OCPSG Research Programme
Research Note
November 2025

Revisiting Moral Contagion Theory in Social Media Data

Calvin Yixiang Cheng, James Rice, Lucca Rallo Vanderchmitt, Başak Bozkurt, and Ryan Ratnam

Oxford Computational Political Science Group
Let Data Inspire.

Revisiting Moral Contagion Theory in Social Media Data

Calvin Yixiang Cheng[1], James Rice[2], Lucca Rallo Vanderchmitt[3], Başak Bozkurt[1], and Ryan Ratnam[1]

[1]Oxford Internet Institute, University of Oxford
[2]Department of Government, University of Essex
[3]Department of Methodology, LSE

Introduction to the Project

The digital transformation of communication has reshaped the ways people perceive and react to moral values in society, with online social networks serving as fast, and temporally salient conduits for the spread of moral discourse. Moral contagion theory, which posits that morally framed content exhibits increased virality, has emerged as an important framework for understanding the diffusion of moral discourse in digital environments, particularly with implications for research on political persuasion, polarization, media psychology and communication strategies (Brady, Wills, Jost, Tucker, & Van Bavel, 2017).

Despite initial empirical support for moral contagion effects, recent investigations have raised questions about the robustness and generalizability of these findings. Critical reanalyses have revealed potential methodological limitations and measurement artifacts that may have biased preivous estimates of these effects (Burton, Cruz, & Hahn, 2021). Such concerns highlight a broader challenge in computational social science: the sensitivity of theoretical conclusions to measurement approaches and analytical choices. The debate surrounding moral contagion illustrates how methodological decisions – particularly those concerning the operationalisation of measuring the spread, longevity, and dispersion of moral content – can significantly affect empirical findings and theoretical interpretations.

This paper addresses these concerns through a three-study design that systematically examines measurement error, temporal dynamics, and causal mechanisms underlying moral contagion effects. Study I focuses on the validity of moral foundations measurement by examining whether the phenomenon can be replicated using different computational approaches, ranging from dictionary-based methods to large language models (LLMs). Study II extends the analysis to a longitudinal setting, evaluating contagion effects assessing the survival rate of moral content. Study III applies double machine learning, to test the causal relationship between moral content and and their virality. In sum, this project aim to resolve the ongoing theoretical debate while providing methodological guidance for future research in computational moral psychology and political communication [1].

Study I: Measurement Bias in Moral Foundations Analysis

This study examined the extent to which measurement bias influences the observed moral moral contagion effect on Twitter (now known as X). Specifically, we compared different approaches to estimate moral foundations expressed in social media posts, including dictionaries (Graham, Haidt, & Nosek, 2009), machine learning approaches(Nguyen et al., 2023), and LLMs, benchmarked against human annotations (Hoover, Johnson, Boghrati, Graham, & Dehghani, 2020; Trager et al., 2022). Then we applied these measurement outputs to nine datasets to re-examine findings on moral contagion reported in previous research (i.e., Brady et al., 2017; Burton et al., 2021).

Dictionary-based approaches, prototypically reliant on static word lists, have been widely used to identify moral content in text (Brady et al., 2017). Despite the advantages in efficiency, interpretability and scalability, the bag-of-words approach is limited in validity and reliability in detecting moral foundations(Burton et al., 2021). More advanced computational techniques – such as machine learning, deep learning, and LLMs that are trained specifically for moral foundation detection – may provide greater

[1] Note that this is an ongoing project. As work in progress, we present an extended abstract of Study I, which has been submitted to the Political Studies Association (PSA) Annual Conference 2026.

validity and accuracy. Our first research question therefore asks: *RQ1. to what extent do machine learning- and LLM-based approaches differ from traditional dictionary-based methods in measuring moral foundations*? Next, to examine how measurement bias may influence estimates of the moral contagion effect, we applied results from different measurement approaches to re-produce results on datasets from previous studies. We therefore ask: *RQ2. to what extent do different measurement methods influence the observed moral contagion effect?*

Specifically, we benchmarked the four techniques - dictionary, MFormer, moral-strength, and GPT-4.1-nano - using a ground truth corpus that was annotated by human to answer RQ1. As for RQ2, we applied the LLMs measurement results to nine politically charged datasets including “Gun Control,” “Same-Sex Marriage,” and “Climate Change” from Brady et al. (2017), and “Women’s March,” “COVID-19,” “Mueller Report,” “MeToo,” “US Election,” and “Post-Brexit” from Burton et al. (2021). The dependent variable throughout the analysis is retweet count, and we used negative binomial regressions as our main analytical tool – consistent with the literature.

Study I: Results

The binary classification results demonstrate consistent performance across methods, with the MFormer achieving the highest F1 score. As shown in Table 1, fine-tuned, encoder only models like MFormer can achieve a comparable performance with decoder only LLMs (i.e., gpt-4.1-nano). Dictionary approaches are worse than language model approaches while retaining comparable performance to a supervised machine learning approach like Moral-Strength.

Method	Accuracy	F1 Score	Recall
MFormer	**0.76**	**0.76**	**0.76**
GPT-4.1-nano	0.72	0.71	0.72
MFD 2	0.70	0.69	0.70
MoralStrength	0.69	0.69	0.69
MFD 1	0.68	0.68	0.68

Table 1: Twitter binary classification results. MFormer outperforms other approaches.

Method	Accuracy	F1 Score	Recall
Mfd1 Standardized	**0.55**	**0.54**	**0.55**
Mf Standardized	0.53	0.55	0.53
Mfd2 Standardized	0.52	0.54	0.52
Ms Standardized	0.52	0.53	0.52
Gpt Mft	0.46	0.49	0.46
Emfd Standardized	0.26	0.19	0.26
Moralbert Standardized	0.26	0.19	0.26

Table 2: Twitter multi-class classification

In contrast, multi-class classification results present a notably different performance – see Table 2. MFD 1 dictionary emerges as the top performer at 54.8% accuracy. The top four methods (MFD 1, MFormer, MFD 2, and moral-strength (ms)) report between 51.8% and 54.8% accuracy, indicating surprising differences among the dictionary, transformer, and LLM methods. F1 scores remain relatively strong, with MFormer achieving the highest F1 at 54.5% despite not having the best accuracy, suggesting superior precision-recall balance. The GPT-4.1-nano model shows mixed accuracy compared to the binary classification (46.1% F1 score), indicating that the model performance may be particularly sensitive to the complexity of the classification task [2]. The eMFD dictionary method and MoralBERT report poor performance at 26.3% accuracy.

Our replication analyses demonstrate that moral contagion effects remain robust across measurement approaches, but effect sizes differ notably. Using LLM-based indicators, moral contagion is statistically significant in seven of nine datasets, with effect sizes substantially larger than those estimated using

[2] See also cross-domain fusion results in Guo, Mokhberian, and Lerman (2023))

dictionary methods. For example, in the COVID-19, Mueller Report, and MeToo datasets, LLM-based measures yielded effects approximately three and six times larger than those reported in previous studies (Table 3.

Dataset	**Dictionary**	**LLMs**
Women's March	1.01	1.03
COVID-19	1.15***	3.35***
Mueller Report	1.28***	3.33***
MeToo	0.91***	1.92***
US Election	1.02	1.17***
Post Brexit	1.02	2.48***
Gun Control	0.98	3.73***
Same Sex Marriage	0.99	0.97
Climate Change	1.04***	2.79***

Table 3: Comparison of the Incidence Rate Ratio (IRR) of negative binominal regressions: dictionary-based vs LLM-based measurements on moral foundation covariates, controlling for emotional and moral-emotional covariates. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. $IRR = 1$ means no effect, $IRR > 1$ refers a positive relationship, otherwise negative relationships.

Conclusion

Study I evaluated multiple computational approaches for detecting moral foundations in text and re-examined the moral contagion effect. Compared to dictionary-based methods, advanced models – particularly LLMs – more accurately captured nuanced moral framing despite its limitations in capturing nuanced moral foundations. Importantly, while measurement bias did not overturn the existence of moral contagion, it did alter effect size estimates, which were consistently larger when measured with LLMs. These results strengthen the empirical foundations of moral contagion theory and highlight the methodological advantages of LLMs in computational social science research.

References

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318.

Burton, J. W., Cruz, N., & Hahn, U. (2021). Reconsidering evidence of moral contagion in online social networks. *Nature Human Behaviour*, *5*(12), 1629–1635.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. doi:10.1037/a0015141

Guo, S., Mokhberian, N., & Lerman, K. (2023). A data fusion framework for multi-domain moral classification. In *Proceedings of icwsm*. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/22145/21924

Hoover, J., Johnson, K., Boghrati, R., Graham, J., & Dehghani, M. (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, *11*(8), 1057–1071.

Nguyen, T. D., Chen, Z., Carroll, N. G., Tran, A., Klein, C., & Xie, L. (2023). Measuring moral dimensions in social media with mformer. arXiv: 2311.10219 [cs.CL]

Trager, J., Ziabari, A. S., Davani, A. M., Golazizian, P., Karimi-Malekabadi, F., Omrani, A., ... Reyes, M., et al. (2022). The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.