1. Provide the working title of your study.

"Moral Contagion or Emotional Echo? A Preregistered Replication of Brady et al. (2017)"

2. Name the authors of this preregistration.

Niklas Bacher

3. List each research question included in this study.

- RQ1: Does moral-emotional language increase the upvote count of Reddit submissions in morally polarized subreddits?
- RQ2: Is the effect of moral-emotional language on submission diffusion stronger than the effect of purely emotional or purely moral language?
- RQ3: Does the effect of moral-emotional language on submission diffusion differ across subreddits?

4. Hypotheses

- **H1:** Reddit posts containing more moral-emotional language will receive more upvotes than posts with less moral-emotional content.
- **H2:** The effect of moral-emotional word usage on Reddit post score will be significantly stronger than the effect of purely moral or purely emotional language.
- **H3:** The effect of moral-emotional language will vary across subreddits. Specifically, the effect will be weaker or even negative in subreddits with moderation norms encouraging balanced or deliberative discourse (e.g., r/NeutralPolitics) compared to subreddits with more polarized or emotionally charged content (e.g., r/IsraelPalestine, r/AbortionDebate).

- **Framework:**

    - All hypotheses will be tested using negative binomial regression models, with score (upvotes minus downvotes) as the outcome.
    - The main predictors are counts of moral words, emotional words, moral-emotional words (intersection of the two lexica).
    - The only control variable is the tokenized length of the post, due to data availability
    - Posts with extreme length values (Token count > 10,000) are excluded as outliers to ensure convergence and avoid skewing the regression.
    - Because Reddit submissions vary greatly in length compared to tweets (which used to be capped at the time Brady et.al did their analysis), I explicitly control for this structural difference by including token_count

5. Dataset Description

Data consists of publicly available Reddit submissions from four subreddits. I selected the subreddits based on their thematic relevance to the original study and variation in discourse norms:

- r/Abortiondebate
- r/CovidVaccinated
- r/IsraelPalestine
- r/NeutralPolitics

6. Data availability

The data were obtained from a publicly available full Reddit data dump provided by the Pushshift project, specifically from this community post: https://www.reddit.com/r/pushshift/comments/1itme1k/separate_dump_files_for_the_top_40k_subreddits/. The files are freely (but somehow tedious I have to admit) available via torrent and cover historical Reddit submissions for the top 40,000 subreddits from June 2005 to December 2024.

The original LIWC Emotion dictionary is not open access and was not available for this study. Instead, I used the wordlists included in the public replication code of Burton et al. (2021), who stated that their implementation reproduced the LIWC categories used by Brady et al. (2017). The moral dictionary was taken from the Moral Foundations OSF repository.

7. How to access the data

I accessed the data using the subreddit-specific .zst compressed files shared via torrent in the Pushshift post linked above. After downloading the files for the relevant subreddits (e.g., r/AbortionDebate, r/CovidVaccinated), I decompressed and filtered posts using zstd and Python.

8. Date of download / access

- Download of .zst dumps: 29 July 2025
- Lexicon-based variable creation and filtering: 29-30 July 2025

All preprocessing was completed prior to preregistration. The final analytic dataset was locked on 31 July 2025.

9. Documentation of Data Collection

Data were collected using Pushshift's API, a widely-used archival resource for Reddit data. No scraping was conducted. Only publicly visible Reddit posts were included. Subreddits were selected to cover a mix of polarized (e.g., r/Palestine) and more deliberative (e.g.,

r/NeutralPolitics) discourse environments. Moderation policies differ across these communities and are hypothesized to influence diffusion dynamics.

10. Codebooks

Not applicable

11/12. Which variables in the dataset are you analyzing?

Dependent Variable

- score: The Reddit post score, calculated as upvotes minus downvotes. This serves as a proxy for message diffusion, analogous to retweet counts in Brady et al. (2017). Note that Reddit's scoring system differs from Twitter's, as it integrates both positive and negative evaluations. However, as there are no negative scores, the score should work as well (and is applicable for a NB Model).

Independent Variables

The following three variables are **mutually exclusive**:

- moral_count: Number of moral words in the post text, based on the MFD dictionary. Z-standardized.
- emo_count: Number of emotional words in the post text, based on LIWC-style dictionary replicated from Burton et al. (2021). Z-standardized.
- moral_emo_count: Number of words that appear in both the moral and emotional dictionaries. Z-standardized.

Control Variable

- token_count: Total number of tokens (words) in the post text (selftext). Z-standardized. This is crucial due to the wide variance in Reddit post length, unlike Twitter.

13. Inclusion / Exclusion Criteria

Inclusion:

- Only Reddit submissions (i.e.posts, not comments) from selected subreddits.
- Only posts in English (based on subreddit default language)
- Posts must contain textual content in the column *selftext.*

Exclusion:

- Posts with a token length (word count) exceeding 10000 are excluded as outliers due to their extreme leverage and to ensure model convergence in negative binomial regression.

14. Missingness/Limitations

1. **Measurement limitations:** The emotional dictionary used in this study is not the original LIWC dictionary, but a reproduction derived from the publicly available replication code of Burton et al. (2021), who claim to match the original Brady et al. (2017) implementation.
2. **Score as a proxy for diffusion:** Reddit score (upvotes minus downvotes) is not identical to Twitter retweet counts.
3. **Subreddit-level heterogeneity:** The subreddit sample is diverse in size. Sample sizes vary as follows:
    - r/AbortionDebate: 7,032 posts
    - r/CovidVaccinated: 14,028 posts
    - r/IsraelPalestine: 5,840 posts
    - r/NeutralPolitics: 4,886 posts
4. **Platform and format differences:** Reddit posts are often way longer and more nuanced than tweets. While post length is statistically controlled (via token length), the structure of Reddit discourse may influence how moral or emotional content affects diffusion.
5. **Selection bias:** Only public posts were included. Removed/deleted or highly downvoted posts may be underrepresented. Additionally, subreddit selection was purposive and may not represent the full ideological spectrum.