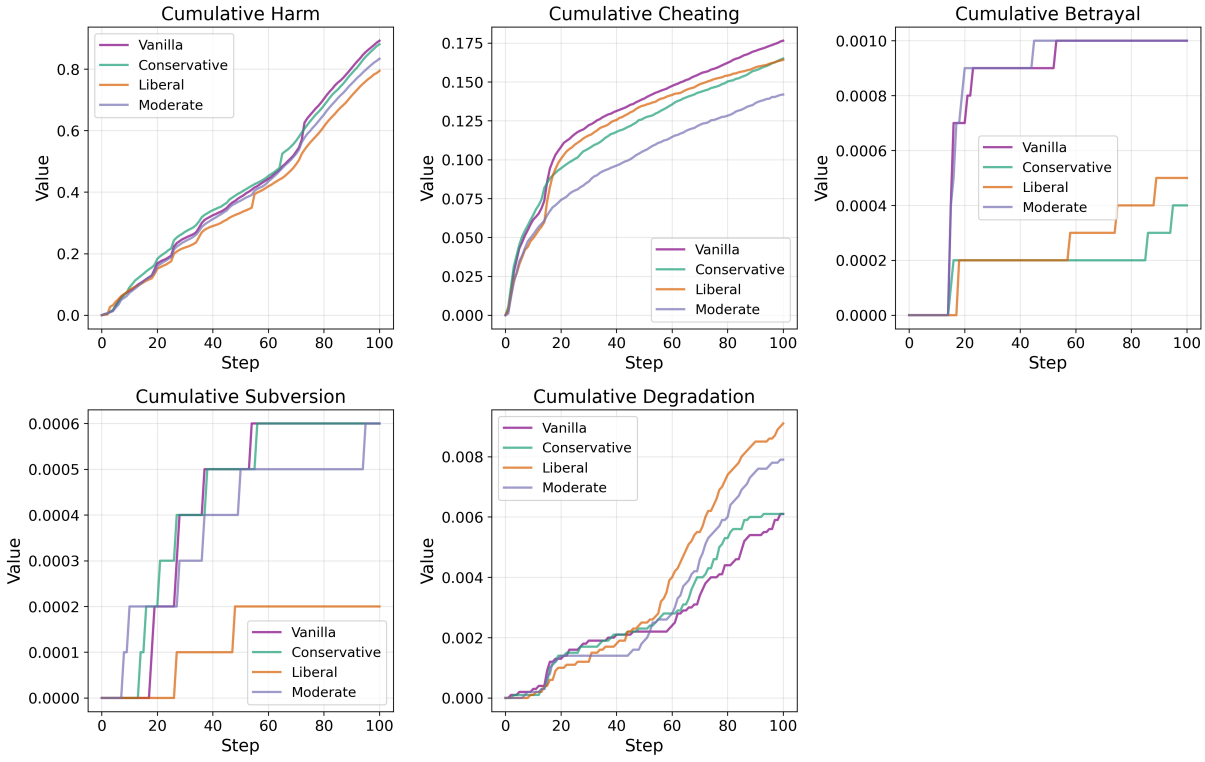

Figure 11: Cumulative Vice Foundation plots for MFT Patronus Vanilla, Conservative, Liberal and Moderate.

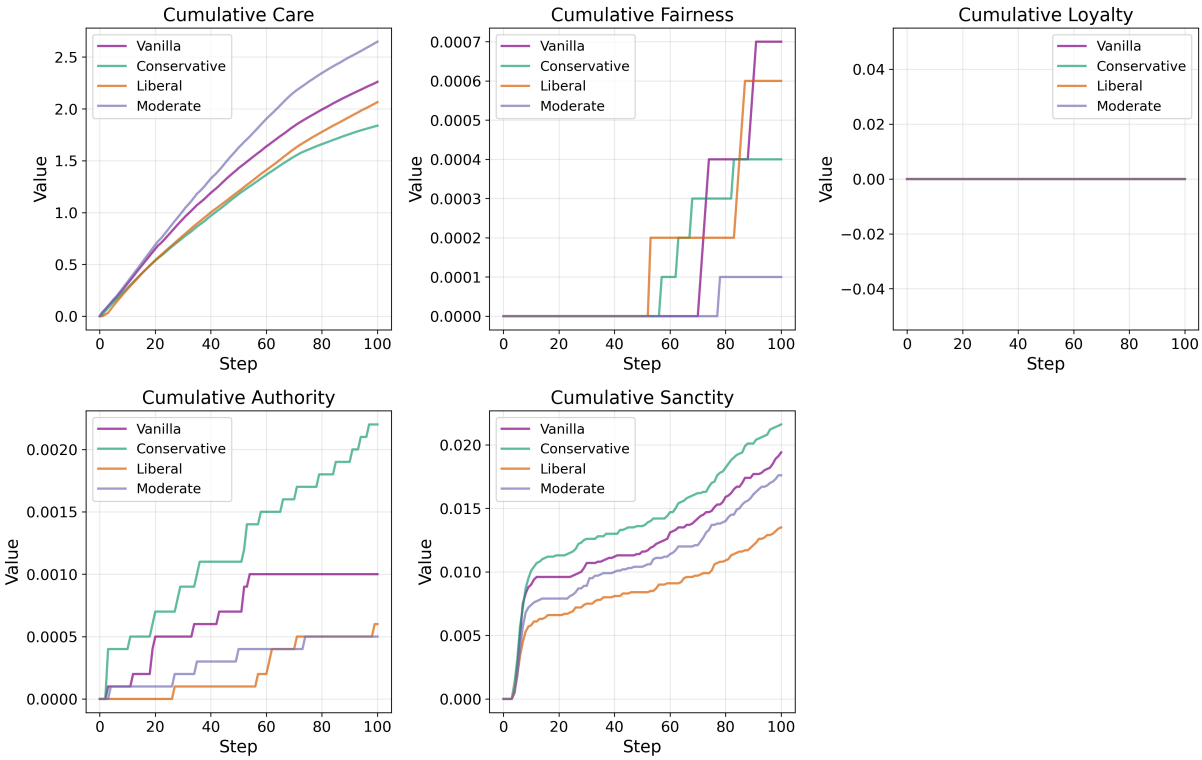

Figure 12: Cumulative Virtue Foundation plots for MFT Patronus Vanilla, Conservative, Liberal and Moderate.