NLP Alignment: A Roadmap to Aligning NLP with AGI Safety

Zhijing Jin (Max Planck Institute & ETH Zürich)
November 2021

"*How can NLP research contribute to the goal of AGI safety?*" This is the question standing at the core of my PhD research. Behind this mission are my commitment to Effective Altruism since my undergraduate in 2019, and my strong passion for building safe natural language processing (NLP) models that are aligned with human values.

With a forecast of the superintelligent AI approaching in the foreseeable future [5, 20], AGI safety research is increasingly urgent and important [21]. However, most existing AGI safety frameworks are based on the reinforcement learning paradigm [1, 11]. In this proposal, I will highlight NLP, an important but neglected domain in the existing landscape of AGI safety. The importance of NLP to AGI safety rises with the recent emergence of very large language models and their impressive capabilities, such as GPT-3 [6], BERT [8] and its variants [7, 18, inter alia], and also increased adoption of NLP in various systems that are used by billions of users such as Alexa, Google Search, Google Translate, and Facebook and Twitter recommendations.

Based on the vast progress in NLP, I believe an urgent need for AGI safety emerges, and, accordingly, I describe my research direction on "*NLP alignment*" as a subcategory of general AI alignment. I outline my research on NLP alignment in terms of two subgoals: outer alignment and inner alignment. The roadmap of my research is summarized in Figure 1.

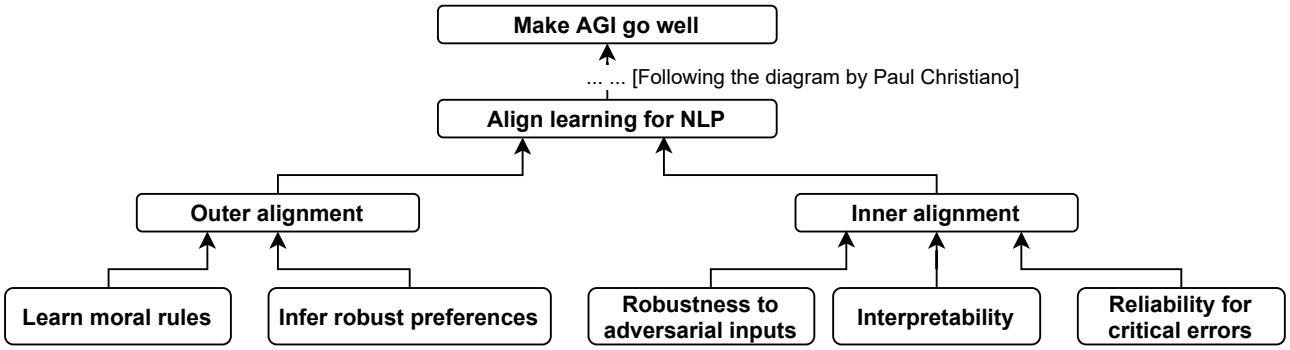

Figure 1: Roadmap of NLP alignment, including outer alignment (with subdirections of learning moral rules and inferring robust preferences), and inner alignment (with subdirections of robustness to adversarial inputs, interpretability, and reliability for critical errors).

1 Outer Alignment

I aim to prevent NLP models that only know how to perform a task well but ignore the true human preferences behind the literal meaning of the task. For example, if an NLP model is instructed to make the user happy, it should not do things like harming another person to make the user laugh. There are two subdirections that I am pursuing to address the outer alignment problem: (1) learning moral rules, and (2) inferring robust preferences of humans.

Learning Moral Rules: Human society uses moral rules and social norms to prevent people from "over-optimizing" their goals. Similarly, to make sure that the NLP models do not do harm in order to achieve a given goal, I need to make the models learn humans' moral values and social norms [10]. I have an ongoing project with Sydney Levine, a postdoc at MIT and Harvard whose expertise is in moral judgment and decision-making processes [15, 16, 17]. Combining with my expertise in NLP, we are making an interdisciplinary effort to make NLP models learn moral rules in a smart and critical way: knowing when to follow and break certain rules, under various subtle contextual changes [2]. For example, telling a lie is

generally considered morally wrong, but telling a white lie can be considered acceptable. We aim to consider the contextual situations, and make the NLP model learn moral decisions with awareness and reasoning of the context.

Inferring Robust Preferences of Humans: Another important goal in outer alignment is to learn the real preferences of humans. Instead of following dogmatic rules, humans have a complicated causality chain behind many decisions, and learning such causality can effectively help approach AGI alignment. With my lab's strength in causal inference, I am interested in making models learn the "causal process" of how humans complete a task. A feasible near-term goal is to learn the causal graph behind humans' language generation process. For example, when a person says a sentence, the words can be a product of the topic, opinion, and intent of the speaker; all these elements cause the utterance. I aim to model the language generation process of humans, specifically factoring in important aspects such as the intent of humans behind the literal words. Such intent can potentially be learned via latent variable causal graphs [23] and disentanglement, similar to the work I previously surveyed in text style transfer [12].

2 Inner Alignment

Together with outer alignment, I also want to make sure the inner working of NLP models fits our expectations. In the following, I will elaborate three threads of my research efforts: (1) testing NLP model robustness via adversarial attack and contrast sets (the *robustness-to-adversarial-inputs* goal), (2) causally explaining NLP model decisions (the *interpretability* goal), and (3) increasing NLP model reliability against critical errors such as social biases and logical fallacies (the *reliability-for-critical-errors* goal).

Testing NLP Model Robustness: I want NLP models to be robust against two types of adversarial attacks: perturbation (where the model should keep its prediction when the input is slightly paraphrased), and contrast sets (where the model should change its prediction when the input is changed to imply another output). My efforts in this direction can be seen in my previous papers which are broadly cited in the field of NLP model robustness, including an AAAI 2020 paper working on perturbation of BERT and other common NLP models [13], and an EMNLP 2020 paper working on contrast sets for aspect-based text classification [24].

Causally Explaining NLP Model Decisions: I want to interpret how NLP models make their decisions. For example, if a text is classified as hate speech, I want to know which features the model uses to identify this. I plan to build on existing work of causal interpretations of NLP models [3, 9, 22], as well as causal inference methods to identify latent variables in the causal graph [23]. Together with the two tools, I have ongoing work to learn the latent variable causal graph that the model uses to make its prediction.

Increasing NLP Model Reliability against Critical Errors: There are two types of critical errors that I want to eliminate from NLP models. The first one is social biases such as gender and racial biases. Based on an NLP dataset with annotations of these demographics [4], I have a near-term project to use causal discovery methods to check whether the model leverages these unwanted biases for its prediction. This method will be extensible to a broad range of tests to expose the unwanted spurious correlations that the model picks up [19].

The second critical error that I want to eliminate from models is logical fallacies. When testing with GPT-3, I found it easy to generate text to persuade people by logically fallacious arguments such as false analogy, and using correlation to imply causation. It is a very dangerous tendency to generate logically fallacious text for persuasion or propaganda purposes. To tackle this, I have done a preliminary study on logical fallacy detection for NLP models [14], where I contribute the first dataset of logical fallacies in NLP, and identify the urgent needs to push the next generation of NLP models to be logically reliable.

Conclusion

This effort towards NLP alignment is a challenging, long-term pursuit that requires tremendous effort. I am very dedicated to integrating this pursuit into my life-long career path in academic research, and continue to explore this topic in both breadth and depth.

Acknowledgements: For this proposal, I sincerely thank the many constructive suggestions from Geoffrey Irving (DeepMind), Mrinmaya Sachan (ETH), Rada Mihalcea (University of Michigan), and Cynthia Xin Chen (CHAI).

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016.

[2] Edmond Awad, Sydney Levine, Andrea Loreggia, Nicholas Mattei, Iyad Rahwan, Francesca Rossi, Kartik Talamadupula, Joshua Tenenbaum, and Max Kleiman-Weiner. When is it morally acceptable to break the rules? A preference-based approach. In *12th Multidisciplinary Workshop on Advances in Preference Handling (MPREF 2020)*, 2020.

[3] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *CoRR*, abs/2102.12452, 2021.

[4] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561, 2019.

[5] Nick Bostrom. *Superintelligence*. Dunod, 2017.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[7] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[9] Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. Causalm: Causal model explanation through counterfactual language models. *Comput. Linguistics*, 47(2):333–386, 2021.

[10] 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, Online, November 2020. Association for Computational Linguistics.

[11] Evan Hubinger. An overview of 11 proposals for building safe advanced AI. *CoRR*, abs/2012.07532, 2020.

[12] Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. Deep learning for text style transfer: A survey. *CoRR*, abs/2011.00416, 2020.

[13] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8018–8025. AAAI Press, 2020.

[14] Zhijing Jin, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. Detecting logical fallacies: From quiz to climate change news. 2021.

[15] Sydney Levine, Max Kleiman-Weiner, Nicholas Chater, Fiery Cushman, and Josh Tenenbaum. The cognitive mechanisms of contractualist moral decision-making. In *CogSci*, 2018.

[16] Sydney Levine, Max Kleiman-Weiner, Laura Schulz, Josh Tenenbaum, and Fiery Cushman. What if everybody did that?: Universalization as a mechanism of moral decision-making. In *CogSci*, page 2125, 2019.

[17] Sydney Levine, Max Kleiman-Weiner, Laura Schulz, Joshua Tenenbaum, and Fiery Cushman. The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences*, 117(42):26158–26169, 2020.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[19] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3428–3448. Association for Computational Linguistics, 2019.

[20] Toby Ord. *The precipice: Existential risk and the future of humanity*. Hachette Books, 2020.

[21] Stuart Russell. *Human compatible: Artificial Intelligence and the problem of control*. Penguin, 2019.

[22] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias. *CoRR*, abs/2004.12265, 2020.

[23] Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[24] Xiaoyu Xing, Zhijing Jin, Di Jin, Bingning Wang, Qi Zhang, and Xuanjing Huang. Tasty burgers, soggy fries: Probing aspect robustness in aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3594–3605. Association for Computational Linguistics, 2020.