Re-visiting the Political Deepfake Videos Misinformation the Public But No More than Other Fake Media (2021)

Yilin Zhu, Jialin Zhao, Renjing Liu

24 2021

1 Abstract

2 Introduction

With the continuous development of Artificial intelligence and deep learning in scientific research and applications, technologies have greatly affected many aspects of our lives. Recently, the deepfake, a term derived from the combination of "deep learning" and "fake," which leverages open-source deep learning technology to produce a kind of inveracious content, have been used to fabricate video and audio recordings of political elites around the world, showing and saying things they have never done. And some reports have reported these videos could even lead to a coup attempt, posing remarkable threats and potential harm on the political democracy (Bio 2019). Moreover, some also expressed concerns that deepfake technology may also undermine our ability to detect both honest and dishonest claims in the real word (Rini 2020).

Based on these impacts and concerns caused by deepfake, Soubhik Barari, Christopher Lucas and Kevin Munger (Barari, Lucas, and Munger 2021a) compared the power of persuasion between deepfake video and an identical political scandal presented in some other comparable media formats such as textual headlines or audio recordings, and they demonstrated that deepfakes are no more effective than the same misinformation produced by other media formats, being not uniquely powerful at deception or affective manipulation (Barari, Lucas, and Munger 2021a). Additionally, they (Barari, Lucas, and Munger 2021a) also investigated the heterogeneity by participants' characteristics. It is noticeable that the authors (Barari, Lucas, and Munger 2021a) registered the older adults as the subgroups with hypothesized susceptibility to the deepfake detection, assuming they might be unable to evaluate the accuracy of digital information. And in their first experiment (exposure), which compares the attitudinal effects (the extent of fake in the clips and the change in effect to the target of the clips) of scandal clips by different subgroups of respondents, they found 'although the elderly are more likely to be effectively triggered by the Warren fake media clippings relative to those below 65, there is no detectable difference in deception nor across media.'(Barari, Lucas, and Munger 2021a)

However, in the second experiment (Detection), which examines the ability of respondents to discriminate between real and fake clippings, they (Barari, Lucas, and Munger 2021a) did not perform comparisons across different age groups, but compared the outcome differences across several motivated reasoning factors (cognitive reflection, political knowledge, digital knowledge and partisan identity) instead (Barari, Lucas, and Munger 2021a).

On the basis of what Soubhik Barari, Christopher Lucas and Kevin Munger (Barari, Lucas, and Munger 2021a) have investigated for heterogeneity by the elderly and those below 65. We adjust the code, registering the young people (18-24 years old) as the subgroups with hypothesized susceptibility to deepfake detection, to perform comparisons in outcomes of dual experiments across young people and those older than 24. The reproducing materials provided by Soubhik Barari (Barari, Lucas, and Munger 2021b), allow us to produce this narrow replication of Soubhik Barari, Christopher Lucas and Kevin Munger (Barari, Lucas, and Munger 2021a).

There are some reasons motivating us to perform this relocation. According to (Barnwell 2020), young people ageing between 18 and 24 years old, are the largest age group who engage with countless images and videos that have been edited or digitally altered. For instance, a lot of young people are exposed to ironic TikTok videos. And hence, these people are more likely to touch with deepfake every day, compared with older adults. However, we might suspect whether they are aware of or have they thought deeply about the disinformation online? Moreover, we are also wondering about their ability to discriminate between real and fake information.

We acknowledge and are grateful for the replication materials provided by the authors of Soubhik Barari, Christopher Lucas and Kevin Munger (Barari, Lucas, and Munger 2021a). The remainder of this paper is formulated as follows. Section 2 will briefly talk about the experimental design. Section 3 will generally discuss the data and model relevant to our replication. Section 4 will summarize our associated findings, and finally Section 5 will provide more discussions for our findings.

3 Experimental Design

Barari, Lucas and Munger did 2 experiments in this study. The first experiment is an exposure experiment of a 2 x 6 pairing factorial design. Some of the randomly chosen respondents are informed about deepfakes while the rest are not. Then they are all exposed to a single fictitious news feed which is randomly chosen from any one of the following 6 situations: text, audio, skit, video, ad, no clip at all (the control group), and the videos used are deepfakes (Barari, Lucas, and Munger 2021a).

During the experiment, a natural "news-feed" environment is implemented: each respondent watches 6 media clips in order and the third news feed is a fictitious one, and the rest clips are real. An actress is paid to perform as Elizabeth Warren in the skit and uses the same audio recording as the audio condition. The deepfake video is constructed from the footage used in the skit condition. This setting is very important because it enhances the internal validity of this experiment. If video conditions present different results compared to the audio and the text conditions but not with the skit condition, then this means the audiovisual information is most persuasive, no matter if the video is fictionalized or not (Barari, Lucas, and Munger 2021a).

The following experiment is a detection experiment: same respondents are asked to scroll through 8 news videos which allows us to find the within-individual deception rate. Before the task, half of the respondents are debriefed about whether they are exposed to deepfake videos in the first experiment, while the other half are not. Later, half of the subjects are also provided an accuracy prime (Barari, Lucas, and Munger 2021a). For this experiment, since subjects may watch the same deepfake videos before, an upward bias may exist here although none of them mention this in open feedback. All respondents are randomly assigned into 3 situations with different levels of deepfakes used to construct the video: 75% (high-fake), 25% (low-fake) and 0% (no-fake) (Barari, Lucas, and Munger 2021a).

4 Data and Model

The experiments described above are two survey experiments on the Lucid survey research platform, which eventually collects a sample size of 5,750 valid respondents out of 17,501 total participants in September 2020 and October 2020 (Barari, Lucas, and Munger 2021a). Thus the dataset includes the responses from these subjects in both experiments, plus any characteristics that can potentially affect the deepfake deception and appeal. Namely these parameters include some demographic information such as age, gender, education, household income, race, and ethnicity. These characteristics are hypothesized to be highly relevant to the experiment results due to their correlation with digital literacy, internet usage, political knowledge and partisanship (Barari, Lucas, and Munger 2021a).

Since this is a Lucid survey experiment, Barari, Lucas and Munger also introduce a series of "technology checks" to ensure that respondents are actually able to watch and listen to videos. Besides the technology checks, pre-experimental attention checks are also applied here to do a brief review on whether the answers to some basic demographic characteristics (e.g., gender and age) from respondents are consistent with the characteristics provided by Lucid. If the two answers do not match up, these respondents are then labeled as "low-quality" respondents, which will be dropped later when applying the statistical models as a robustness measure (Barari, Lucas, and Munger 2021a).

Another safeguard for the validity of this experiment is the representativeness of the data. Barari, Lucas, & Munger adjust the distribution of the originally collected data by using raking to calculate post-stratification weights, which tries to duplicate the demographics traits in the most recent Current Population Survey (CPS). Similar to the "low-quality" labels, weighted regression also acts as a robustness measure in later analyses. Remember that some demographics are correlated to the experiment results, so this step is very important because it eliminates the bias in results and enhances the external validity of this experiment.

Three methods in total are used in each discussion of the experiment results and they are reproduced in a very similar way as Barari, Lucas and Munger did. While their research question focuses on comparing the deepfake videos with other media formats, our research narrows it down to examine the differences between age groups. So our models compare the differences between age groups instead of which media condition the subject exposed to as Barari, Lucas and Munger did.

First there is a figure showing the value of the estimate (namely, the mean value) and the 95% confidence interval of the estimate for different age groups. The 95% confidence interval is constructed by adding and subtracting 1.96*estimate, which requires the assumption of normality. Since the data is sufficiently large, and each response is independent, so the normality assumption is relaxed here and thus this analysis is reliable. Next is a non-parametric test (t-test) to directly compare the difference in mean values among various age groups, which also leads to valid interpretation based on the same reasons as the confidence interval. Finally multiple linear regressions are performed as a robustness measure (Barari, Lucas, and Munger 2021a).

In the exposure experiment, the regression model examining the age effect on deception is estimated via (Barari, Lucas, and Munger 2021a):

$$Believei =$$

$$iAgegroupi+$$

$$Xi+$$

i

(1)

where Believe*i* is the extent of belief (from 1-5) that clipping was not fake or doctored, Agegroupi splits the sample into 5 groups based on respondents' age: 18-24, 25-34, 35-44, 45-64 and 65+, and the age group of

18-24 acts as a reference category in the regression results. Xi is a vector of covariates including the device platform, media condition (treat), gender, education, cognitive resources (CRT), measures of digital literacy, political knowledge and internet usage, if the respondent is a sexist. εi is the error term and β is a coefficient vector for the covariates. The model examining the age effect on affect is equivalent to that for measuring deception, except Favor$i$, the favorability as the outcome.

In the detection experiment, the key model (Barari, Lucas, and Munger 2021a) we use to test the effect of age on detection accuracy via the specification is also similar with what we have defined, except DetectAcc$i$, the ability to detect between real and fake clippings. Barari, Lucas, & Munger uses R, and our reproduction is also done in R (R Core Team 2020).

5 Results

```
## `summarise()` has grouped output by 'agegroup'. You can override using the `.groups` argument.
```

```
## `summarise()` has grouped output by 'type', 'group'. You can override using the `.groups` argument.
```

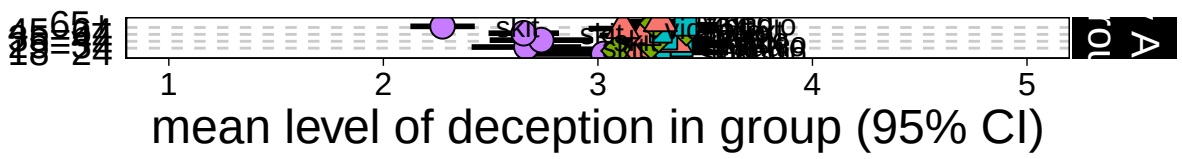
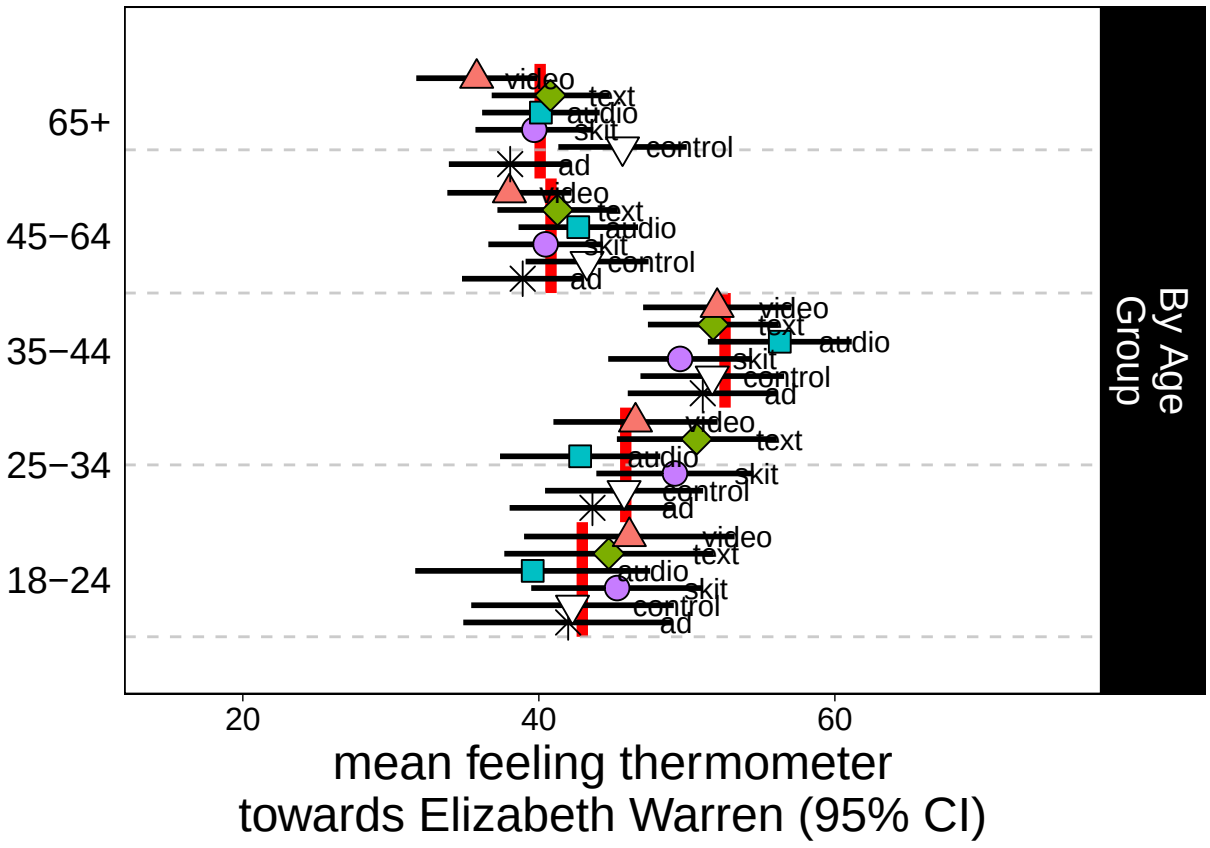


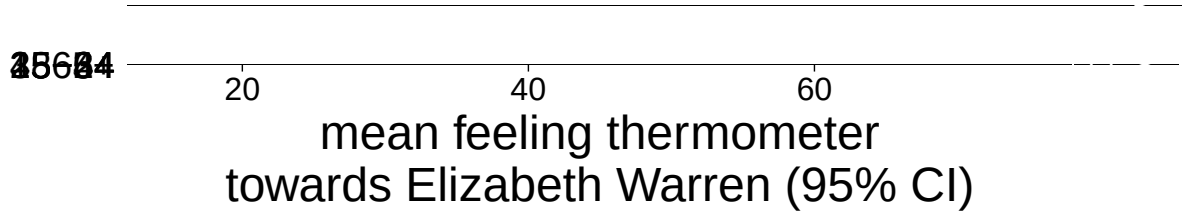
Figure 1: Marginal Means in Exposure Experiment Deception Outcomes

1) shows the mean level of deception in each age group where the outcome is the extent of believing the clipping is fake or doctored, on a scale of 1-5. The p-values are respectively 0.392, 0.7338, 0.8289, and 0.8863 when we compare the 18-24 age group with age 25-34, age 35-44, age 45-64, and age 65+. Finally the regression results of extent of belief on age groups and covariates are presented in [ref: Table 1].


```
## `summarise()` has grouped output by 'agegroup'. You can override using the `.groups` argument.
```

```
## `summarise()` has grouped output by 'type', 'group'. You can override using the `.groups` argument.
```


65+
45−64
35−44
25−34
18−24
video
text
audio
skit
control
ad
By Age Group
20
40
60
mean feeling thermometer
towards Elizabeth Warren (95% CI)



We then want to examine the favorability heterogeneity for different subgroups. From the topline null results shown in the **??**), we can see that young respondents are easier to be affectively triggered by fake clippings than participants aging between 25 and 44, especially presenting a significant difference with the 35-44 age subgroup (δ = -8.66, t= -12.16, p value <0.05). The results also indicate the same finding with what authors have investigated, showing that the elderly are more likely to be affectively triggered by the inveracious media clippings than the young people (δ =3.42, t =2.0196, p value <0.05). The results regression model [Table 2] also shows the same insights.

6 Discussion

7 Limitation and Future Works

8 Appendix

8.1 Appendix A

8.2 Appendix B

Reference

Barari, Soubhik, Christopher Lucas, and Kevin Munger. 2021a. "Political Deepfake Videos Misinform the Public, but No More Than Other Fake Media."

———. 2021b. "Political Deepfake Videos Misinform the Public, but No More Than Other Fake Media." https://github.com/soubhikbarari/Political-Deepfakes-Fx.

Barnwell, Paul. 2020. "Are Deepfake Videos a Threat to Democracy?" https://www.commonsense.org/education/articles/are-deepfake-videos-a-threat-to-democracy.

Bio. 2019. "The Bizarre and Terrifying Case of the 'Deepfake' Video That Helped Bring an African Nation to the Brink." https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rini, Regina. 2020. "Deepfakes and the Epistemic Backstop." *Philosopher's Imprint 20*.