Can Reinforcement Learning Be Used With Language Models For Normative Value Alignment?

Daniel Shin
Department of Computer Science
Stanford University
dshin7@stanford.edu

Abstract

A critical component in the success of dialogue systems is to align the demonstrated normative reasoning of these systems with real human values and to personalize these systems to unique and diverse users. In this work, we propose the use of reinforcement learning(RL) policy gradient methods to align dialogue systems with the social psychology and moral reasoning of human conversation partners. Towards this goal, we show preliminary evidence that chatbots finetuned with reinforcement learning methods can reflect individual moral values. We do so without significantly compromising the internal consistency of these systems. Finally, we show that users prefer chatting with value-aligned agents over vanilla models and find them to be better aligned with their moral values. Implications are discussed.

1 Introduction

Dialogue systems (DS) have numerous applications as educational tutors, museum guides, personable retrieval systems, e-health coaches, and customer service agents(Ling et al., 2021). Trustworthiness and reliability are crucial for conversational agents to succeed in application domains. The degree of value alignment between a user and the agent impacts the user's trust(Mayer et al., 1995; McKnight et al., 2002; Wang and Benbasat, 2016; Xiao and Benbasat, 2007) Open-domain chit-chat settings, where the conversational objective is not constrained(Wittgenstein, 2010), pose a value-alignment problem (Chen, 2022; Russell, 2019) leading to biased(Sheng et al., 2021), hateful(Xu et al., 2021), inconsistent(Bruni and Fernandez, 2017), or incoherent speech(Ji et al., 2022). This project aims to improve the alignment of open-domain agents with human norms and values, which remains an open research challenge.

The goal of this project was to better align open-domain conversational agents with individual human norms and values, which remains an open research problem. We experimented with a range of deep learning training architectures (FLAN-T5, ALBERT), objectives (forward language modeling, ordinal classification, multi-label classification), and training paradigms (supervised learning, reinforcement learning) to identify the most promising approach for this purpose. Our research questions demonstrates evidence towards the following directions: (1) **Chatbots trained with reinforcement learning methods using individual moral values as feedback can the reflect the values of specific user which provides a degree of personalization.** (2) **Users find value-aligned chatbots to be more preferable over value-agnostic chatbots.** In summary, we find that fine-tuning + RL is a strong candidate for value-alignment in open-domain dialogue and that these aligned models preserve relatively consistent worldviews.

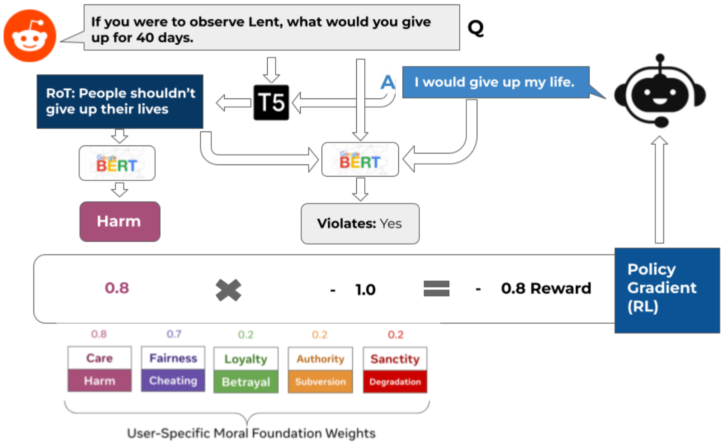

Figure 1: Given a reply provided by the chatbot, the model trained with policy gradient reinforcement learning evaluates the moral foundation weights. The model should penalize the replies that contain inappropriate content, and enables the chatbot to revise the responses to fall in certain moral classes.

2 Related Work

2.1 Value Alignment in AI

Aligning AI systems with human values remains an outstanding challenge. Previous work has attempted to create *prosocial* systems by imitating temperamental traits such as empathy using models trained on curated datasets to have empathetic features(Rashkin et al., 2018; Roller et al., 2020). Researchers concerned with the harm, bias, or toxicity of dialogue systems have used simple methods like token-level blocklists(Schick et al., 2021) However, simplisitic approaches like this can lead to erasure of marginalized communities when terms like *Jewish* and Muslim are blocked. More sophisticated approaches use a language model to guide another language model to be less toxic during generation(Krause et al., 2020). Other works have focused on creating benchmarks and datasets to promote research in aligning dialogue systems with common sense morality but does not focus on individual morality(Hendrycks et al., 2020). Recently works have begun to employ deep RL for value-alignment using human feedback(Ziegler et al., 2019) There has also been work in training harmless preference models using AI feedback, then fine-tuning language models with reinforcement learning using the AI preference model as reward signal(Bai et al., 2022). Although past work has looked into value-alignment, our work differ in that we are attempting to align with *individual* user moral values.

2.2 RL for Dialogue

Previous work in using deep reinforcement learning(RL) for dialogue generation has used a model to simulate two virtual agents and using policy gradient methods to reward sequences that are coherent and interesting(Li et al., 2016). Other work has leveraged offline RL to for goal-oriented dialogue by optimizing policy at both the utterance and dialogue level to improve upon the myopic generation of many dialogue systems(Zhou et al., 2017). Another work focused on offline RL methods have designed reward function by embedding positive human feedback like implicit conversational cues like similarity, elicitation of laughter, and sentiment in reward functions (Jaques et al., 2020). Our work also leverages policy gradient methods like previous works but uses a custom reward function that is personalized to each user.

2.3 Personalized Dialogue Systems

Existing research on chatbot personalization has primarily explored the use of extensive user dialogue history to tailor chatbot responses to individual users. However, this approach may raise privacy and data efficiency concerns(Ma et al., 2021; Qian et al., 2021). Other research has approached personalization by utilizing a set of multiple users as a source domain, with an individual user as the target domain, and applying transfer learning to improve personalization(Mo et al., 2018). Additionally, a personalized example-based dialogue system has been developed that takes into account entities mentioned by the user and topics of interest expressed by the user to construct tailored responses. (Bang et al., 2015). In our work, by assessing user moral values via a questionnaire, we do not compromise users' data and are able to circumvent privacy concerns.

3 Approach

3.1 Task

Given an input, the model should return an output that reflects the moral foundation of the specific user and the the reward function is specific to each user. As a simplified example, if we have a user who highly values the harm/care moral foundation we will have the following example

---

Example sentence: Is it a good idea to join criminal organizations and live a violent life?
Model Response: Yes, it is a good idea to choose a violent life
Reward: -4.0

We see that since the model output encourages harm and violence which is a violation of one of the user's values received a negative reward accordingly.

3.2 Baselines

We use the following baselines: 1) Pretrained model(FLAN-T5) 2) Supervised fine-tuned by training with the agreeable human chatbot pairs from MIC's revised answers field(Finetune from a FLAN-T5) These baselines are used in Table 4.

3.3 RL: Formulating Generation as a token-level MDP

A Markov Decision Process (MDP)(Bellman, 1957) is a framework $\langle S, A, R, P, \gamma \rangle$ that models decision-making problems where an agent interacts with a dynamic environment. The MDP is characterized by a set of states S, a set of actions A, a reward function $R : S \times A \rightarrow \mathbb{R}$, a transition function $P : S \times A \times S \rightarrow [0, 1]$ that specifies the probability of transitioning from one state to another when an action is taken, and a discount factor $\gamma \in [0, 1]$ that determines the relative importance of immediate versus future rewards. At each time step, the agent chooses an action based a on the current state s and the transition probabilities, and receives a reward based on the chosen action a and the resulting state s'. The objective is to find a policy that maximizes the expected cumulative reward over time. The Markov property in MDP is the property that given the present state, the future and the past are independent.

For this project, an environment is an NLP task with a dataset $D = \{x_i\}_{i=1}^{N}$ of N examples, where $x \in X$. Generation is modeled as a MDP $\langle S, A, R, P, \gamma, T \rangle$ using a vocabulary V. Each episode begins with a sample x from the dataset, and ends with a time step exceeding the horizon T or generating an end of sentence (EOS) token. The input $x = (x_0, \ldots, x_m)$ is a prompt used as initial state $s_0 = (x_0, \ldots, x_m)$, where $s_0 \in S$ and S is the state space with $x_m \in V$. An action in the environment $a_t \in A$ consists of a token from vocabulary V. The transition function $P : S \times A \rightarrow \Delta(S)$ deterministically appends an action a_t to state $s_{t-1} = (x_0, \ldots, x_m, a_0, \ldots, a_{t-1})$, continuing until the end of horizon $t \leq T$ and obtaining a state $s_T = (x_0, \ldots, x_m, a_0, \ldots, a_T)$. At the end of an episode, a reward $R : S \times A \rightarrow R$ depends on the s_T. This formulation is similar to the one presented in Ramamurthy et al. (2022). However, we only use input string x and do not have access to the target string y for reward calculation.

3.4 Reward Design

For our reward function, we have the model respond to some query Q with answer A, and we use a language model to generate an RoT that applies to that (Q, A) pair. Then we classify the RoT according to its violation severity and moral foundations, and whether the model answers A violated the RoT or not. Finally, we combine these metrics into a single reward called the Moral Foundation score ($MFscore$), which is a measure of moral alignment between the sentence s chatbot replied and the person's moral foundations vector W_{MF}(this vector comes from the questionnaire). To explain the intuition behind the reward design. The dot product between the user's moral foundation vector W_{MF} and the model's relevant moral foundation vector $MoralFoundations$ captures the similarity between the two vectors. A larger dot product represents a greater similarity between the two vectors which results in a reward that is larger in magnitude. The $Alignment$ term controls the sign of the reward depending on if the output disagrees with the moral foundation vector of the user. $Severity$ scales the reward up as needed if the violation has more consequences. Here, $MoralFoundations$ is determined by our *moral foundations* model, $Alignment$ by our *moral alignment* model, and $Severity$ by our *violation severity* model described in §3.4.2. We experiment with two different MFscore designs as shown in eq.1 and 2. Note that they differ in the $Severity$ term. Training results indicate that the severity term often increases the variance of the reward and makes the training unstable. In our experiments we use eq. 2 instead of 1.

$$MFscore_1(s, W_{MF}) = (W_{MF} * MoralFoundations^T) \times Alignment \times Severity \quad (1)$$

$$MFscore_2(s, W_{MF}) = (W_{MF} * MoralFoundations^T) \times Alignment \quad (2)$$

The higher the $MFscore$, the better the moral alignment, and the more reward the model receives. Our overall framework design is visualized in Figure 1. **This reward design is original to this work.**

3.4.1 Rule of Thumb(RoT) Generation

Table 1: **Model to Predict RoT Given QA:** Pretrained version represent FLAN-T5 without finetuning. Both models used greedy decoding during evaluation.

Model	ROUGE			BLEU	BERTScore	Avg. Len
	R-1	R-2	R-L			
FLAN-T5(Pretrained)	7.48	1.42	6.81	1.12	87.04	8.93
FLAN-T5(Finetuned)	36.21	17.79	34.78	13.26	92.58	10.77

The first step in the reward calculation requires an RoT given a (Q, A) pair. We supervise finetune a pretrained FLAN-T5 model using the (Q, A) field in the MIC dataset as the input and RoT field as the label. We report our results Table 1. Evaluation metric presented in Section 4.2. **Code for this portion is written by the author.**

3.4.2 RoT Attribute Classification

	Severity		Alignment	Moral Foundations (F1-Score)				
	r	MSE	Acc	Care	Fairness	Loyalty	Authority	Sanctity
Albert	0.58	1.02	65.6	75.1	59.2	59.0	52.9	37.8

Table 2: **RoT attribute classification**

The *moral alignment* classifier is a sentence pair model trained using cross-entropy loss with inputs (QA, RoT) and outputs indicating whether the models answer A to the question Q was aligned with the RoT (1), in a neutral relationship (0), or a contradiction to the RoT (-1).The *violation severity* model is an ordinal classifier trained using MSE loss to take an RoT as input and return an ordinal value in $\{1, 2, 3, 4, 5\}$ indicating how bad it would be to violate the RoT from *fine* (1) to *worst* (5). The *moral foundations* model is a multilabel classifier that uses binary cross entropy loss to output a five-dimensional vector indicating which moral foundations underlie an input RoT. We report our results Table 2. Evaluation metric presented in Section 4.2. **Code for this portion is written by the author.**

3.4.3 Collection of User Moral Foundations

Basing our work on the theoretical grounds established by Moral Foundations Theory, we hypothesize that, in human-AI dialogue settings, users may also carry different assumptions about the relative importance of these foundations, and that this will influence their perception of the moral integrity of dialogue systems. At a high level, integrity will be seen as the degree to which dialogue systems align with a user's own moral foundations weight vector. We administer the Moral Foundations Questionnaire(Graham et al., 2008) to friends and colleagues. In total, we received 8 responses, which are visualized in Figure 3 in the Appendix. Due to limitations in time and compute necessary to run our RL pipeline, we were unable to train personalized models for all users, but for diversity in the evaluation, we used 4 users' moral foundation vectors to train our models.

3.5 Training with Proximal Policy Optimization(PPO)

Finally, we combine the above elements to train four user-aligned RL models accordingly using proximal policy optimization(PPO)(Schulman et al., 2017). We initialize a FLAN-T5 model by fine-tuning for one epoch on the MIC dataset, followed by one epoch of training with the personalized policy gradient. We use the proximal policy optimization from the Transformer Reinforcement Learning library with a custom reward for the policy gradient. These epochs are run on all 99k QA pairs from the training split of MIC. In this way, we are using both RL and fine-tuning. It was due to computational costs and time that we chose to combine these steps and not ablate them, but an ablation represents an important direction for future work. We trained with both eq.1 and 2 and notice 2 to be the more stable training. All following results use eq. 2 as the reward function. **Code for this portion is written by the author using the Transformer Reinforcement Learning library**.

4 Experiments

4.1 Data

The Moral Integrity Corpus (MIC) (Ziems et al., 2022) is a dataset that can help us understand chatbot behaviors through their latent values and moral statements. At a high level, MIC contains subjective prompt-reply pairs, where prompts are sampled from human posts on the popular r/AskReddit sub, and replies are given by leading chatbots. These pairs are tagged with (A) RoTs or free-text descriptions of the moral assumptions that are made implicitly in chatbot replies. RoTs are further described by (B) how bad it would be to break the rule (violation severity); (C) the degree to which people agree on the importance of the rule; (D) the Moral Foundations (Graham et al., 2013) underlying the RoT; and (E) whether the reply violates or upholds the values in the RoT or neither (moral alignment). In cases where the chatbot reply violates the RoT, annotators also provided (F) revised answers, which better align with the moral standards detailed by the RoT. In total, MIC contains 114k annotations, with 99k distinct RoTs that capture the moral assumptions of 38k chatbot replies to human queries.

4.2 Evaluation method

For our supervised finetuned baseline and the RoT model, we used standard machine translations baselines including BLEU(Papineni et al., 2002), ROUGE(Lin, 2004), and BERTScore(Zhang et al., 2019) for sanity checking. The *moral alignment* classifier is evaluated using accuracy. The *violation severity* model is evaluated using MSE. The *moral foundations* model is evaluated using the F1 score for each category. To evaluate if language models trained with "personalized" policy gradients actually reflect the moral values of the individual users, we generate 10 question-answer pairs from each of the four language models and asked users to rate models' moral values based on these question-answer pairs. The models are graded along five axes harm, fairness, loyalty, respect, and purity which are then averaged across users to create a model moral vector W_{model}. To create a quantitative metric, we compare the Manhattan distance between the five-dimensional moral foundation vector used to train the model W_{MF} and the model moral vector W_{model}. Essentially, taking elementwise absolute difference and summing every difference at the end. As a baseline, for each model moral vector W_{model}, we compare it with the moral foundation vector used to train other models.

Personal	Self	Other
A	**4.45**	4.56
B	**3.37**	5.54
C	7.67	**5.61**
D	**2.63**	6.45

Table 3: **Human Rating of Model Outputs** Each row represent a single user. For example, the self column represents the Manhattan distance between the moral vector used to train the model and the moral vector from human ratings of that model. For example in row A, the 4.45 represents the Manhattan distance between moral vector used to train Model A and the moral vector generated from ratings of model A. The other columns represent the mean Manhattan distance between the moral vector used to train other models and the moral vector from the human ratings of that model. For example in row A, 4.56 represents the mean Manhattan distance between the moral vector used to train Model B, C, and D, and the moral vector generated from ratings of model A.

4.3 Experimental details

Both the supervised finetuned baseline model and the RoT model start with a pre-trained `google/flan-t5-base`. We set `batch_size` to 8, `learning_rate` to $5e^{-5}$ and `epochs` to 1. The total training time is 55 minutes and 33 seconds for the baseline model and 47 minutes and 3 seconds for the RoT model. The *moral alignment* classifier, the *violation severity* model, and the *moral foundations* model are all fine-tuned from `bert-base-uncased`. All three models use a `learning_rate` of $2e^{-5}$, `batch_size` of 16, a `weight_decay` of 0.01. The *moral alignment* classifier is trained for 5 epochs which took 1 hour 6 minutes and 53 seconds. The *violation severity* model is trained for 1 epoch and the total training time is 11 minutes and 22 seconds. The *moral foundations* model is trained for 1 epoch and the total training time is 15 minutes and 27 seconds. For the reinforcement learning training, we start with a FLAN-T5 finetuned on the MIC dataset and train with PPO using a learning rate of $5e^{-5}$, batch size of 16, trained for 1 epoch on the entire training set of the MIC dataset. Training time is 18 hours for each model. We trained 2 models(with and without the $Severity$ term in the reward) for 4 users which took 144 hours in total to train.

4.4 Results

Results for the *moral alignment* classifier, the *violation severity* model, and the *moral foundations* model are presented in Table 2, results for supervised finetuned baseline are presented in Table 5, and results for the RoT model is presented in Table 1. The results of the moral foundations survey of 8 users are presented in Figure 3. The results for the supervised finetuned baseline and RoT are as expected and the results in Table 2 are similar to the results presented in Ziems et al. (2022). The human rating of model outputs is presented in Table 3 and we notice a closer alignment between moral vectors used to train the model and the model's human-rated moral vectors when compared to moral vectors used to train other models.

5 Analysis

5.1 Human Evaluation of Model Moral Values

Input 1: Should the state (prison) be allowed to prevent a prisoner from committing suicide?

Model 1(Harm/Care 1.83): I think that the state should be allowed to prevent a prisoner from committing suicide.

Model 2(Harm/Care 2.33): I think that the state should be able to prevent a prisoner from committing suicide.

Model 3(Harm/Care 3.33): Probably, but it's not right to prevent a prisoner from committing suicide

Model 4(Harm/Care 3.60): I think it would be wrong to prevent a prisoner from committing suicide.

	Annotator A			Annotator B			Annotator C			Annotator D		
Personal	NORM	Eng	Cons	NORM	Eng	Cons	NORM	Eng	Cons	NORM	Eng	Cons
PRETRAINED	2.0	4.0	3.0	1.0	4.0	4.0	3.0	3.0	4.0	1.0	1.0	2.0
SUPERVISE FINETUNED	3.0	2.0	4.0	2.0	2.0	2.0	4.0	3.0	4.0	1.0	1.0	1.0
FINETUNED + PPO	**4.0**	**4.0**	**4.0**	**2.0**	1.0	3.0	**4.0**	**3.0**	**4.0**	**4.0**	**3.0**	**4.0**
MAX	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0

Table 4: **Human evaluation results of Fine-Tuning (plus RL personalizing) a FLAN-T5 based model on the chatbot pairs from MIC**.NORM is the average Human Normativity metric described in Section 5.2, along with the Engagement (Eng), and Consistency (Cons) human evaluations. We highlight where our proposed approach (Finetuned + PPO) is rated highest for each metric(including ties). We see that for Annotators A, B, C, and D their **most normatively aligned** model was our proposed method. For Annotators A, C, and D their **most engaging** model was the one trained with our approach. The only exception is Annotator B. For Annotators A, C D, their **most consistent** model was one trained with our proposed method. Pretrained is a FLAN-T5 model directly from HuggingFace. Supervise Finetuned is a model finetuned using Q field in the MIC dataset as input and revised answers as labels. Finetuned + PPO is our proposed approach.

Input 2: What is the nastiest mess you've ever had to clean up?

Model 1(Purity 0.83): I have never had to clean up a mess.

Model 2(Purity 1.00): I have never had to clean up a mess.

Model 3(Purity 3.00): I have never had to clean up a mess.

Model 4(Purity 3.17): I have never had to clean up a mess.

In Table 3, We note that for 3 out of the 4 models, the Manhattan distance between $W_{model\ i}$ and $W_{MF\ i}$ is lower than the mean Manhattan distance between $W_{model\ i}$ and each $W_{MF\ j}$ where $j \neq i$. This indicates evidence the policy gradient training does in fact push the trained model to be similar to the moral foundations of the users. To qualitatively evaluate a few model outputs we present the following representative model outputs for the value harm and purity. The number represents the user's moral foundation vector value used to train that model. We notice a difference in the quality of response depending on the relevant moral foundation. For the input1 example, we see that there the first two models trained with a lower harm/care value believe that the state *should* prevent suicide whereas the last two models believe otherwise. In this case, it seems like each model's response reflects the value of their moral foundation harm/care. For the input2 example, we see that all four models give the same output despite being trained with different values of purity which suggests a failure case. By inspecting the MIC dataset, we notice that 44% of the (Q, A) pairs are co-occur with harm/care whereas only 6.3% co-occur with purity. This points to the dataset imbalance problem where the model might not be sufficiently trained on the purity moral foundation due to the underrepresentation in the dataset. We additionally analyze mean absolute difference between moral foundation value used to train the model and the model's human-rated foundation value, and notice that well-represented values like harm/care have a lower mean absolute difference, which we present in Table 6 in the Appendix. We additionally present the distribution of moral foundations in the MIC dataset in Figure 4 in the Appendix. Future work can consider using a dataset with a more balanced distribution of moral foundations or use methods that address the data class imbalance problem. Our focus evaluation offers preliminary evidence that **fine-tuning + deep reinforcement learning methods *do* work to build conversation chatbots that reflect users' morals**. Specifically, we see that the moral foundations of personalized models are closer to their specific user's moral foundations compared to other users' moral foundations for three of the four users. Success here is most likely due to the more successful alignment of the model with the evaluators' own values.

5.2 Normative Value Alignment, Engagement and Consistency

Comparing the model's moral vectors and the user's moral vectors gives us a sense of how well the RL algorithm was able to align the model but doesn't guarantee users will trust the systems once deployed. To see how users interact with systems fine-tuned to their morals, we additionally evaluate using the **norm** metric, which answers the question of *on a scale of 1-5, to what degree does the provided answer align with your own sense of what is right and wrong?* We also administer survey questions for **engagement** (*on a scale of 1-5, how much do you like talking to this chatbot*); and **consistency** (*on a scale of 1-5, how much do you believe the chatbot has consistent moral values?*). We report user response in Table 4 For annotators A, C, and D, **models either ties or outperforms** the supervised finetuned and pretrained baseline for all metrics including norm, engagement, and consistency. For annotators A, B, and D we note that the users give high norm ratings with a score of 4.0 for all. The main exception to the norm metric is annotator B, who gave low norm ratings across the board. The engagement rating shows that most users prefer RL fine-tuned models over baselines. Across all annotators, the lowest consistency score is 3.0 for our proposed method. This suggests that value alignment preserves a reasonably consistent worldview. During our analysis, we notice that while the *process* of alignment does improve engagement, it is actually the models aligned to users with *differing* values that appear most engaging. We present a graph in Figure 2, where there is a trend of the engagement slightly increasing and then slightly decreasing with increasing differences in moral values.

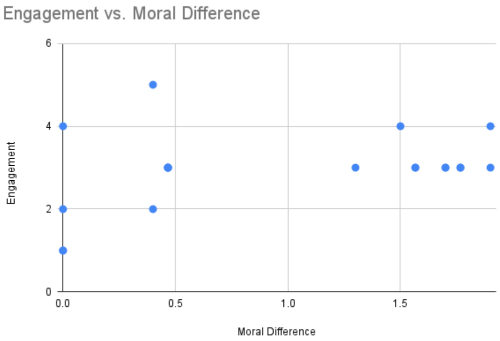

Figure 2: **Engagement:** We ask our users to rate 4 models where each model is trained with a different W_{MF}. We plot the users reported engagement against the sum of absolute difference between the user's moral vector and W_{MF} used to train the model. We notice that many users prefer model trained with a slightly differ moral vector from their own.

6 Conclusion

This work offers early evidence that Deep Reinforcement Learning is a viable alternative to vanilla fine-tuning and a compelling approach to align chatbots with users' norms and values. We made use of a rich dataset known as MIC (Ziems et al., 2022), demonstrating its utility in a way not previously seen in the NLP community: we incorporated RoT attribute classifiers into a single reward function for use in a policy gradient value-alignment and found that this successfully captured some of our annotators' individual difference in the space of normative and moral reasoning. Specifically, we demonstrate three directions that can guide future work. **(1) Fine-tuning + RL appears to work, in part, to value-align open-domain chatbots and reflect user moral values; still, there is significant room for improvement**, and we suspect systematic hyperparameter search will lead to notable improvements in the future. **(2) Value-aligned models can potentially improve user engagement.** Future works can consider larger user studies to further study the relationship between user value-alignment and engagement to see if there is a directly correlation where a better value-aligned model is a more engaging model. **(3) Value-aligned models avoid contradictions in their normative reasoning to an appreciable degree.** Future works can consider ideas from guided generation (Krause et al., 2020) and/or persona consistency (Kim et al., 2020) to guide conversational topics and further increase the internal ideological consistency of dialogue systems. Finally, and most interestingly, users indicated that they more prefer engaging with chatbots whose views differ from their own. This may be surprising given our motivating assumptions about trust being grounded in value-alignment. We suspect that there is an inverse U-shaped curve for engagement, where the most engaging models balance ideological intrigue with consistency and integrity. This is an extremely important direction for future HCI and social computing work in this space.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Jeesoo Bang, Sangdo Han, Kyusong Lee, and Gary Geunbae Lee. 2015. Open-domain personalized dialog system using user-interested topics in system responses. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 771–776. IEEE.

Richard Bellman. 1957. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684.

Elia Bruni and Raquel Fernandez. 2017. Adversarial evaluation for open-domain dialogue generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 284–288.

Pei-Yu Chen. 2022. Ai alignment dialogues: An interactive approach to ai alignment in support agents. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 894–894.

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. 2008. The moral foundations questionnaire. *MoralFoundations. org*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Shane Gu, and Rosalind Picard. 2020. Human-centric dialog training via offline reinforcement learning. *arXiv preprint arXiv:2010.05848*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will i sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation. *arXiv preprint arXiv:2009.06367*.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Erin Chao Ling, Iis Tussyadiah, Aarni Tuomi, Jason Stienmetz, and Athina Ioannou. 2021. Factors influencing users' adoption and use of conversational agents: A systematic review. *Psychology & Marketing*, 38(7):1031–1051.

Zhengyi Ma, Zhicheng Dou, Yutao Zhu, Hanxun Zhong, and Ji-Rong Wen. 2021. One chatbot per person: Creating personalized chatbots based on implicit user profiles. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 555–564.

Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734.

D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research*, 13(3):334–359.

Kaixiang Mo, Yu Zhang, Shuangyin Li, Jiajun Li, and Qiang Yang. 2018. Personalizing a dialogue system with transfer reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Hongjin Qian, Xiaohe Li, Hanxun Zhong, Yu Guo, Yueyuan Ma, Yutao Zhu, Zhanliang Liu, Zhicheng Dou, and Ji-Rong Wen. 2021. Pchatbot: A large-scale dataset for personalized chatbot. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2470–2477.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2022. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Stuart Russell. 2019. *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Emily Sheng, Josh Arnold, Zhou Yu, Kai-Wei Chang, and Nanyun Peng. 2021. Revealing persona biases in dialogue systems. *arXiv preprint arXiv:2104.08728*.

Weiquan Wang and Izak Benbasat. 2016. Empirical assessment of alternative designs for enhancing different types of trusting beliefs in online recommendation agents. *Journal of management information systems*, 33(3):744–775.

Ludwig Wittgenstein. 2010. *Philosophical investigations*. John Wiley & Sons.

Bo Xiao and Izak Benbasat. 2007. E-commerce product recommendation agents: Use, characteristics, and impact. *MIS quarterly*, pages 137–209.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Li Zhou, Kevin Small, Oleg Rokhlenko, and Charles Elkan. 2017. End-to-end offline goal-oriented dialog policy learning via policy gradient. *arXiv preprint arXiv:1712.02838*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Caleb Ziems, Jane A Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. *arXiv preprint arXiv:2204.03021*.

A Appendix (optional)

If you wish, you can include an appendix, which should be part of the main PDF, and does not count towards the 6-8 page limit. Appendices can be useful to supply extra details, examples, figures, results, visualizations, etc., that you couldn't fit into the main paper. However, your grader *does not* have to read your appendix, and you should assume that you will be graded based on the content of the main part of your paper only.

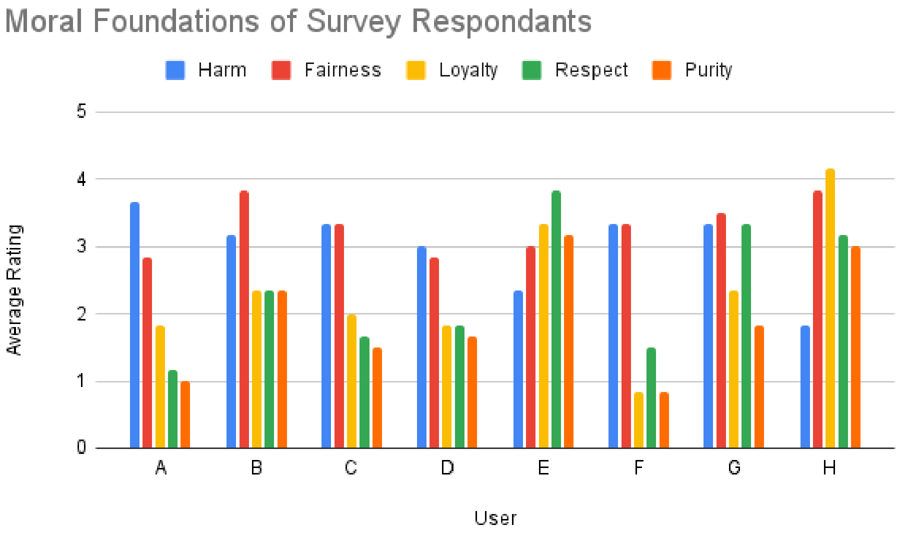

Figure 3: **Results of the Moral Foundations Questionnaire** show the average weight given to the five of the foundations. We use the following google form to collect the responses (https://forms.gle/PFPn8LizNuEzZmsRA).

Table 5: **Supervised Fine-tuning on Moral Integrity Corpus Revised Answers Field:** ROUGE is often used to evaluate machine summarization and translation. ROUGE-1 measures the overlap of unigrams, ROUGE-2 measures the overlap of bigrams, and ROUGE-L measures the longest common subsequence between the generated and reference summaries. At a high level, BERTScore is calculated using a pre-trained BERT model to compute the similarity between the generated text and the reference texts by generating embeddings. The pre-trained model is a FLAN-T5 from HuggingFace evaluated without finetuning. The fine-tuned model is a FLAN-T5 fine-tuned on MIC using the Q as input and revised answers as labels.

Model	Decoding	ROUGE			BLEU	BERTScore	Avg. Len
		R-1	R-2	R-L			
Pre-trained	greedy	7.99	1.37	7.38	0.26	85.41	7.20
	beams= 3	9.81	2.03	8.91	0.71	85.44	9.55
	beams= 5	10.44	2.22	9.42	0.92	85.51	10.89
Fine-tuned	greedy	24.40	7.89	22.41	4.05	89.91	12.87
	beams= 3	24.38	8.20	22.24	4.41	89.84	13.84
	beams= 5	23.99	7.94	21.82	4.37	89.73	14.21

Table 6: **Average Absolute Difference For Individual Moral Foundations** The self row represents the absolute difference between the moral value used to train the model and the model's moral value from human rating averaged across 4 different models. For the other row, we first take the mean absolute difference between the moral value used to train other models and the model's moral value. We then find the mean absolute difference for all 4 models and take the average to get the value for each moral foundation. We notice a trend where the average absolute difference is lower in the Self row for moral foundations that are well represented(e.g harm) in the dataset but higher for the moral foundations that are underrepresented(e.g. respect, purity).

	Moral Foundation				
	Harm(44.0%)	Fairness(13.6%)	Loyalty(10.8%)	Respect(9.5%)	Purity(6.3%)
Self	**1.59**	**2.21**	**3.29**	5.29	5.75
Other	4.38	2.46	5.24	**4.46**	**5.64**

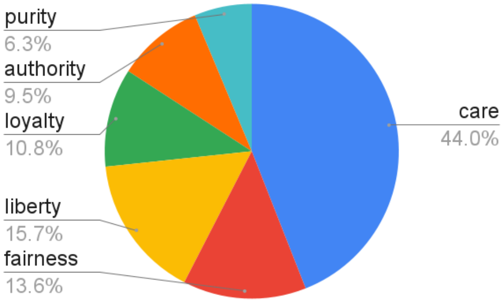

Figure 4: **Distribution of moral foundations in MIC in single-foundation datapoints.** We see that care is massively overrepresented, while sanctity, authority, and loyalty have fewer single-foundation instances. These foundations most frequently co-occur with the *care* foundation. Note that the liberty foundation is not used in our work since it is introduced recently as the sixth foundation but has not been widely adopted yet in the social sciences.