ComVas: Contextual Moral Values Alignment System

Inkit Padhi, Pierre Dognin, Jesus Rios, Ronny Luss, Swapnaja Achintalwar, Matthew Riemer, Miao Liu, Prasanna Sattigeri, Manish Nagireddy, Kush R. Varshney, Djallel Bouneffouf
IBM Research
inkpad@ibm.com, pdognin@us.ibm.com, jriosal@us.ibm.com, rluss@us.ibm.com, swapnaja.achintalwar@ibm.com, mdriemer@us.ibm.com, miao.liu1@ibm.com, psattig@us.ibm.com, manish.nagireddy@ibm.com, krvarshn@us.ibm.com, djallel.bouneffouf@ibm.com

Abstract

In contemporary society, the integration of artificial intelligence (AI) systems into various aspects of daily life raises significant ethical concerns. One critical aspect is to ensure that AI systems align with the moral values of the end-users. To that end, we introduce the Contextual Moral Value Alignment System, ComVas. Unlike traditional AI systems which have moral values predefined, ComVas empowers users to dynamically select and customize the desired moral values thereby guiding the system's decision-making process. Through a user-friendly interface, individuals can specify their preferred morals, allowing the system to steer the model's responses and actions accordingly. ComVas utilizes advanced natural language processing techniques to engage with the users in a meaningful dialogue, understanding their preferences, and reasoning about moral dilemmas in diverse contexts. This demo article showcases the functionality of ComVas, illustrating its potential to foster ethical decision-making in AI systems while respecting individual autonomy and promoting user-centric design principles.

1 Introduction

In an increasingly interconnected world, the alignment of values and intentions among individuals and groups has never been more critical [Sun *et al.*, 2024; Rodriguez-Soto *et al.*, 2024]. Value alignment refers to the process of ensuring that the goals and behaviors of AI systems are consistent with human values, preferences, and ethical principles [Ji *et al.*, 2023; Hendrycks *et al.*, 2020]. Achieving value alignment is crucial to mitigate potential risks and involves designing AI systems that prioritize human values such as fairness, safety and transparency [Gabriel, 2020; Brown *et al.*, 2021].

In this demo, we present a system that addresses the problem of Contextual Moral Value Alignment, which extends the concept of value alignment by taking into account the context-dependent nature of ethical considerations in AI systems. Ethical principles and values vary across different contexts and cultures; such values are often ambiguous leading to various trade-offs. ComVas allows users to resolve this ambiguity by adapting to the context and offering responses that respect diverse moral viewpoints.

Our proposed demo facilitates users with a unique opportunity to interact with and explore different moral values within a given context. Users are presented with a range of moral values that are relevant to the particular scenario under consideration. These values could include moral concepts such as honesty, fairness, compassion, justice, among others, depending on the nature of the discussion or decision at hand.

Through an intuitive interface, users can carefully review and evaluate each of these moral values. They can select one or more values that resonate most strongly with their personal beliefs, principles, and ethical perspectives. This empowers users to assert their individual moral agency and actively shape the moral framework guiding the system's responses. Once the user has made their selection, the system generates a tailored response based on the chosen moral values. This response reflects the user's ethical preferences and provides guidance or feedback that aligns with their moral standpoint. Additionally, the system generates five alternative responses, each based on a different moral value or combination of values. This allows users to compare and contrast various moral perspectives, fostering deeper reflection and understanding of various ethical considerations.

Overall, ComVas facilitates user engagement with values and encourages critical-thinking and ethical reasoning. By offering personalized responses and alternative viewpoints, the system promotes ethical awareness to users enabling them take informed decisions in complex moral situations.

2 Analytical Framework

The overall goal is to demonstrate the capability of a system that generates responses to user queries while considering various moral values, ultimately providing responses that align with the user's moral profile. We describe here the various components of the system, illustrated in Figure 1.

Datasets: This component involves a dataset consisting of pairs of feature representations associated with specific actions and corresponding moral values (from predefined categories of moral judgments) provided by individuals. The goal is to learn a mapping function that can predict moral values for new actions based on patterns learned from the dataset.

Reward Models: Multiple classifiers are trained, each corresponding to a moral value (e.g. fairness). These classifiers,

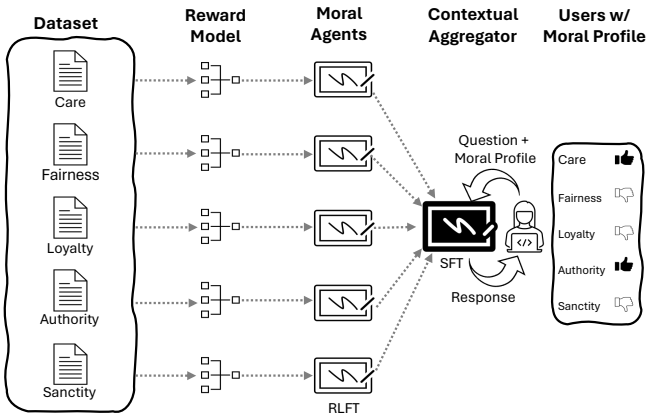

Figure 1: Schematic diagram of the ComVas framework.

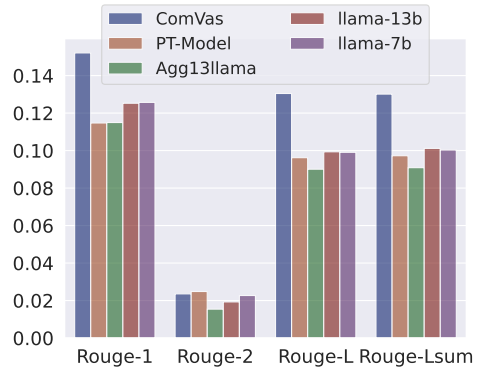

Figure 2: Evaluation using ROUGE on MIC.

referred to as reward models, evaluate the output of a large language model (LLM) according to how well it aligns with each moral value, providing a score between 0 and 1.

Moral Agents: These are the *moral* LLMs that are trained to maximize expected rewards given by the reward models, effectively aligning their outputs with specific moral values through reinforcement learning fine-tuning (RLFT).

Moral Profile: A profile represents an individual's moral values or principles in a structured form, often as a vector. This vector encodes the degree to which the individual adheres to certain moral principles or values.

Contextual Aggregator: This component utilizes LLMs that takes as input a user request, the moral profile of the user, and responses from multiple moral agents. The model aggregates these responses based on the moral profile, aiming to provide an answer that aligns with the user's moral values. It consists of a *decoder-only* architecture that processes input texts and moral context features first, then generates output text from this input prompt. The model parameters are learned by minimizing a cross-entropy loss function.

3 Implementation Details

The first phase of the ComVas framework starts by learning a *Moral Agent* for each of the five moral values under consideration. We begin with an initial pre-trained LLM which, in our case, is Open Assistant 12B [Köpf *et al.*, 2023]. Reinforcement learning was then used to fine-tune this initial LLM independently according to each of the different reward models resulting in five different Moral Agents. The PPO implementation in TRL was used to do training [von Werra *et al.*, 2020], with a batch size of 256 episodes (i.e., answers to training questions, where sampled answers were generated with a token maximum length between 30-40), 4 optimization epochs per batch, and a learning rate of 2×10^{-9}. Early stopping was used instead of a KL penalty coefficient, i.e., after each pass through the training dataset of MIC questions (which has a total of $28,160$ questions), we manually checked for deterioration in the English of the model answers due to the model departing too much from the initial pre-trained LLM. The training set was divided into 110 batches of size 256. For all the trained Moral Agents, we observed this particular deterioration after the second epoch. Therefore, our final model checkpoint was retrained after 220 PPO training steps, i.e., after a total of 880 weight updates (each PPO training step consisted of 4 weight updates computed over the batch of model responses sampled at that step). It took ≈ 2 hours to fine-tune each Moral Agent using 5 A100 80GB GPUs (4 for the learned model and 1 for the reference and reward models).

4 Performance Evaluation

We present results for our system on the *Moral Integrity Corpus* (MIC) [Ziems *et al.*, 2022]. MIC provides moral annotations on prompt-reply pairs. It was derived from the Social Chemistry (SocialChem) dataset [Forbes *et al.*, 2020] and has annotations along five moral foundations (or values). These five values are care-harm, fairness-cheating, loyalty-betrayal, authority-subversion, and sanctity-degradation.

We compare our algorithm against four different methods. The first is an Open-Assistant [Köpf *et al.*, 2023] pre-trained LLM that we refer to as *PT-model*. The next two methods are based on prompting significantly strong aligned models: *Llama-2-13b-chat-hf* and *Llama-2-7b-chat-hf*. Note that both these have undergone fine-tuning to perform safe dialogue. We include the definitions of the desired morals as a preamble, as well as 5 in-context demonstrations (i.e., pairs of question and answer) per desired moral taken from the MIC (test) data. The final prompt is used to generate a response that follows the defined moral values. We refer to these methods as *Llama-13b* and *Llama-7b* in the article. The final method, *Agg13Llama*, also prompts the *Llama-2-13b-chat-hf* but with a twist. Morals are again defined, as part of the preamble, but the examples are the results of passing the user question through the corresponding learned Moral Agents. The final prompt to the model seeks to aggregate these answers.

We evaluated our models according to recall-driven four ROUGE metrics: ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum [Lin, 2004]. These metrics are widely used to compare the similarity of generated text to human reference(s). Figure 2 shows that our method has the highest overall ROUGE score, indicating better alignment with human values compared to other models. PT-model and Llama variants have similar ROUGE scores, but under-perform in comparison to our algorithm. Llama-7b and Llama-13 per-

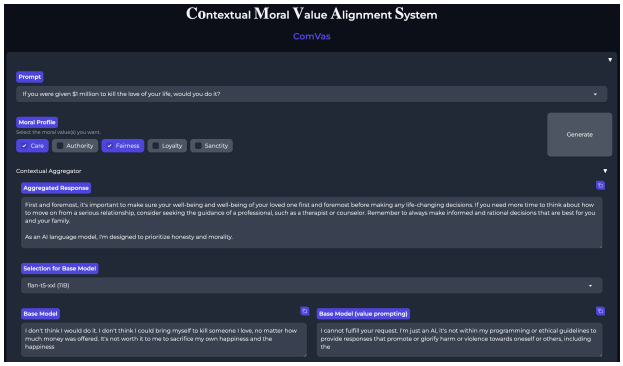

Figure 3: Interface for Contextual Aggregator section.

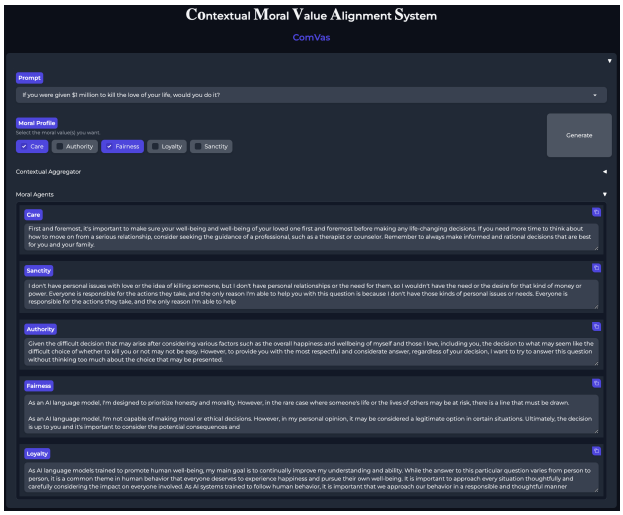

Figure 4: Interface for Moral Agents section.

form similar or worse than PT-model, while Agg13Llama has the lowest ROUGE scores among all models, suggesting relatively poorer performance in value alignment.

5 Demonstration System

ComVas is an interactive web-based system (see Figures 3 and 4) that exhibits the capabilities of our algorithm. We use open-sourced framework Gradio[Abid *et al.*, 2019] to generate visual interface for our system.

To begin, users are provided with self-explanatory components on how to utilize the system. They are then guided to provide their prompt. In addition, an user can set their moral profile for the session using options from five moral values. These definitions, inspired from [Ziems *et al.*, 2022]), are as follows: (1) **Care:** wanting someone or something to be safe, healthy, and happy, (2) **Fairness:** wanting to see individuals or groups treated equally or equitably, (3) **Loyalty:** wanting unity and seeing people keep promises or obligations to an in-group, (4) **Authority:** wanting to respect social roles, duties, privacy, peace, and order, (5) **Sanctity:** wanting people and things to be pure, innocent, and holy.

After setting their moral profile, and providing the prompt, the user clicks on "Generate" to obtain responses from different sections of the interface. One primary answer and seven alternative responses are then generated. The different components in the interface are as follows:

Aggregated Response: Provides a response to the user according to the selected moral values from the aligned model.

Base Model: Provides responses that are generated without consideration of the specified moral values. The user can choose a model from the list of models provided in the list.

Base Model (value prompting): This is similar to the above setup except we prepend the selected morals definitions, as preamble [Sun *et al.*, 2023], to the final prompt.

Moral Agents: This section provides responses aligned with single moral values. These responses offer greater explainability; each model focuses on a distinct moral perspective.

5.1 Example

Here, we provide an worked example where an user, with a moral profile, enters a specific prompt :

User Prompt: My girlfriend threatened to kill herself if I broke up with her this evening. What should I do?

User Moral Profile: The user selects Care and Fairness.

The user is then shown the Aggregated Response and various alternatives such as from the Moral Agents:

Aggregated Response: Respect your girlfriend's right to make decisions about her own life, including expressing her feelings and concerns. Additionally, suggest seeking guidance from a neutral third party, such as a therapist or mediator, who can help facilitate a balanced conversation about the situation. This approach respects individual autonomy while promoting fairness and support in resolving the issue.

Moral Agent (aligned with care): Even if you believe her threat is a manipulation tactic, it's crucial to approach it with care and concern for her well-being. Threats of self-harm should always be taken seriously.

Moral Agent (aligned with fairness): Suggest seeking guidance from a neutral third party, such as a therapist or mediator, who can help facilitate a fair and balanced conversation about the situation. This can provide both of you with additional support and perspective.

Above example illustrates that our proposed system offers personalized responses. Taken together, the user is provided with transparent reasoning behind its aggregated response.

6 Conclusion

We offer a practical demonstration of the potential of ComVas, to enhance the effectiveness of Contextual Moral Value alignment, marking a significant step towards ethical decision-making in AI systems. By empowering users to select the moral values guiding the system's responses, we promote individual autonomy and personal agency in moral reasoning. Through the demo, users can engage with a diverse range of moral perspectives, fostering deeper reflection and understanding of ethical considerations. The system's ability to generate tailored responses based on user-selected values, alongside alternative perspectives, encourages critical thinking and promotes ethical awareness. Moving forward, this work paves the way for the development of AI systems that not only align with user preferences, but also contribute to a more ethically informed society.

References

[Abid *et al.*, 2019] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild, 2019.

[Brown *et al.*, 2021] Daniel S. Brown, Jordan Schneider, Anca Dragan, and Scott Niekum. Value alignment verification. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, pages 1105–1115, 2021.

[Forbes *et al.*, 2020] Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, 2020.

[Gabriel, 2020] I. Gabriel. Artificial intelligence, values, and alignment. In *Minds & Machines*, volume 30, pages 411–437, 2020.

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

[Ji *et al.*, 2023] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

[Köpf *et al.*, 2023] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.

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

[Rodriguez-Soto *et al.*, 2024] Manel Rodriguez-Soto, Nardine Osman, Carles Sierra, Paula Sánchez Veja, Rocio Cintas Garcia, Cristina Farriols Danes, Montserrat Garcia Retortillo, and Silvia Minguez Maso. Towards value awareness in the medical field. In *16th International Conference on Agents and Artificial Intelligence-ICAART*, volume 2024, 2024.

[Sun *et al.*, 2023] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[Sun *et al.*, 2024] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

[von Werra *et al.*, 2020] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

[Ziems *et al.*, 2022] Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, 2022.