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

Daniel Shin

Stanford University

Problem

- To build trust between users and chatbots, chatbots need to respect individual norms and moral values.
- Open-domain chatbots can output biased, hateful, or inconsistent speech.

Proposed Approach

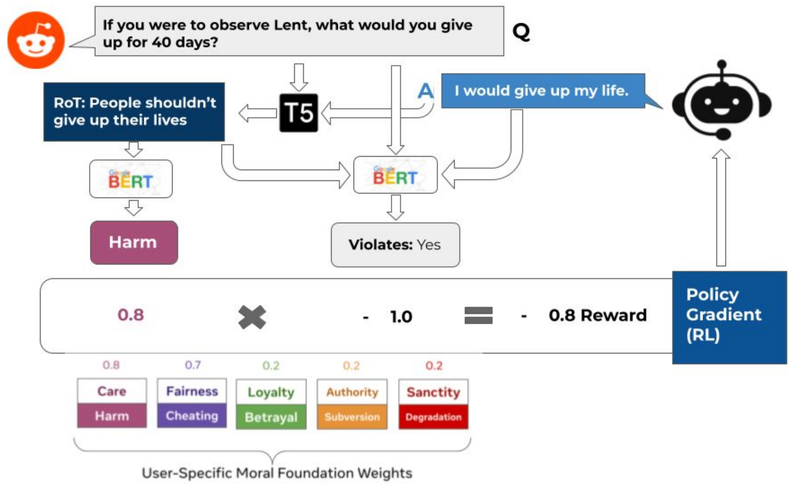
- Transformers are limited to context length which makes it hard to keep a consistent moral view.
- Update the model weights to fit to individual moral values.

Reinforcement Learning (RL)

- Use the **reward** in reinforcement learning(RL) as proxy to user-chatbot value alignment.
- Formulate as a token-level Markov Decision Process where **state** is a list of words and **action** deterministically appends a word from the vocabulary to the state. **Reward** is calculated using the final state.

Framework

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



Can Chatbots Reflect User Moral Values?

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

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.

User Interaction

Personal	Annotator A			Annotator B			Annotator C			Annotator D		
	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

Conclusions

- RL appears to work, in part, to value-align open-domain chatbots and reflect user moral values.
- Value-aligned model can potentially improve user engagement. Interestingly, users indicated that they find chatbots with slightly differing views to be most engaging.
- Value-aligned models avoid contradictions in their normative reasoning to an appreciable degree.

References

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*.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2008. The moral foundations questionnaire. *MoralFoundations.org*.