### Study Information

1. **Utilizing Personality to Quantify Online Echo Chambers**
2. Kiran Roye, Almog Simchon
3. Description
   1. Social media has become subject to political polarization. One of the main mechanisms theorized to spread this polarization is the echo-chamber. However, there is no consensus as to the extent to which echo-chambers drive polarization. Therefore, we will create a measurement of the strength of echo-chambers on 4 distinct Reddit communities, distinguished by levels of polarization and political-leaning, to determine which communities are most self-similar, and thus are more of an echo-chamber.
4. Hypotheses (required)
   1. The polarized communities will have a higher mean echo chamber coefficent, even the non-politically polarized community

### Design Plan

In this section, you will be asked to describe the overall design of your study. Remember that this research plan is designed to register a single study, so if you have multiple experimental designs, please complete a separate preregistration.

1. Study type
   1. Observational Study
2. Blinding (required)
   1. Blinding describes who is aware of the experimental manipulations within a study. Mark all that apply.
      1. No blinding is involved in this study.
3. Is there any additional blinding in this study?
4. Study design (required)
   1. We have a between subject design analyzing personality vectors of four reddit communities

### Sampling Plan

In this section we’ll ask you to describe how you plan to collect samples, as well as the number of samples you plan to collect and your rationale for this decision. Please keep in mind that the data described in this section should be the actual data used for analysis, so if you are using a subset of a larger dataset, please describe the subset that will actually be used in your study.

1. Existing data (required)
   1. Preregistration is designed to make clear the distinction between confirmatory tests, specified prior to seeing the data, and exploratory analyses conducted after observing the data. Therefore, creating a research plan in which existing data will be used presents unique challenges. Please select the description that best describes your situation. Please do not hesitate to contact us if you have questions about how to answer this question ([prereg@cos.io](mailto:prereg@cos.io)).
      1. Registration prior to creation of data: As of the date of submission of this research plan for preregistration, the data have not yet been collected, created, or realized.
      2. Registration prior to any human observation of the data: As of the date of submission, the data exist but have not yet been quantified, constructed, observed, or reported by anyone - including individuals that are not associated with the proposed study. Examples include museum specimens that have not been measured and data that have been collected by non-human collectors and are inaccessible.
      3. Registration prior to accessing the data: As of the date of submission, the data exist, but have not been accessed by you or your collaborators. Commonly, this includes data that has been collected by another researcher or institution.
      4. Registration prior to analysis of the data: As of the date of submission, the data exist and you have accessed it, though no analysis has been conducted related to the research plan (including calculation of summary statistics). A common situation for this scenario when a large dataset exists that is used for many different studies over time, or when a data set is randomly split into a sample for exploratory analyses, and the other section of data is reserved for later confirmatory data analysis.
      5. Registration following analysis of the data: As of the date of submission, you have accessed and analyzed some of the data relevant to the research plan. This includes preliminary analysis of variables, calculation of descriptive statistics, and observation of data distributions. Please see cos.io/prereg for more information.
2. Data collection procedures (required)
   1. Both comments and text-based posts will be collected in Python with Reddit’s API using the PRAW library. Comments and posts will be collected from a right leaning community (r/conservative), a left-leaning community (r/progressive), two non-political communities composed of individuals with a loyalty to the subject of the community (r/NYYankees) and lastly an apolitical group with theorized low group membership (r/boston). Comments will be associated with a user ID and a post ID, as well as the community. All posts on each subreddit within the 6 months, as well as all comments under these posts, will be collected. Then, to obtain a personality model for each comment and post, Contextualized Construct Representation will be performed.
3. Sample size (required)
   1. Assuming 1 post per hour on each subreddit, we will collect around 17,000 posts combined from all 4 communities. Assuming an average of 100 comments per post, we will also collect around 1.7 million comments.
4. Stopping rule (optional)
   1. If your data collection procedures do not give you full control over your exact sample size, specify how you will decide when to terminate your data collection.
   2. **Example**: We will post participant sign-up slots by week on the preceding Friday night, with 20 spots posted per week. We will post 20 new slots each week if, on that Friday night, we are below 320 participants.
   3. **More information**: You may specify a stopping rule based on p-values only in the specific case of sequential analyses with pre-specified checkpoints, alphas levels, and stopping rules. Unacceptable rationales include stopping based on p-values if checkpoints and stopping rules are not specified. If you have control over your sample size, then including a stopping rule is not necessary, though it must be clear in this question or a previous question how an exact sample size is attained.

### Variables

In this section you can describe all variables (both manipulated and measured variables) that will later be used in your confirmatory analysis plan. In your analysis plan, you will have the opportunity to describe how each variable will be used. If you have variables which you are measuring for exploratory analyses, you are not required to list them, though you are permitted to do so.

1. Measured variables (required)
   1. The single outcome variable of this study will be the echo-chamber coefficients of each comment in each community the each subreddit, measured as the cosine similarity between the personality of each comments and the post it was commented under.

### Analysis Plan

You may describe one or more confirmatory analysis in this preregistration. Please remember that all analyses specified below must be reported in the final article, and any additional analyses must be noted as exploratory or hypothesis generating.

A confirmatory analysis plan must state up front which variables are predictors (independent) and which are the outcomes (dependent), otherwise it is an exploratory analysis. You are allowed to describe any exploratory work here, but a clear confirmatory analysis is required.

1. Statistical models (required)
   1. We will use a one-way between subjects ANOVA to our data. The categorical independent variable is the community, whereas the dependent variable is the echo-chamber coefficient.
2. Exploratory analysis (optional)
   1. We expect that there may be a difference in personality measures between the different communities, therefore we will look for the differences between communities for each measure as well.