Morality is Non-Binary: Building a Pluralist Moral Sentence Embedding Space using Contrastive Learning

Jeongwoo Park*, Enrico Liscio*, and Pradeep K. Murukannaiah

TU Delft, Delft, the Netherlands

Introduction

The study of human morality is garnering increasing attention in the field of NLP [1, 8, 9, 18, 21]. Existing works typically treat morality as a score ranging from right to wrong [2, 12, 17]. Such a binary approach to morality has been shown to be *emergent* in state-of-the-art models [19], which exhibit a moral direction that maps actions from "do's" to "don'ts" without being explicitly trained on data with moral annotations. However, this simplistic approach to morality does not capture the nuances of moral judgment [20]. Pluralist moral philosophers argue that morality can be deconstructed into a finite number of elements, referred to as *moral values* [5]. For instance, the debate on immigration touches on the moral values of fairness ("Everyone should be given equal opportunities") and in-group loyalty ("I worry about the preservation of our identity")—how each of us prioritizes fairness vs. loyalty influences our moral judgment in this debate.

Contribution We summarize our work published at EACL'24 [16], where we investigate whether a pluralist approach to morality is emergent in language models. To this end, we map a pluralist morality approach to a sentence embedding space, a multi-dimensional representation that allows us to geometrically and visually inspect the relationships between moral values inside a language model. Our experiments show that pluralist morality is not emergent—human-provided labels are necessary to train language models to discern pluralist morality.

Method and Experiments

We experiment with the Moral Foundation Twitter Corpus (MFTC [6]), composed of over 35k tweets from seven domains, ranging from MeToo to Hurricane Sandy. Each tweet is annotated with the values of the Moral Foundations Theory (MFT [5]), which postulates that morality can be deconstructed into 10 moral values (see Figure 1). We assess how the values are mapped in a sentence embedding space, a representation underlying language models that maps sentences as points in high-dimensional space, grouping semantically similar ones together.

* Equal Contribution.

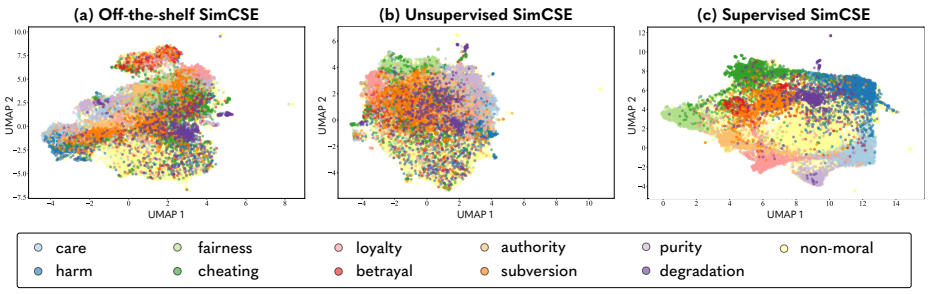

Fig. 1. UMAP plot of with off-the-shelf SimCSE model (a, left), unsupervised SimCSE approach (b, middle), and supervised SimCSE approach (c, right).

We use SimCSE [4] to train the embedding space by minimizing the distance between tweets with the same label and maximizing it between those with different labels. In its unsupervised variant, SimCSE treats each tweet as having a different label. In its supervised variant, SimCSE uses the MFTC labels to train the embedding space. We compare (a) an off-the-shelf SimCSE embedding space, and the embeddings trained with (b) the unsupervised and (c) the supervised SimCSE approaches. These variants help us determine whether a pluralist approach to morality emerges in (a) an off-the-shelf model that was not exposed to morally loaded data, and a model that has been exposed to morally loaded data (b) but not to the respective human labels or (c) with the respective labels.

Results and Discussion

We inspect the resulting embedding spaces through an intrinsic and extrinsic evaluation. First, we examine how well the embedding space distinguishes between different moral values within the MFT. We compute the similarity between tweets with similar and different moral labels and visualize the embedding space by mapping it into two dimensions through UMAP [13] (Figure 1). Next, we validate the embedding spaces against unseen data, i.e., the MFTC test set and the Moral Foundations Dictionary 2.0 (MFD2.0 [3]), an external dictionary of moral terms. Both intrinsic and extrinsic evaluations show that the supervised approach forms distinct clusters of moral values and can generalize to unseen data, while the off-the-shelf and unsupervised methods fail to do so.

Our experiments show that a pluralist approach to morality can be captured in a sentence embedding space, but also that human labels are necessary to successfully train the embeddings. Our work represents the starting point for incorporating a pluralist approach to morality in language models to allow them to map and reflect the diversity of human judgment [7, 10, 11, 14, 15]. However, our results constitute a warning that self-supervision alone is not sufficient to capture the complexity of human morality.

References

1. Alshomary, M., Baff, R.E., Gurcke, T., Wachsmuth, H.: The Moral Debater: A Study on the Computational Generation of Morally Framed Arguments. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. pp. 8782–8797. ACL '22, ACL, Dublin, Ireland (2022)
2. Forbes, M., Hwang, J.D., Shwartz, V., Sap, M., Choi, Y.: Social chemistry 101: Learning to reason about social and moral norms. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. pp. 653–670. EMNLP '20, ACL, online (2020)
3. Frimer, J.A.: Moral foundations dictionary 2.0 (Dec 2019), https://osf.io/ezn37/
4. Gao, T., Yao, X., Chen, D.: SimCSE: Simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6894–6910. EMNLP '21, ACL, Online and Punta Cana, Dominican Republic (2021)
5. Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S.P., Ditto, P.H.: Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. In: Advances in Experimental Social Psychology, vol. 47, pp. 55–130. Elsevier, Amsterdam, the Netherlands (2013)
6. Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A.M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., Moreno, G., Park, C., Chang, T.E., Chin, J., Leong, C., Leung, J.Y., Mirinjian, A., Dehghani, M.: Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. Social Psychological and Personality Science **11**, 1057–1071 (2020)
7. Lera-Leri, R.X., Liscio, E., Bistaffa, F., Jonker, C.M., Lopez-Sanchez, M., Murukannaiah, P.K., Rodriguez-Aguilar, J.A., Salas-Molina, F.: Aggregating value systems for decision support. Knowledge-Based Systems **287**, 111453 (2024)
8. Liscio, E., Araque, O., Gatti, L., Constantinescu, I., Jonker, C., Kalimeri, K., Murukannaiah, P.K.: What does a Text Classifier Learn about Morality? An Explainable Method for Cross-Domain Comparison of Moral Rhetoric. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. pp. 14113–14132. ACL '23, ACL, Toronto, Canada (2023)
9. Liscio, E., Dondera, A.E., Geadau, A., Jonker, C.M., Murukannaiah, P.K.: Cross-Domain Classification of Moral Values. In: Findings of the Association for Computational Linguistics: NAACL 2022. pp. 2727–2745. ACL, Seattle, USA (2022)
10. Liscio, E., Lera-Leri, R., Bistaffa, F., Dobbe, R.I., Jonker, C.M., Lopez-Sanchez, M., Rodriguez-Aguilar, J.A., Murukannaiah, P.K.: Value Inference in Sociotechnical Systems. In: Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems. pp. 1774–1780. AAMAS '23, IFAAMAS, London, United Kingdom (2023)
11. Liscio, E., Siebert, L.C., Jonker, C.M., Murukannaiah, P.K.: Value Preferences Estimation and Disambiguation in Hybrid Participatory Systems. Journal of Artificial Intelligence Research pp. 1–33 (2024)
12. Lourie, N., Le Bras, R., Choi, Y.: Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 13470–13479. AAAI '21 (2021)
13. McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: Uniform manifold approximation and projection. Journal of Open Source Software **3**(29) (2018)
14. van der Meer, M., Falk, N., Murukannaiah, P.K., Liscio, E.: Annotator-Centric Active Learning for Subjective NLP Tasks. In: Proceedings of the 2024 Conference

on Empirical Methods in Natural Language Processing. pp. 1–19. EMNLP '24, ACL, Miami, USA (2024)
15. van der Meer, M., Vossen, P., Jonker, C.M., Murukannaiah, P.K.: Do Differences in Values Influence Disagreements in Online Discussions? In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 15986–16008. EMNLP '23, ACL, Singapore (2023)
16. Park, J., Liscio, E., Murukannaiah, P.K.: Morality is Non-Binary: Building a Pluralist Moral Sentence Embedding Space using Contrastive Learning. In: Findings of the Association for Computational Linguistics: EACL 2024. pp. 654–673. ACL, St. Julian's, Malta (2024)
17. Pyatkin, V., Hwang, J.D., Srikumar, V., Lu, X., Jiang, L., Choi, Y., Bhagavatula, C.: ClarifyDelphi: Reinforced Clarification Questions with Defeasibility Rewards for Social and Moral Situations. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. pp. 11253–11271. ACL '23, Toronto, Canada (2023)
18. Reinig, I., Becker, M., Rehbein, I., Ponzetto, S.: A survey on modelling morality for text analysis. In: Findings of the Association for Computational Linguistics ACL 2024. pp. 4136–4155. ACL, Bangkok, Thailand (2024)
19. Schramowski, P., Turan, C., Andersen, N., Rothkopf, C., Kersting, K.: Large pretrained language models contain human-like biases of what is right and wrong to do. Nature Machine Intelligence **4**, 258–268 (2022)
20. Telkamp, J., Anderson, M.: The Implications of Diverse Human Moral Foundations for Assessing the Ethicality of Artificial Intelligence. Journal of Business Ethics **178**, 961–976 (2022)
21. Vida, K., Simon, J., Lauscher, A.: Values, Ethics, Morals? On the Use of Moral Concepts in NLP Research. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 5534–5554. ACL, Singapore (2023)