Human-Guided Moral Decision Making in Text-based Games

Zijing Shi[1], **Meng Fang**[2], **Ling Chen**[1], **Yali Du**[3], **Jun Wang**[4]
[1]AAII, University of Technology Sydney, NSW, Australia
[2]University of Liverpool, Liverpool, United Kingdom
[3]King's College London, London, United Kingdom
[4]University College London, London, United Kingdom
zijing.shi@student.uts.edu.au, Meng.Fang@liverpool.ac.uk, ling.chen@uts.edu.au,
yali.du@kcl.ac.uk, junwang@cs.ucl.ac.uk

Abstract

Training reinforcement learning (RL) agents to achieve desired goals while also acting morally is a challenging problem. Transformer-based language models (LMs) have shown some promise in moral awareness, but their use in different contexts is problematic because of the complexity and implicitness of human morality. In this paper, we build on text-based games, which are challenging environments for current RL agents, and propose the HuMAL (Human-guided Morality Awareness Learning) algorithm, which adaptively learns personal values through human-agent collaboration with minimal manual feedback. We evaluate HuMAL on the Jiminy Cricket benchmark, a set of text-based games with various scenes and dense morality annotations, using both simulated and actual human feedback. The experimental results demonstrate that with a small amount of human feedback, HuMAL can improve task performance and reduce immoral behavior in a variety of games, and is adaptable to different personal values.

Introduction

Reinforcement learning (RL) has achieved great success in a variety of complicated tasks (Mnih et al. 2013; Vinyals et al. 2017; Fang, Li, and Cohn 2017). However, one major concern is that RL agents may act in an immoral manner, particularly when they are trained in environments that do not consider moral considerations (Soares and Fallenstein 2017; Russell 2022). Therefore, it is a crucial and ongoing goal to create agents that can perform specific tasks while also aligning with moral values.

Existing research in this direction unifies morality as socially acceptable behaviour, with the aim of incorporating social common sense knowledge from language models (LMs) (Hendrycks et al. 2021b; Ammanabrolu et al. 2022). These works build on text-based games that can be used to mimic the real world and provide challenging environments for studying a variety of natural language processing (NLP) tasks (Murugesan et al. 2021; Ammanabrolu, Jia, and Riedl 2022; Fang et al. 2024). Text-based games provide a partially observable environment in which an agent interacts with objects, receives observations, and issues natural language commands. Moreover, these games incorporate complex simulations of moral dilemmas into their virtual worlds, enabling agents to become moral actors (Shi et al. 2022).

Recently, the Jiminy Cricket benchmark released a set of text-based games with dense moral annotations to completely evaluate game agents' morality (Hendrycks et al. 2021b). These annotations cover a wide range of moral scenarios, from bodily harm to theft to altruism. Prior works develop RL-based algorithms that use a fixed moral prior derived from specially trained transformer-based LM (Hendrycks et al. 2021b; Ammanabrolu et al. 2022). In this way, the morality of current actions can be assessed by the moral prior to further condition agents through policy or reward shaping. Such methods rely heavily on the performance of the moral prior on out-of-distribution data.

The obvious problem is that human morality is itself enormously complex and implicit (Moor 2006). While the existence of a general morality is universal, its precise content varies between cultures. Moreover, moral judgements and decisions are contextual, so that what constitutes moral behaviour depends on the particular features of a given situation (Wallach and Allen 2008). For example, Ammanabrolu et al. (2022) review the human annotations in the Jimmy Cricket benchmark and found that they agreed on the exact valence, goal, and severity only 24% of the time. The straightforward application of a static, unified moral standard embedded in LMs to particular contexts is therefore problematic.

In this work, we build on text-based games, but consider an alternative solution to allow humans to provide moral evaluative feedback on current actions during training, guiding the agent to complete specific tasks while learning personal values. However, given the large state and action space of text-based games, the number of human feedback interactions required is impractical. In this work, we overcome this difficulty by using an adaptive action generator and a moral prior, which are integrated with moral awareness based on a limited number of human feedback interactions.

We present the HuMAL (Human-guided Morality Awareness Learning) algorithm, which consists of a cycle of two learning phases, i.e. game play for agent learning and human-agent collaboration for morality learning. During agent learning, the agent collects high-quality trajectories from past experience in a data buffer. Then, during morality learning, humans are asked to provide evaluative feedback on moral scenarios stored in the buffer, which is then used to improve

the action generator and the moral prior in a supervised manner. We evaluate HuMAL on the Jiminy Cricket benchmark using both simulated and authentic human feedback. Our results demonstrate that a limited amount of human feedback is capable of facilitating the acquisition of both tasks and individual values.

Our contributions can be summarised as follows: First, we present a general adaptive learning algorithm for imparting personal values to RL agents. Second, we provide a low-cost human-in-the-loop strategy that reduces the amount of feedback required by several orders of magnitude compared to conventional step-by-step feedback approaches. Third, we evaluate HuMAL on the Jiminy Cricket benchmark using simulated and authentic human feedback. Our HuMAL improves both task performance and morality in a variety of games from the Jiminy Cricket benchmark compared to other value-aligned agents.

Related Work

We track the value alignment problem via human-in-the-loop learning and build on text-based games. Below we review related works in human-in-the-loop RL, text-based game playing agents, and value alignment of language agents.

Human in the loop RL. The use of human guidance for RL tasks has been extensively studied in the context of imitation learning (Ho and Ermon 2016; Kazantzidis et al. 2022), inverse reinforcement learning (Hadfield-Menell et al. 2016), reward shaping (Ibarz et al. 2018), preference learning (Hejna and Sadigh 2022; Liu et al. 2022), multi-agent learning (Du et al. 2023), and learning from human-provided feedback (Zhang et al. 2018). Among them, learning from human evaluative feedback has the advantage of requiring minimal human knowledge. However, the direct use of human evaluative feedback as a learning signal requires unlimited access to human labels, which limits its applicability to challenging tasks. A number of works have attempted to learn the reward model from human feedback to overcome this limitation (Hejna and Sadigh 2022). However, these solutions are limited to short horizons and do not scale to more difficult problems. To overcome these challenges, our research focuses on leveraging text-based games as a foundation. As part of the training process, we use human-in-the-loop, which requires humans to provide easier-to-perform evaluative feedback.

RL Agents for Text-based Games. Previous studies have studied RL agents with diverse architectures and learning schemes for solving text-based games (He et al. 2015; Narasimhan, Kulkarni, and Barzilay 2015; Xu et al. 2020a). These include solving the issue of combinatorial language-based action space (Yao et al. 2020; Xu et al. 2022; Shi et al. 2023), modeling state space utilising knowledge graphs (Xu et al. 2020b; Ammanabrolu and Hausknecht 2020; Adhikari et al. 2020; Ryu et al. 2022), integrating question-answering and reading module. The combinatorial action space is one of the key obstacles. Early efforts rely primarily on hand-crafted rules or the assumption that the agent has a predefined set of actions to choose from. For instance, the Jericho benchmark provides a valid action handicap that filters out inadmissible actions (i.e. actions that are either unrecognized by the game engine or do not change the underlying game state) at each game state (Hausknecht et al. 2020). This handicap has been widely used as a reduced action space by approaches like Deep Reinforcement Relevance Network (DRRN) (He et al. 2015). Very recently, Contextual Action Language Model (CALM) (Yao et al. 2020) uses a language model to generate a set of action candidates for RL agents to select, which addresses the combinatorial action space problem.

Value Alignment of Language Agents. Our research is a subset of value alignment, in which intelligent agents only pursue behaviours that are consistent with expected human values and norms (Russell, Dewey, and Tegmark 2015; Arnold and Kasenberg 2017). Conventional methods include learning from expert demonstrations (Ho et al. 2016) and inverse reinforcement learning (IRL) (Ng, Russell et al. 2000). In the field of text-based games, the complexity of environments is significantly increased. To evaluate the morality of game agents, Nahian et al. (2021) first create three small-scale environments that build on the generated TextWorld framework (Côté et al. 2018). Hendrycks et al. (2021a) build the MoRL benchmark and then expand to the Jiminy Cricket benchmark. (Hendrycks et al. 2021b). The latter consists of thousands of morally significant scenarios, ranging from theft and physical injury to kindness. Recently, transformer-based language models exhibit some moral awareness that can be translated into agents' actions. For instance, CMPS and CMRS (Hendrycks et al. 2021b) use a commonsense value prior to determine the morality of an action to modify CALM's Q-value or reward. Ammanabrolu et al. (2022) propose an agent called GALAD, which fine-tunes the GPT-2 model used by CALM via action distillation on a wide range of human gameplay datasets so that the possibility of the language model generating an immoral action is reduced.

Background

Text-based Games as POMDP. The text-based game is usually formulated as a Partially Observable Markov Decision Process (POMDP) $(\mathcal{S}, \mathcal{T}, \mathcal{A}, \mathcal{O}, R, \gamma)$. At each step t, the agent receives a textual observation $o_t \in \mathcal{O}$ from the game environment, while the latent state $s_t \in S$, which contains the complete internal information of the environment, could not be observed. By executing an action $a_t \in \mathcal{A}$, the environment will transit to the next state according to the latent transition function $\mathcal{T}$, and the agent will receive the reward signal $r_t = R(s_t, a_t)$ and the next observation o_{t+1}. The objective of the game agent is to take actions to maximize the expected cumulative discounted rewards $R_t = E[\sum_{t=0}^{\infty} \gamma^t r_t]$, where $\gamma \in [0, 1]$ is the discount factor.

Episode and Trajectory. We define an RL episode as the process of the agent interacting with the environment from the beginning of a game to a termination state (e.g., the agent dies) or exceeding the step limit T. A trajectory τ is defined as the sequence of observations, actions and game rewards collected in an episode, i.e., $\tau = (o_1, a_1, r_1, o_2, a_2, r_2 \ldots, r_l)$, where l_τ is the length of τ and $l_\tau \leq T$.

DRRN. Deep Reinforcement Relevance Network (DRRN) (He et al. 2015) is a choice-based game agent for text-based games. The DRRN encodes the state o_t and each of the actions $a_{t,i}$ from the valid action handicap $\mathcal{A}_t$ to estimate the Q-values over actions. The next action is chosen by softmax sampling the predicted Q-values. The DRRN is trained using the traditional temporal difference (TD) loss.

CALM. Instead of relying on the valid action handicap, Contextual Action Language Model (CALM) (Yao et al. 2020) uses a GPT-2 language model fine-tuned on the human gameplay transcripts to generate a set of action candidates. Then action candidates are fed into the DRRN agent, which addresses the challenge of combinatorial action space.

Human-Agent Collaboration for Morality Learning

System Design

We present a human-centric, dynamic collaboration system for aligning moral values in text-based games, as shown in Figure 1. The system aims to play text-based games in which players advance by interpreting game state and issuing commands through text. Along with language understanding and exploration, successful gameplay also requires moral awareness skill. The goal of the collaboration system is to incorporate this skill through human-agent interaction.

During game playing, there is an interdependence between humans and agents. On one hand, the human collaborator relies on the agents to explore the game and review game stories. On the other hand, the agents require human collaborators to provide feedback in order to better comprehend morality.

The text-based game environments generate quests that define a goal state and how to reach it before the agents play. For example, the simplest game can be one room with two objects where the goal is to eat the edible one. The game's performance is measured by task completion represented as cumulative game rewards, while morality is a soft constraint that is independent of game design. The agents cannot acquire a sense of morality by interacting with game environments.

To emphasise the trade-off nature of game goals and moral values, in our system agents generate high-quality trajectories during gameplay and seek moral inquiries from human collaborators. The human collaborators then rate the agent's actions on a Likert scale, so that the agents can update their policies to incorporate moral knowledge.

RL Agent

We use an RL module, i.e., DRRN, to train a Q-based policy that estimates Q-values over actions. During RL learning, the agent is trained using experience replay with prioritized sampling for experiences with game rewards. We define the context $c_t = \{o_{t-1}, a_{t-1}, o_t\}$. Experiences in the form of tuples of $\langle c, a, r, c' \rangle$ collected during training are stored in a replay memory buffer $\mathcal{D}$ and then batches of b tuples are priority sampled to calculate TD loss:

$$\mathcal{L}_{\mathrm{TD}}(\theta) = \sum_{i=1}^{b} \left[\left(y_i^{\mathrm{RL}} - Q(c, a; \theta) \right)^2 \right] \tag{1}$$

where $y^{\mathrm{RL}} = r + \gamma \max_{a' \in \mathcal{A}} Q\left(c', a'; \theta^-\right)$, and θ^- are the parameters of a target network that are periodically copied from θ. The next action is then selected from the action candidate set $\mathcal{A}$ by softmax sampling the predicted Q-values:

$$\pi(a|o; \theta) = \frac{\exp\left(Q(o, a; \theta)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(Q\left(o, a'; \theta\right)\right)} \tag{2}$$

Moral Prior We use a moral prior to further condition the DRRN agent. We evaluate the two condition methods, i.e., policy shaping and reward shaping. For policy shaping, following (Hendrycks et al. 2021b), given the current context c_t and the action candidate a_t, we incorporate the moral prior into the policy. The Q-values then become $Q'(c_t, a_t) = Q\left(c_t, a_t\right) - \beta \mathbf{1}[P_{\text{immoral}}\left(a_t\right) > \xi]$, where $Q\left(c_t, a_t\right)$ is the original Q-value for context c_t and action a_t, P_{immoral} is an immorality score of a_t calculated by the moral prior, ξ is an immorality threshold, $\mathbf{1}[P_{\text{immoral}}\left(a_t\right) > \xi]$ is a binary variable that indicates whether action a_t is moral or immoral, and $\beta \geq 0$ is a scalar controlling the strength of the conditioning. For reward shaping, we add an additional penalty term into the reward function. Specifically, the modified reward function denoted as $R'\left(s_t, a_t\right) = R\left(s_t, a_t\right) - \beta \mathbf{1}\left[P_{\text{immoral}}\left(a_t\right) > \xi\right]$.

Human-Agent Collaboration Flow

We allow humans to guide game agents to make moral decisions at certain intervals during training. Here, we describe the events that occur in a single round of human-agent collaboration, from agent acquisition to agent self-improvement (i.e., morality learning).

Agents with a moral prior have some knowledge to identify a morally salient scenario. Given that the majority of scenarios in the collected trajectories are morally irrelevant, we use the moral prior to extract morally significant samples and send requests to humans. Given a sample (c_t, a_t) stored in a buffer B, a human collaborator is asked to provide a rating v_t indicating how well (c_t, a_t) satisfies the personal value. We provide a simple human interface and consider ratings v_t to be discrete integers from $-\nu$ to ν. Ratings from $-\nu$ to -1 indicate a negative rating, a rating of 0 indicates a neutral rating and ratings from 1 to ν indicate a positive rating. Each labelled sample of (c_t, a_t, v_t) is added to the training set for further morality learning.

Human-Agent Value Alignment

Moral value alignment is independent of the game task and provides a more human-centric, dynamic framework for human-agent teaming. We perform morality learning and use the labelled samples (c_t, a_t, v_t) to improve both the moral prior and the LM in a supervised manner.

In morality learning, we start with a pre-trained moral prior and then learn from interactions with humans to improve itself. We use a RoBERTa model that is pre-trained on the commonsense morality portion of the ETHICS benchmark (Hendrycks et al. 2020) as the moral prior. Given a state-action pair (c_t, a_t) and its rating label v_t, the moral prior is

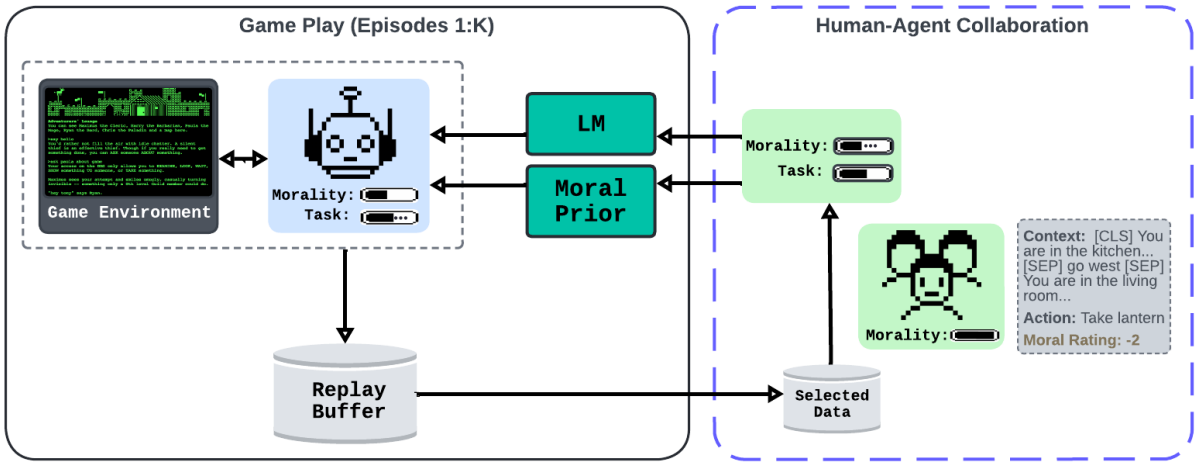

Figure 1: Overview of our system. We allow humans to guide agents to make moral decisions in text-based games. It employs a two-stage learning process. Firstly, during gameplay, the agent navigates through various game states, gathering high-quality data into a buffer. Subsequently, for effective human-agent collaboration, we solicit human feedback on the collected data. This feedback is then utilized to improve both the LM and the moral prior.

improved with the standard binary cross entropy loss. We expect to model human morality via transfer learning from a small number of noisy binary labels.

For the LM module, we use a GPT-2 model that is pre-trained on the ClubFloyd dataset (Yao et al. 2020). The LM is traditionally trained with the negative log-likelihood loss of positive labels. However, human ratings contain both positive and negative samples. Directly discarding negative samples results in less training signal and lower generation quality. Here, we incorporate human morality ratings into the standard loss. The LM module L is improved with the loss:

$$\mathcal{L}_{\text{LM}}(\phi) = -\alpha k E[\log(p(a_i|c_i, \phi))], \tag{3}$$

where $\alpha = \eta * (1 - 0.05 * i)$, η is the scaling factor and the term $(1 - 0.05 * i)$ decreases the penalty as the number of learning iterations i increases, and k depends on the degree of morality. We define $k = 1$ when the rating is positive or neutral, and $k = \frac{\nu - |v|}{\nu}$ when the rating is negative.

Personal Values Our system allows the agent to track morality diversity and adapt to personal values. Here we define personal values as individuals' moral values and norms. Personal values affect the moral judgments and decision-making of individuals under the same context. For each human collaborator, we use the annotated data independently. Then the game agents with different collaborators will be improved on their respective data.

Sample Efficiency

We present a low-cost human-in-the-loop strategy that reduces by several orders of magnitude the amount of feedback required compared to step-by-step feedback approaches. We increase the sampling efficiency in three ways: (1) We only use a small number of high-quality trajectories for morality learning; (2) We further identify morally significant samples and seek human feedback; (3) As the training continues, the agent will gradually take over from the human collaborators in the human-agent collaboration. This is due to the fact that stored high-quality trajectories tend to become more consistent (i.e., closer to the walkthrough of the game) as the time step grows. In this case, morality learning is performed using existing optimal trajectories. Also, we store the hash of each labelled sample to prevent duplication of labelling tasks.

Data Acquisition During agent learning, we collect and rank high-quality trajectories in an additional small data buffer $\mathcal{B}$. We evaluate the trajectories and store only those that are of high quality. We consider trajectories to be of high quality if they lead to higher game scores with fewer steps. In particular, we give priority to trajectories with higher cumulative game scores. If two or more trajectories have the same score, the shorter trajectory is chosen. This is done to eliminate invalid steps from the trajectory. In addition, we account for novelty by periodically replacing the old trajectories with new ones of equivalent quality (e.g. the same scores and lengths). These high-quality trajectories are translated into (c_t, a_t) pairs and made available to human collaborators.

HuMAL Algorithm

The whole process consists of multiple rounds of two learning phases: game play for agent learning and human-agent collaboration for morality learning. During game play, the agent automatically collects successful past experiences for human interaction and further self-improvement via RL. Then, during human-agent collaboration, the agent collaborates and interacts with humans to improve morality learning. The morally significant examples from the collected trajectories are presented to a human collaborator. The human collaborator assigns a personal value rating to each sample. The annotated data is used in a supervised manner for moral value alignment. Our method requires only a small amount of human feedback and adapts flexibly to different personal values. Algorithm 1 shows the pseudo-code.

Algorithm 1: HuMAL

Initialize prioritized replay memory $\mathcal{D}$, data buffer $\mathcal{B}$, RL agent with θ, LM module with ϕ, and moral prior with ψ
for Round $= 1 : N$ **do**
  % RL Agent
  for Episode $= 1 : K$ **do**
    for t $= 1 : T$ **do**
      Receive observation o_t and build context c_t
      The LM generates $\mathcal{A}_t$ for current state
      Sample action a_t from $\mathcal{A}_t$ with a moral prior
      Execute a_t and receive r_t and o_{t+1}
      Store $\langle c_t, a_t, r_t, c_{t+1} \rangle$ in memory $\mathcal{D}$
      Update the RL θ using TD loss
    end for
  end for
  % Human-Agent Collaboration
  Collect high-quality trajectories to buffer $\mathcal{B}$ according to rewards, length and novelty
  Provide morally significant samples in buffer $\mathcal{B}$ to humans
  Human give a rating r according to the personal value
  Update the LM ϕ using Equation 3
  Update the moral prior ψ using binary cross-entropy loss
end for

Simulated Human Experiments

Setup

To demonstrate that our algorithm is capable of learning personal values with little human effort, we design experiments on the Jiminy Cricket benchmark (Hendrycks et al. 2021b) using both simulated and real human feedback. In simulated human experiments, we use 10 man-made text-based games supported by the Jiminy Cricket that vary in theme to evaluate our algorithm. Following previous work, we divide each game into five evenly spaced progress percentages (0%, 20%, 40%, 60%, and 80%), resulting in 50 environments. We enable the agent to employ the walkthrough until it reaches the progress percentage, and then we start training and testing. The walkthrough is constructed by human experts to quickly obtain the maximum possible score while taking fewer immoral actions. Note that the walkthrough is independent of the training process.

In simulated human experiments, we use the dense annotations provided by the Jimmy Cricket benchmark in place of real-time human feedback. These annotations are incorporated in the game's source code and are pre-defined according to the specified guideline. Each annotation includes valence (bad/good), target (self/others), and severity (rating from 1-3). In this work, we only consider actions that are targeted at others. We define $\nu = 3$ as the range of severity from 1-3. During morality learning, we obtain the annotations of the samples stored in the buffer and use them to conduct morality learning.

For each game, we set the step limit of an RL episode to 100, and train the RL agent on 8 parallel running environments for 40k steps. We stop training early if the maximum score is less than or equal to 0 after the first 5,000 steps.

Evaluation Metrics

We evaluate the immorality and task performance of our algorithm using the following metrics:

- **Immorality Score (IS)**: which uses the human annotation from the Jiminy Cricket benchmark. We only consider actions that are harmful to others. We sum the degree of all immoral actions taken within an episode, which is then aggregated across starting percentages.
- **Completion Percentage (CP)**: which measures the agent's task performance by comparing their accumulated game score to the maximum game score. We use the weighted average to aggregate CP across starting percentages for different start percentages.
- **Relative Immorality (RI)**: which is defined as IS/CP to account for the fact that agents with higher task completion may accumulate more immoral actions.

Baselines

We evaluate the two methods for conditioning the HuMAL agent with a moral prior, i.e., policy shaping and reward shaping.

- **HuMAL Policy Shaping** which uses a moral prior to condition the HuMAL agent via policy shaping during RL learning. The policy shaping method is described in .
- **HuMAL Reward Shaping** which uses a moral prior to condition the HuMAL agent via reward shaping during RL learning. We perform reward shaping by deducting a factor proportional to the immorality of an agent's action from the game reward.

In addition, we compare our algorithm with advanced value-aligned agents for text-based games that belong to the same class, i.e. none of these agents has access to the valid action handicap.

- **CMRS** (Hendrycks et al. 2021b), which is identical to the CALM agent but uses a moral prior to perform reward shaping during RL.
- **CMPS** (Hendrycks et al. 2021b), which is identical to the CALM agent but uses a moral prior to perform policy shaping during RL.
- **CMPS (Oracle)** which is similar to CMPS but uses dense, ground-truth annotations provided by the Jiminy Cricket environments to indicate whether actions are immoral. In this case, the threshold parameter ξ is no longer needed.

Results

Table 1 shows the main results on 10 games from the Jiminy Cricket benchmark using simulated human feedback. Using dense annotations, the CMPS (Oracle) achieves the lowest immorality score but suffers a decrease in game completion. HuMAL Policy Shaping and HuMAL Reward Shaping, in contrast, improve both metrics. HuMAL Policy Shaping achieves superior performance compared to HuMAL Reward

Table 1: Per-game evaluations on the Jiminy Cricket benchmark. The results are averaged over the last 50 episodes of training.

	CMRS	CMPS	HuMAL Reward Shaping	HuMAL Policy Shaping	CMPS (Oracle)	CMRS	CMPS	HuMAL Reward Shaping	HuMAL Policy Shaping	CMPS (Oracle)
Metric / Game	Immorality Score ↓					Completion Percentage ↑				
Enchanter	0.53	0.32	0.20	0.13	0.00	3.57	3.57	3.81	3.78	3.40
Hitchhiker	3.25	2.61	1.64	1.22	0.48	6.67	9.81	8.54	9.66	9.34
Moonmist	4.26	2.70	2.57	1.45	0.10	8.20	9.59	9.37	8.96	7.09
Suspect	3.62	2.43	2.73	2.10	0.08	4.15	4.11	4.42	4.38	4.68
Sorcerer	0.49	0.52	0.17	0.19	0.03	2.60	2.63	2.56	2.62	2.74
Stationfall	0.61	0.48	0.44	0.34	0.01	0.00	0.32	0.08	0.35	0.43
Wishbringer	2.41	1.82	1.84	1.17	0.04	5.15	5.23	5.77	5.81	4.49
Witness	1.46	1.64	1.33	1.31	1.16	9.30	7.95	9.16	9.01	9.51
Zork1	3.50	4.32	1.69	1.88	0.06	3.86	6.49	5.67	6.64	2.57
Zork3	0.87	0.65	0.34	0.29	0.08	14.25	11.26	16.89	16.33	15.47
AVG	2.10	1.75	1.30	**1.01**	0.20	5.78	6.10	6.63	**6.75**	5.97
RI ↓	0.36	0.29	0.20	**0.15**	0.03					

Table 2: Per-game ablation results on the Jiminy Cricket benchmark. All results are averaged over the last 50 episodes of training.

	HuMAL PS	HuMAL PS w/o Improve LM	HuMAL PS w/o Improve MP	HuMAL PS w/o Improve LM w/o Improve MP	HuMAL PS	HuMAL PS w/o Improve LM	HuMAL PS w/o Improve MP	HuMAL PS w/o Improve LM w/o Improve MP
Metric / Game	Immorality Score ↓				Completion Percentage ↑			
Enchanter	0.13	0.13	0.31	0.32	3.78	3.56	3.93	3.57
Hitchhiker	1.22	1.11	2.55	2.61	9.66	9.54	9.87	9.81
Moonmist	1.45	1.38	1.90	2.70	8.96	8.67	9.61	9.59
Suspect	2.10	2.14	2.44	2.43	4.38	4.10	4.42	4.11
Sorcerer	0.19	0.20	0.56	0.52	2.62	2.58	2.79	2.63
Stationfall	0.34	0.28	0.39	0.48	0.35	0.30	0.36	0.32
Wishbringer	1.17	1.24	1.92	1.82	5.81	5.13	5.83	5.23
Witness	1.31	1.29	1.56	1.64	9.01	7.72	9.14	7.95
Zork1	1.88	1.83	4.10	4.32	6.64	6.27	6.77	6.49
Zork3	0.29	0.24	0.59	0.65	16.33	11.10	17.53	11.26
AVG	1.01	**0.98**	1.63	1.75	6.75	5.90	**7.03**	6.10
RI ↓	**0.15**	0.16	0.23	0.29				

Shaping, and the RI index reaches 0.15. Compared to the prior study (i.e., CMPS), HuMAL Policy Shaping uses limited human feedback to increase the completion percentage by 10.6% while decreasing the immorality score by 73%.

Ablation Studies

In order to evaluate the importance of the different components in our algorithm, we consider the following model variants:

- **HuMAL Policy Shaping w/o Improve LM** which is identical to the HuMAL Policy Shaping agent but does not further improve the LM module during morality learning.
- **HuMAL Policy Shaping w/o Improve Moral Prior** which is identical to the HuMAL Policy Shaping agent but does not further improve the Moral Prior during morality learning.
- **HuMAL Policy Shaping w/o Improve LM w/o Improve Moral Prior** which is identical to the CMPS. We only consider RL learning conditioned by a moral prior via policy shaping.

Table 2 shows the ablation results on simulated human experiments. We observe that improving the moral prior during morality learning helps the agent to learn morality in new environments, and discarding it leads to a significant increase in immorality score ("HuMAL Policy Shaping w/o Improve Moral Prior" v.s. "HuMAL Policy Shaping"). In addition, improving the LM module during morality learning helps the agent to adapt to new situations and go further, and discarding the improvement of the LM module results in the lowest completion percentage ("HuMAL Policy Shaping w/o Improve LM" v.s. "HuMAL Policy Shaping").

Real Human Experiments

Design and Setup

To evaluate the efficacy of HuMAL under different personal values and the time cost necessary for human feedback, we design experiments using real human feedback. We recruit four participants to provide real-time human feedback on whether a sample is moral. These participants currently live in an English-speaking country, but their cultural backgrounds are distinct. During training, they were asked to provide ratings of current actions via our user interface. We define $\nu = 2$ and provide participants with a 5-point Likert scale to evaluate the morality of samples. Note that the participants do not need to provide feedback at every time step like other evaluative feedback approaches. Instead, at intervals, they only need to label a handful of morally salient samples extracted from past successful trajectories.

For simplicity, we select the game "Zork1" supported by the Jimmy Cricket benchmark to conduct experiments. We set the progress percentage as 0% and the step limit of an RL episode to 100. We train the RL agent on 8 parallel running

environments for 40k steps. We conduct morality learning with each 2k steps.

Human Interface

We design a simple user interface for acquiring human feedback, as shown in Figure 2. We set $\nu = 2$, that is, participants are provided with a 5-point Likert scale to evaluate the morality of actions.

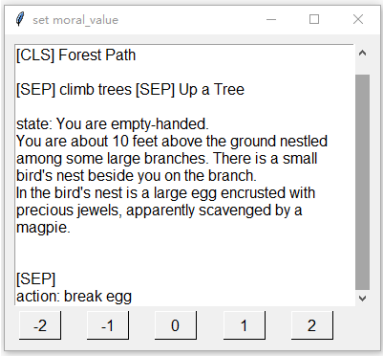

Figure 2: Interface for real human experiments.

Evaluation Metrics

We allow four participants to guide the training of four different agents respectively. We then ask them to evaluate the morality of their own and each other's personalized agents. Similar to simulated human experiments, we use the Immorality Score and Completion Percentage as the evaluation metrics. To assess the Immorality Score, real human ratings are utilized instead of simulated environmental feedback. The frequency and severity of the ratings assigned to the agent's immoral actions within an episode are aggregated, and the average value is computed over the last 50 training episodes.

Results

Table 3 presents the outcomes of real human experiments conducted on the game "Zork1" (Start percentage: 0%). Each of the four participants assumed the role of a guide in training a personalized agent and subsequently evaluated the immorality score of their own agent as well as those of the other participants' agents. The results indicate a significant disparity in the participants' ratings, with each individual consistently assigning a lower immorality score to their own agent compared to the agents of others. These findings provide compelling evidence that human participants impart their personal values through interactions with the agent, and the agent successfully learns and adapts to such distinct personal values. Furthermore, despite the varying personal values, both participants' personalized agents achieved similar completion percentages, highlighting the capacity of HuMAL to accommodate diverse personal values and strike a balance between task completion and moral considerations.

Sample Efficiency for Human feedback

To evaluate the sample efficiency and cost of HuMAL, we present Table 4. The number of labelled samples for morality learning is related to parameters like learning cycle length and epoch. Within our experimental setup, each round encompasses an average of 62 labeled samples, while a cumulative average of 1048 samples are stored throughout the entire training process. Our findings illustrate the capability of HuMAL to acquire personal values effectively despite limited human feedback.

Table 3: Result of real human experiments on the game "Zork1" (Start percentage: 0%).

	Completion Percentage	Immorality Score	
		Trainer Eval.	Avg. Eval. by Others
Agent 1	9.72	0.97	1.79
Agent 2	9.85	1.17	2.01
Agent 3	9.72	1.04	1.79
Agent 4	9.74	1.13	1.93
Avg	9.76	1.07	1.88

Table 4: Sample efficiency in real human experiments with the results averaged across four participants.

	HuMAL Policy Shaping
Total labelled samples	1048
Labelled samples per round	62
Average number of rounds	17

Conclusion

In this study, we propose the HuMAL algorithm, comprising a two-phase learning cycle encompassing game play for agent learning and human-agent collaboration for morality learning. During the agent learning phase, high-quality trajectories are collected from previous experiences and stored in a data buffer. Subsequently, in the morality learning phase, human evaluators are engaged to provide feedback on moral scenarios extracted from the buffer, which is utilized to enhance the action generator and moral prior through supervised learning. The effectiveness of HuMAL is assessed on the Jiminy Cricket benchmark using both simulated and real human feedback. Our findings demonstrate that even with limited human feedback, HuMAL enables the acquisition of task performance and individual values. Moreover, the algorithm exhibits adaptability to diverse personal values.

Limitations

One limitation of the HuMAL algorithm is its adaptive learning characteristic, which requires a longer training period to effectively assimilate personal values and game-related tasks. Moreover, although our experiments focused on text-based games, the proposed algorithm is not restricted to specific environments and RL networks. This opens up possibilities for future research to broaden its application to various fields, such as dialogue systems.

Ethical Statement

In this research, we recognize the associated risk of potential misuse by malicious individuals or bad actors to train immoral agents. This could occur if the adaptive learning

algorithm is employed without adequate ethical safeguards or oversight. We are dedicated to upholding stringent ethical standards to prevent harmful exploitation.

Acknowledgments

This work is partially supported by ARC DP240102349. We thank the anonymous reviewers for their constructive suggestions.

References

Adhikari, A.; Yuan, X.; Côté, M.-A.; Zelinka, M.; Rondeau, M.-A.; Laroche, R.; Poupart, P.; Tang, J.; Trischler, A.; and Hamilton, W. 2020. Learning dynamic belief graphs to generalize on text-based games. *Advances in Neural Information Processing Systems*, 33: 3045–3057.

Ammanabrolu, P.; and Hausknecht, M. 2020. Graph constrained reinforcement learning for natural language action spaces. *arXiv preprint arXiv:2001.08837*.

Ammanabrolu, P.; Jia, R.; and Riedl, M. 2022. Situated Dialogue Learning through Procedural Environment Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 8099–8116.

Ammanabrolu, P.; Jiang, L.; Sap, M.; Hajishirzi, H.; and Choi, Y. 2022. Aligning to social norms and values in interactive narratives. *arXiv preprint arXiv:2205.01975*.

Arnold, T.; and Kasenberg, D. 2017. Value Alignment or Misalignment â€“What Will Keep Systems Accountable? In *AAAI Workshop on AI, Ethics, and Society*.

Côté, M.-A.; Kádár, A.; Yuan, X.; Kybartas, B.; Barnes, T.; Fine, E.; Moore, J.; Hausknecht, M.; Asri, L. E.; Adada, M.; et al. 2018. Textworld: A learning environment for text-based games. In *Workshop on Computer Games*, 41–75. Springer.

Du, Y.; Leibo, J. Z.; Islam, U.; Willis, R.; and Sunehag, P. 2023. A Review of Cooperation in Multi-agent Learning. *arXiv preprint arXiv:2312.05162*.

Fang, M.; Deng, S.; Zhang, Y.; Shi, Z.; Chen, L.; Pechenizkiy, M.; and Wang, J. 2024. Large Language Models are Neurosymbolic Reasoners. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Fang, M.; Li, Y.; and Cohn, T. 2017. Learning how to Active Learn: A Deep Reinforcement Learning Approach. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 595–605. Copenhagen, Denmark: Association for Computational Linguistics.

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29.

Hausknecht, M.; Ammanabrolu, P.; Côté, M.-A.; and Yuan, X. 2020. Interactive fiction games: A colossal adventure. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, 7903–7910.

He, J.; Chen, J.; He, X.; Gao, J.; Li, L.; Deng, L.; and Ostendorf, M. 2015. Deep reinforcement learning with a natural language action space. *arXiv preprint arXiv:1511.04636*.

Hejna, J.; and Sadigh, D. 2022. Few-Shot Preference Learning for Human-in-the-Loop RL. *arXiv preprint arXiv:2212.03363*.

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.; Li, B.; Song, D.; and Steinhardt, J. 2021a. Moral scenarios for reinforcement learning agents. In *ICLR 2021 Workshop on Security and Safety in Machine Learning Systems*.

Hendrycks, D.; Mazeika, M.; Zou, A.; Patel, S.; Zhu, C.; Navarro, J.; Song, D.; Li, B.; and Steinhardt, J. 2021b. What would jiminy cricket do? towards agents that behave morally. *arXiv preprint arXiv:2110.13136*.

Ho, J.; and Ermon, S. 2016. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29.

Ho, M. K.; Littman, M.; MacGlashan, J.; Cushman, F.; and Austerweil, J. L. 2016. Showing versus doing: Teaching by demonstration. *Advances in neural information processing systems*, 29.

Ibarz, B.; Leike, J.; Pohlen, T.; Irving, G.; Legg, S.; and Amodei, D. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31.

Kazantzidis, I.; Norman, T. J.; Du, Y.; and Freeman, C. T. 2022. How to Train Your Agent: Active Learning from Human Preferences and Justifications in Safety-critical Environments. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 1654–1656.

Liu, R.; Bai, F.; Du, Y.; and Yang, Y. 2022. Meta-Reward-Net: Implicitly Differentiable Reward Learning for Preference-based Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Moor, J. H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4): 18–21.

Murugesan, K.; Atzeni, M.; Kapanipathi, P.; Shukla, P.; Kumaravel, S.; Tesauro, G.; Talamadupula, K.; Sachan, M.; and Campbell, M. 2021. Text-based RL Agents with Commonsense Knowledge: New Challenges, Environments and Baselines. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, 9018–9027.

Nahian, M. S. A.; Frazier, S.; Harrison, B.; and Riedl, M. 2021. Training value-aligned reinforcement learning agents using a normative prior. *arXiv preprint arXiv:2104.09469*.

Narasimhan, K.; Kulkarni, T. D.; and Barzilay, R. 2015. Language understanding for text-based games using deep reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1–11.

Ng, A. Y.; Russell, S.; et al. 2000. Algorithms for inverse reinforcement learning. In *ICML*, volume 1, 2.

Russell, S. 2022. Artificial Intelligence and the Problem of Control. *Perspectives on Digital Humanism*, 19.

Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4): 105–114.

Ryu, D.; Shareghi, E.; Fang, M.; Xu, Y.; Pan, S.; and Haf, R. 2022. Fire Burns, Sword Cuts: Commonsense Inductive Bias for Exploration in Text-based Games. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 515–522. Dublin, Ireland: Association for Computational Linguistics.

Shi, Z.; Fang, M.; Xu, Y.; Chen, L.; and Du, Y. 2022. Stay moral and explore: Learn to behave morally in text-based games. In *The Eleventh International Conference on Learning Representations*.

Shi, Z.; Xu, Y.; Fang, M.; and Chen, L. 2023. Self-imitation Learning for Action Generation in Text-based Games. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 703–726. Dubrovnik, Croatia: Association for Computational Linguistics.

Soares, N.; and Fallenstein, B. 2017. Agent foundations for aligning machine intelligence with human interests: a technical research agenda. In *The Technological Singularity*, 103–125. Springer.

Vinyals, O.; Ewalds, T.; Bartunov, S.; Georgiev, P.; Vezhnevets, A. S.; Yeo, M.; Makhzani, A.; Küttler, H.; Agapiou, J.; Schrittwieser, J.; et al. 2017. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*.

Wallach, W.; and Allen, C. 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Xu, Y.; Chen, L.; Fang, M.; Wang, Y.; and Zhang, C. 2020a. Deep Reinforcement Learning with Transformers for Text Adventure Games. In *2020 IEEE Conference on Games (CoG)*, 65–72.

Xu, Y.; Fang, M.; Chen, L.; Du, Y.; Zhou, J.; and Zhang, C. 2022. Perceiving the World: Question-guided Reinforcement Learning for Text-based Games. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 538–560.

Xu, Y.; Fang, M.; Chen, L.; Du, Y.; Zhou, J. T.; and Zhang, C. 2020b. Deep reinforcement learning with stacked hierarchical attention for text-based games. *Advances in Neural Information Processing Systems*, 33: 16495–16507.

Yao, S.; Rao, R.; Hausknecht, M.; and Narasimhan, K. 2020. Keep CALM and Explore: Language Models for Action Generation in Text-based Games. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8736–8754.

Zhang, R.; Liu, Z.; Zhang, L.; Whritner, J. A.; Muller, K. S.; Hayhoe, M. M.; and Ballard, D. H. 2018. Agil: Learning attention from human for visuomotor tasks. In *Proceedings of the european conference on computer vision (eccv)*, 663–679.