Pre-Registration - 17.08.2021

Can crowd-based user judgements against misinformation backfire?

Jonas Stein, University of Groningen
Vincenz Frey, University of Groningen
Arnout van de Rijt, European University Institute

Background

This experiment investigates whether crowd-based content veracity judgements can be an efficient way to identify misinformation (Kim et al., 2019; Pennycook et al., 2019). We hypothesize that when false information is judged by subgroups of susceptible, like-minded people – as they would be found in so called echo chambers (Del Vicario et al., 2016; Vosoughi et al., 2018) – biased judgements can grant false information with the necessary early support to later convince other, initially skeptical members of a group. Conversely, we expect that when skeptical and susceptible individuals judge the veracity of a message in an alternating order while being informed about others' previous judgements, the community provides checks-and-balances that increase subjects' propensity to correctly identify true and false messages.

Design

We let 80 bipartisan groups of 25 liberal and 25 conservative subjects judge the veracity of 20 ideologically charged true and false informational messages. Within each group, subjects judge messages in a sequential manner so that each subject starts only once the previous subject has finished judging all messages. True messages are operationalized as the central finding of a published scientific article, whereas false messages represent the inverse finding of a scientific article. Prior to the experiment, we ensure through pretesting that messages have a conservative or liberal connotation, with intentionally liberal (conservative) messages being more likely believed to be true by liberals (conservatives).

Our experiment manipulates the order in which liberal and conservative individuals contribute to judgement sequences, and whether individuals can see previous judgements or not. Through these manipulations, it provides a controlled test of how individual propensity to make correct judgements is affected when previous judgements are visible, and when individual biases in favor of or against certain messages are correlated with the order of judgements. We implement three experimental conditions: First, an independence, or control condition, in which subjects make judgements on the veracity of messages without being able to see earlier judgements in the sequence (20 independent sequences containing 25 liberal and 25 conservative subjects each). In the other two conditions, subjects act under social influence, meaning they can see what earlier judgements in a sequence had been given. The first social influence condition lets 25 susceptible subjects, whose ideology aligns with the connotation of a message, make judgements first. The 25 skeptical subjects, whose ideology diverges from the connotation of a message, make judgements after susceptible subjects made their decisions (20 sequences in which 25 conservatives judge first and 20 sequences in which 25 liberals judge first). In the other social influence condition, 25 liberals and 25 conservatives make alternating judgements, so that the sequence order is uncorrelated with subject ideology (20 sequences).

We recruit a total of 4000 liberal and conservative subjects from the United States via Amazon Mechanical Turk and Prolific, screening people for their self-identified ideology and inviting

them to our own experimental online environment. Moderates, who identify as neither liberal or conservative, are excluded from the study. Subjects are remunerated $1.5 flat-fee for their participation. The experiment is granted ethical approval by the Ethics Board of the European University Institute, Florence. Data is collected in the second half of 2021.

Hypotheses

H1: If the probability of individual, independent judgements being correct exceeds 0.5, the overall fraction of correct judgements increases in the alternating-order scenario in comparison to the independence scenario.

H2: If the probability of correct, independent judgements from susceptible individuals exceeds 0.5, the overall fraction of correct judgements in the susceptible-first scenario increases in comparison to the independence scenario.

H3: If the probability of correct, independent judgements from susceptible individuals is lower than 0.5, the overall fraction of correct judgements in the susceptible-first scenario decreases in comparison to the independence scenario.

H4: In in the susceptible-first scenario, the individual propensity of a correct judgement decreases in i's position in the sequence for susceptible individuals (H4a) and increases in i's position for skeptical individuals (H4b). This is only the case if the probability of a correct, independent judgement from a susceptible individual is lower than 0.5.

Analysis

Our central outcome variable is the fraction of correct judgements in a sequence, which is being compared across treatment conditions. To test Hypothesis 1, we compare the fraction of correct judgements of the sequences in the alternating rating condition with the fraction of correct judgements of the sequences in the independence condition. To test Hypothesis 2 and 3, we compare the fraction of correct judgements in the susceptible-first condition with the fraction of correct judgements in the independence condition.

We only compare judgements from messages meeting the scope conditions of our hypotheses. This means that for Hypothesis 1, we only use judgements from messages where the proportion of correct choices in the independence condition is greater than 0.5. For Hypothesis 2, we only use messages where the proportion of correct choices among susceptible subjects in the independence condition is greater than 0.5. For Hypothesis 3 and 4, only judgements from messages are used where the proportion of correct choices among susceptible subjects in the independence condition is smaller than 0.5.

For Hypotheses 1-3, we use non-parametric tests to compare treatment effects across conditions. For Hypothesis 4, we use a regression analysis in which we regress the fraction of correct judgements by the subjects' position in the sequence. To test for opposing effects among susceptible and skeptical subjects in Hypothesis 4, we include an interaction term of subject ideology and subjects' position in the sequence.

References

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559.

Kim, A., Moravec, P. L., & Dennis, A. R. (2019). Combating Fake News on Social Media with Source Ratings: The Effects of User and Expert Reputation Ratings. *Journal of Management Information Systems*, *36*(3), 931–968. https://doi.org/10.1080/07421222.2019.1628921

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. (2019). Understanding and reducing the spread of misinformation online. *Unpublished manuscript: https://psyarxiv. com/3n9u8*.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559