Pre-registration - How Well Does Accuracy Nudges Work on DeepFakes?

1. Introduction

Shifting attention to accuracy through nudges has been proposed as a scalable intervention to reduce the spread of online misinformation. These interventions have proven effective in reducing the spread of false information related to health and political issues (Pennycook et al., 2021a; Pennycook & Rand, 2022; Offer-Westort, Rosenzweig, & Athey, 2023). However, recent advances in generative artificial intelligence have made these systems capable of creating increasingly credible messages, known as "deepfakes", raising concerns about their potential use by malicious actors for influence operations (Goldstein, 2023; Dufour et al., 2024). *How well does accuracy nudge interventions work on DeepFakes sharing intentions?* This study aims to answer this question using a simplified version of the experimental method proposed by Pennycook, et al. (2021b).

2. Experimental Design:

Manipulations: An accuracy nudge intervention, consisting of a similar evaluation prompt used by Offer-Westort, Rosenzweig & Athey (2023), will be applied to a random half of the sample at the beginning of the survey. Participants will be presented with a neutral news headline along with the following instructions: *"Now we would like you to pretest a news headline for our future studies. We are interested in whether people think this headline is truthful or not. [placebo headline] In your opinion, is this headline truthful?"*. Asking this single question has been consistently found to improve sharing discernment, regardless of the content topic (Pennycook & Rand, 2022). Participants in the control group will not be exposed to this intervention. [Link to the nudge stimuli]

Content evaluation task: Participants will encounter four stimuli in random order: a deepfake video of a dog carrying a baby like a human, a real video of a dog interacting with a baby, a false headline related to the mpox emergency, and a real article headline also related to the mpox emergency. For each stimuli, they will be asked *"If you came across this video on the internet, would you share it with someone via Messenger, WhatsApp or any messaging app?"* and *"If you came across this video on the internet, would you share it publicly on your Facebook, Instagram, Twitter or other social media feed or stories?"*. This differentiation follows Offer-Westort et al. (2023), aknowledging that preferences for sharing contents can differ between private and public channels. They will also be asked *"What do you think about the video/publication? Leave a comment"*, with a response limit of 300 characters and a minimum of 20 characters. The responses will be analyzed to infer whether the participant believed the content or not (or if it is undefinable). Lastly, they will be asked if they have encountered the video/publication before [Links to the real and false sets of stimuli].

Measures: After the content evaluation task, reflective style will be measured using three reworded items of the CRT used by Pennycook & Rand (2019). An established question about political preferences for Colombian subjects will be asked. Participants will answer

"How often do you use the internet and social networks?" on a scale of 1 to 5. This question aims to measure digital literacy. Next, following Pennycook et al. (2021a) we will ask if they sometimes consider sharing health-related news or cute content via private and public channels. Education level will be measured by asking *"How many formal years of education do you have?"* (<5, 5-11, 12-16, >16). Lastly, people will be asked his age and gender [Link to the complete questionnaire].

Materials The survey will be implemented using Otree and Heroku for online administration. The deepfake and real video were obtained from Youtube, the real headline was retrieved from an internet search and the false headline was retrieved from a fact-checking website. The complete questionnaire and set of stimuli can be found at Deepfakes_stimuli_public.

Sample size: Power calculations following Sullivan (n.d.), at a 5% level of significance and 80% power, suggest a required sample size of n=7,743 based on effect sizes reported by Offer-Westort, Rosenzweig, & Athey (2023), and n=1,568 based on effect sizes reported by Pennycook & Rand (2022). Nonetheless, given that participant incentives are self-funded and the study have an exploratory character, I will aim to collect from 300 to 500 responses.

Participants: A random sample, balanced for gender and age, will be drawn from a database of over 5,000 individuals interested in participating in incentivized research activities. This database is managed by the E-Socials research group, of which I am a member. To ensure only one response per person, invitations will be sent with a unique identification code, valid for a single response.

Incentives: The survey is expected to take around 10 minutes. Subjects who complete the survey enter a raffle with three chances of winning $150,000 COP.

3. Hypotheses

H1: Accuracy nudge will reduce sharing intentions (and belief) of false content. Participants exposed to the accuracy nudge condition will be significantly less likely to share (and believe) the deepfake video and the misinformation headline compared to participants in the control condition. This will be measured by lower sharing intentions in private and public channels, as well as belief ratings derived from open-ended answers, for both the deepfake video and the false article headline.

H2: The accuracy nudge will have a different effect on the deepfake video versus the false headline.
The accuracy nudge will affect sharing intentions differently for the deepfake video compared to false article headlines, regardless of presentation order. This is based on the notion that the visual persuasiveness of deepfakes may influence individuals' ability to detect falsehood.

I will also explore the effect of controls, namely familiarity with the content, education, age, cognitive style, digital literacy and political preferences. Of particular interest of mine is the examination of political preferences, as I hypothesize that individuals with extremist views, whether on the left or right, may be more akin to misinformation than those with centrist views, although it is not a central objective of this study.

4. Analysis plan

Since the two types of content (dog-child videos and health-related news headlines) may inherently differ in the preferences that drive sharing behavior, they may not be directly comparable. Therefore, analyses for each content type will be conducted separately.

Outcome Variables:

1. Private sharing intentions for the deepfake video and misinformation headline (binary: 1 = willing to share, 0 = not willing to share).
2. Public sharing intentions for the deepfake video and misinformation headline (binary: 1 = willing to share, 0 = not willing to share).
3. [Optional] Belief ratings for the deepfake video and the headline (categorical: not believed, undefined, believed).

A. Descriptive statistics and correlation matrix:

I will compute and report descriptive statistics for key variables (e.g., sharing intentions, truthfulness ratings, demographic variables) in treatment, control and overall sample. Next, build a correlation matrix to examine basic relationships between variables.

B. Test H1:

All sets of analyses to test H1 will be conducted using both, the whole sample and the subsamples of people that indicated to sometimes consider sharing health-related news or cute content on social media via private and public channels, following Pennycook et al. (2021).

Bar plot and Independent Samples t-Test:

- Similar to Figure 2 of Pennycook et al. (2021), two separate bar plots with 95% IC whiskers will be plotted for both private and public sharing intentions. Each bar plot will correspond to a type of content (videos or headlines). The y axis corresponds to the likelihood to share (%), and the x axis indicates content veracity (True vs. False). Control and treatment conditions bars will be depicted using different colors (see Fig. C1 for an example).
- I will conduct independent samples t-tests to compare the mean private and public sharing intentions between the treatment and control conditions for both the videos and news headlines separately. This will allow me to assess the intervention's impact on each type of misinformation individually.

Regression Analysis:

- Perform logistic and probabilistic regression analyses to estimate the likelihood of privately and publicly sharing the deepfake video or the misinformation headline (binary outcome: share/don't share), with the treatment condition as the primary independent variable, adding controls in a hierarchical manner. Results of these analyses will be presented in a similar manner to tables 1 and 2 of Offer-Westort et al., (2023).

Effect Size:

- Cohen's d for t-tests and odds ratios for logistic regressions will be estimated to quantify the magnitude of the nudge intervention's effect.

C. Test H2:

As in the previous analyses, tests of H2 will be conducted using both, the whole sample and the subsamples of people that indicated to sometimes consider sharing health-related news or cute content on social media.

Comparisons of Effects:

- First, simple comparisons of Cohen's d and odds ratios obtained in the previous analyses will be reported.
- To directly test if the nudge has a significantly different effect on the deepfake video compared to the false article headline, a Z-test will be performed to compare the coefficients for the nudge condition across the regression models. This will allow for a formal comparison of the strength of the nudge's effect across content types.

Logistic Regression with Interaction terms

- To further explore the differences, a logistic regression model with interaction terms between the nudge condition and the type of content (deepfake vs. false headline) will be run to explore potential differential effects of the nudge on the two types of content. However, given that these content types may not be inherently comparable, this interaction model will be treated as an exploratory analysis.

Acknowledgements

I deeply thank Juan Felipe Ortiz-Riomalo, Lina María Restrepo Plaza, Allison Benson, Natalia Perez, and David Becerra for the kind comments and suggestions on this document.

References

Offer-Westort, Rosenzweig & Athey (2023) Battling the Coronavirus Infodemic Among Social Media Users in Africa | Stanford Graduate School of Business

Pennycook & Rand (2019) Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning - ScienceDirect

Pennycook et al. (2021a) - Shifting attention to accuracy can reduce misinformation online | Nature

Pennycook et al (2021b) - A Practical Guide to Doing Behavioral Research on Fake News and Misinformation | Collabra: Psychology | University of California Press (ucpress.edu)

Pennycook & Rand (2022) - Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation | Nature Communications

Sullivan (s.f.) - Power and Sample Size Determination. Boston University School of Public Health.