Understanding Trends in Moralization and Demoralization Using Natural Language

Curtis Puryear, William J. Brady, & Nour Kteily
Department of Management and Organizations, Kellogg School of Management, Northwestern University

Moral right and moral wrong often appear stable and universal, with similar principles recurring across time and place. But which values and issues are most central to morality can also vary greatly within cultures over time. Currently, a large literature has helped reveal the aspects of human morality that are relatively universal, but moral change is much less well understood. As discussions of moral issues have shifted to digital environments—with faceless strangers and concrete social rewards (e.g., likes and shares) for users with the "hottest takes"—it is especially important to understand whether and how cultural morality changes. Here we leverage a large social media corpus spanning 9 years (*N* = 10.3M tweets) to 1) quantify the degree of semantic change in morality 2) understand how the content of the moral domain has changed and 3) test whether the prevalence of moral discussions has changed. Thus, the present work offers an approach to quantifying and interpreting moral change, as well as revealing whether moralized discussions are growing more prevalent in society.

Method

Data: Every tweet from 12.8K users from 2013 to 2021 (*N* = 10.3M)

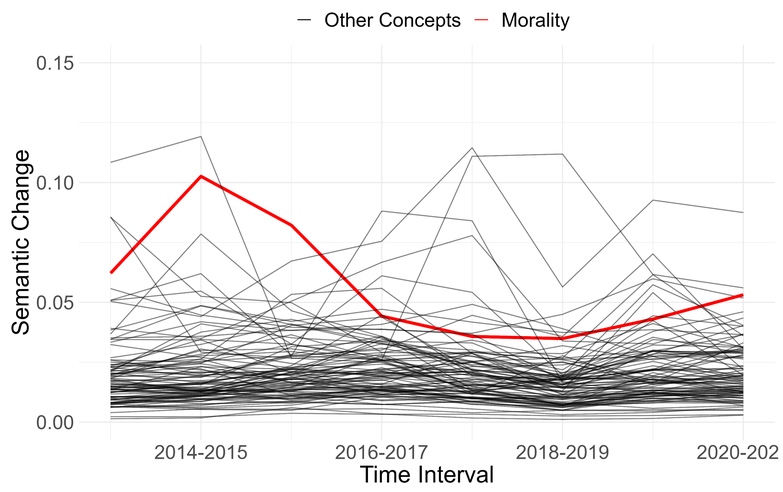
Measuring Moral Change:

- Train diachronic word embeddings the tweets, representing the meaning of concepts within each year
- Represent morality by averaging embeddings of four general morality words ("moral", "immoral", "ethical", & "unethical") within each year
- Moral change = the cosine distance between the moral concept representations between two time points

Measuring Moral Content

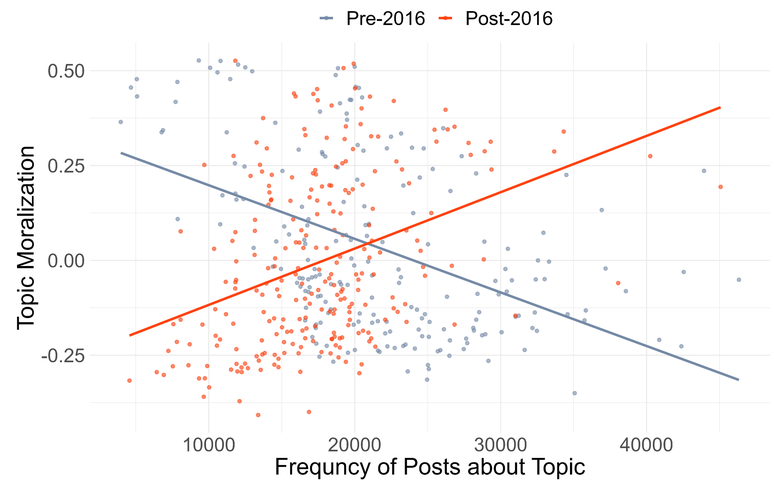
- the more semantically similar a concept is to the moral concept representation, the more it is a *moral* content
- We examine two types of content:
  - 95 concepts based upon seed-words from LIWC-22 dictionaries and from values/activities described by participants in previous work (Boyd et al., 2015)
  - 50 topics extracted from our own data via BERTopic, which extracts diachronic topics by clustering on document embeddings and then fine tunes the topics on each year within the corpus

Morality is a Dynamic Concept



Quantifying the amount of semantic change from year to year for all 95 concepts. Morality undergoes the second most semantic change on average. Examples of other concepts that consistently underwent semantic change were "Group Success", "Risk", and "Severe Weather"

Moral Discussions Became More Prevalent

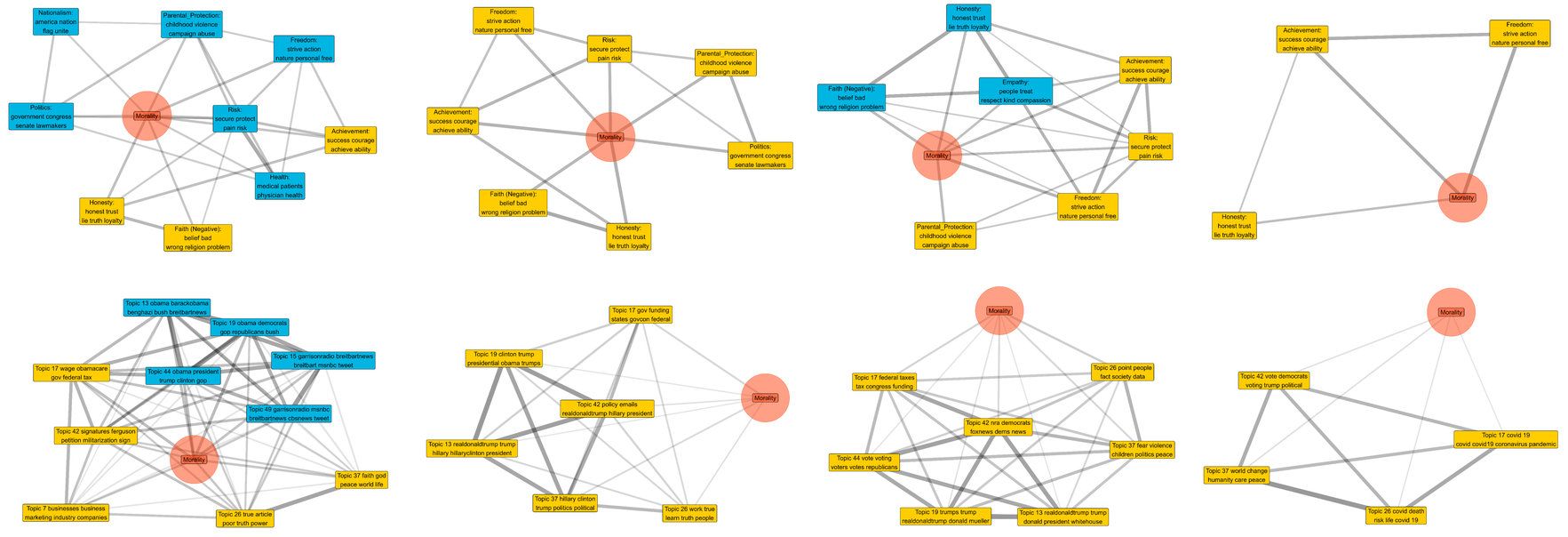


Each point represents a topic from our corpora. In red are topics from years prior to 2016 and in blue are topics from 2016 and later. The y-axis represents how semantically similar each topic is to morality. The x-axis represents how many tweets belong to each topic. Prior to 2016, posts about moralized topics were less prevalent than posts about less moralized topics, but this pattern has reversed.

Mapping Changes in Moral Content

Pre-selected Concepts

Topics from our Data



2014 **2016** **2018** **2020**

Depicted are concepts (from LIWC-22 dictionaries and activities/values from Boyd et al 2015) and discussion topics (extracted from our data via BERTopic) that were most semantically similar to a general concept representation of morality (i.e., cosine similarities > .3). Edges represent cosine similarities between concepts/topics. Colors of concepts/topics represent groups of concepts identified by a community detection algorithm.

Conclusions

Morality appears to be a relatively dynamic concept.

Semantic changes in morality appear to reflect changes in moral content

- Moral content appears to have become less diverse over time
- Moral content became more narrowly focused upon discussion of COVID and values relevant to COVID

Regardless of how the meaning of morality is changing, the prevalence of moralized discussions on social media has substantially increased in recent years

References

Boyd, R., Wilson, S., Pennebaker, J., Kosinski, M., Stillwell, D., & Mihalcea, R. (2015). Values in words: Using language to evaluate and understand personal values. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 9, No. 1, pp. 31-40).

Di Carlo, V., Bianchi, F., & Palmonari, M. (2019, July). Training temporal word embeddings with a compass. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 6326-6334).

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.