Affect, Not Deception: Testing the Effects of Video Manipulations on Political Campaigns *

Pre-Analysis Plan

Soubhik Barari [†] and Christopher Lucas [‡] and Kevin Munger [§]

November 19, 2019

Abstract

Technology to create realistic-looking 'deepfake' videos – where existing video data is modified to change what a person appears to be saying – has recently become public. Although there have not yet been reports of the use of this technology to create political misinformation, there is widespread concern that it is inevitable, particularly in the context of the 2020 presidential election. What are the implications of deepfakes for elite communication? We hypothesize two different informational effects: a deceptive effect that misinforms citizens about elite statements, beliefs and judgments and an affective effect that primes viewers about a target elite by portraying them in a satirical, embarassing or scandalous way. Additionally, we hypothesize that the ability to distinguish deepfakes from authentic videos is correlated with politically important moderators, most notably subject age. We test these hypotheses with three experiments. In the first, we manipulate the issue position expressed by a candidate for the 2020 Democratic Presidential nomination. In the second, we create deepfakes the affectively prime subjects, to directly test the affective prime hypothesis. In the third, we ask subjects to distinguish between authentic and deepfake videos.

*We thank the Weidenbaum Center for generously funding this project.

[†]Graduate Student, Harvard University

[‡]Assistant Professor, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130; christopherlucas.org, christopher.lucas@wustl.edu

[§]Assistant Professor, Pennsylvania State University

1 Introduction

We conduct the first study of the effects of different fake news videos on political beliefs. To do so, we train fake videos of candidates for the 2020 Democratic Presidential nomination. We create deepfakes such that the informational and affective content of the video is directly manipulated. Subjects participate in either the information or affective experiments. All subjects participate in an experiment testing subjects' ability to distinguish authentic and deepfake videos, which takes place after the completion of the informational/affective experiments.

2 Theory and Context

Broadly, deepfake videos are real videos of speakers with facial and speech features realistically altered using deep learning neural networks. As of now, there is a conventional production pipeline for producing all deepfakes. First, the user must obtain a corpus of videos - preferably highly standardized with minimal stylistic variation - of the target actor. We hypothesize that politicians and news reporters are highly attractive candidates for selection, since they routinely produce standardized video content in the form of weekly addresses, press releases, campaign ads, and newsroom segments. Next, they must train the deep learning algorithm of choice in the identification of the speakers' facial features. This is the most time- and resource-intensive step, requiring either hours on a GPU-enabled computing cluster or days to weeks of high-performance computation on a standard laptop. Additionally, they may choose to either train a text-to-speech deep learning model which learns to generate realistic voice samples of the speaker from input text or simply have an

impersonator provide the voice inputs. Finally, either given footage of an impersonator's facial features, a recorded voice performance, or some other input, the deep learning model can generate a video of synthetic facial movements which can be combined with audio to produce a "deepfake" of the target actor.

Three qualities in particular distinguish deepfakes from other contemporary forms of fake news. First is medium: deepfakes present information in the form of audiovisual stimuli, in contrast to textual stimuli. Like other forms of audiovisual political media such as political ads and news commentary (Mutz and Reeves 2005; Ansolabehere and Iyengar 1997)), deepfakes have the capacity to attach affective valence to political information in a way that textual fake news cannot. Second is expressed intent: deepfakes, thus far, have been produced by government actors, satirical entertainment news organizations, computer scientists developing deep learning technology, and - in largest circulation - by "lone-wolf" unaffiliated media producers, predominantly on YouTube. Unlike much of recently circulated textual fake news which deceptively mimic the format, style, and source validity of sincere news media (Guess, Nyhan, and Reifler 2018; Allcott, Gentzkow, and Yu 2019), popular deepfakes are explicitly tagged as either satire, entertainment content, or technological demos by the producers themselves. Although many deepfakes have so far explicitly self-presented as entertainment or satirical media, others do not.[1] Moreover, experts warn that there is little stopping adversaries from using deepfakes for widespread deception. Finally, deepfakes distort the speech and actions of a single target actor. Thus misinformation effects can be two-fold: (1) the viewer is misinformed that the actor actually made the lip-synched state-

[1]See, for example: https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/

ment and/or (2) the viewer is misinformed by the factual content of the actor's statement. Just as the source of a textual fake news story might moderate information consumers' responses, the particular target actor of a deepfake may similarly moderate a viewer's response.

Taken altogether, a deepfake is distinctively characterized as a target actor lip-synching to what may be an arousing audiovisual performance generated by a media producer, either sincerely labelled as a performance or insincerely guised as a news video. Given these characteristics, we hypothesize three different attitudinal effects of exposure to political deepfakes.

The first is *deception of information*, in that a viewer sincerely believes that the deepfake video depicts a real statement by the target actor. Deception, of course, is not unique to fake videos, however experts fear that if a video is convincingly photorealistic enough, there is little capacity for factual correction post-exposure. Consequently, deepfakes may mislead viewers about the target actors' issue positions, intentions, judgments or beliefs (e.g., the viewer believes Obama is uncivil) or misinform viewers' by the falsehood of the target actors' manipulating statements (e.g., the viewer believes Obama when he says Trump is going to launch a nuclear attack).

The second is *distrust in information.* If a viewer is not deceived or they are exposed to an online video explicitly labelled as a doctored video, this exposure may still manifest in greater distrust towards all subsequently encountered political information, even from verified news sources. Prior work suggests that exposure to one popularly circulated deepfake did not particularly result in informational deception, but rather confusion and consequent distrust in sincere news media.

We propose a third attitudinal effect which is *affective priming.* Even if a deepfake neither engenders deception nor sows distrust, it can still very effectively mock, parody, or humiliate

the target actor. Alternatively, it could depict the target actor mocking, parodying, or humiliate an out-partisan or an oppositional actor. Although the viewer may identify that the deepfake is a performance, and not reality, the video may still elicit an affective response that primes them to reframe the target actor either positively or negatively depending on whether the target actor is ex-ante favored or opposed and whether they are mocking or being mocked.

3 Hypotheses

Exposure to deepfakes may informationally deceive viewers, depress their trust in all video media, or if, a particularly arousing performance is generated, prime their attitudes towards particular politicians. These effects are likely moderated by the subject's age and digital literacy. We present a few hypotheses on the relative magnitude of these effects across important dimensions:

H_1: The overall rate of deception from exposure to deepfakes will be low amongst all possible viewers. The viewer's digital literacy will negatively moderate their deception.

H_2: Reported distrust in information will be high for viewers who are able to identify that they are viewing a deepfake upon exposure.

H_3: High political knowledge viewers in the informational video experiment will be less likely to update their beliefs and less likely to be deceived.

H_5: Out-partisan viewers will more negatively evaluate a candidate when a deepfake attempts to mock/ridicule them.

To test these hypotheses, we conduct three experiments in a single survey, fielded to 5,000 online survey respondents on Lucid. In the first experiment, we manipulate the issue position of candidates in the Democratic Presidential Primaries. In the second, we test the affect hypothesis through deepfakes that make fun of the candidate. In the third, we test subjects' ability to identify deepfakes, as a test of the deception hypothesis.

Subjects participate in *either* experiment 1 (issue manipulation) or experiment 2 (affect manipulation). *All* subjects participate in the identification experiment, only after completing either experiment 1 or 2.

In the next section, we describe these three manipulations.

4 Experiment 1: Issue Position Manipulation

In this experiments, subjects watch an AARP voter guide video of one of the 2020 primary candidates being asked about their position on a particular issue. We selected a question regarding healthcare expansion since there is variation on the types of issue positions amongst the candidates. The specific prompt that we select reads:

> *How would you update and strengthen Medicare and Social Security to keep them strong for future generations?*

Upon recruitment into our experiment, we assign each subject to one of 9 treatment conditions (focusing on white male candidates for simplicity). In these conditions, subjects observe either an authentic or disenguous statement of their answer to the question noted above, a video or text presentation of that answer, and a liberal/moderate candidate. These specific questions are as follows:

4.1 Treatment Conditions

- Control Video (show the segment of the AARP video starting with question+music, the response, and then fade to black):
  - **Liberal Democrat (Sanders) true response to question:** "Well as the founder of the defending Social Security caucus the answer to that question is pretty easy, we scrap the cap and what we do is make sure we end the absurdity of somebody making millions of dollars a year today paying exactly the same amount into the Social Security trust fund as somebody making a hundred and thirty two thousand nine hundred dollars and when you do that you can, A, expand benefits for lower-income seniors many of whom are struggling on inadequate Social Security benefits and number two we extend the life of Social Security for our kids and our grandchildren by 52 years. That's what we have to do."
  - **Moderate Democrat (Biden) true response to question:** "I would make sure the people making over $400,000 pay the exact same percentage for both Medicare and Social Security that are paid for people making up to $125,000 that will raise billions of dollars over time there'd be a doughnut hole between 125 and 400 but everybody above that they would have to pay the same percentage in both areas, both Medicare and Social Security, and would increase the solvency exponentially."
- Control Text (vignette):
  - **When asked by AARP advocacy group what is your position on [TEXT**

ABOVE], presidential candidate Bernie Sanders responded: [TEXT ABOVE]

- **When asked by AARP advocacy group what is your position on [TEXT ABOVE], presidential candidate Joe Biden responded: [TEXT ABOVE]**

- Treatment Video
  - **Liberal Democrat (Sanders) fake conservative statement:** "Well as the founder of the defending Social Security caucus, I have to be honest with the American people: it is a Ponzi scheme to tell our kids that are 25 or 30 years old today, 'you're paying into a program that's going to be there.' Anybody that's for the status quo with Social Security today is telling a monstrous lie to our kids, and it's not right. That's why I'm calling to overhaul Social Security and replace it with a commonsense market-based approach."
  - **Moderate Democrat (Biden) fake conservative statement:** "I have to be honest with the American people: it is a Ponzi scheme to tell our kids that are 25 or 30 years old today that you're paying into a program that's going to be there. Anybody that's for the status quo with Social Security today is telling a monstrous lie to our kids, and it's not right. That's why I'm calling to overhaul Social Security and replace it with a commonsense market-based approach."
  - **Moderate Democrat (Biden) fake liberal statement:** "I would make sure we scrap the cap and end the absurdity that the people making over $400,000 pay the exact same percentage millions of dollars pay exactly the same amount for both Medicare and Social Security that are paid for by people making up to

$125,000. That will raise billions of dollars over time and increase the solvency exponentially. And as the country that spends the most in the developed world on bureaucracy relative to patient care, we should absolutely be guaranteeing universal healthcare and that's why I strongly endorse Medicare-for-All."

- Treatment Text
  - **When asked by AARP advocacy group what is your position on [TEXT ABOVE], presidential candidate Bernie Sanders responded: [TEXT ABOVE]**
  - **When asked by AARP advocacy group what is your position on [TEXT ABOVE], presidential candidate Joe Biden responded: [TEXT ABOVE]**

4.2 Measurement of Outcomes and Plausible Effect Moderators

Before treatment assigment, we ask the following demographic and political questions:

4.2.1 Demographic pre-treatment questions

D1: Age

D2: Gender

D3: Highest level of education

D4: How often do you use the internet

D5: How often do you use Facebook

D6: Party ID

D7: How often do you read the news online/offline

D8: How much do you trust the news you read online/offline

4.2.2 Political pre-treatment questions

P1: Feeling thermometer toward Biden, Sanders, Warren, Harris, Buttigieg

P2: Fitness for presidency for Biden, Sanders, Warren, Harris, Buttigieg

P3: What best describes your position on reforming Social Security?

After the video, we ask questions P1, P2, and P3 again, and estimate the the difference in differences for these questions across treatment condition pre/post-treatment as our primary quantity of interest.

5 Experiment 2: Affect Manipulation

In experiment 2, we test the hypothesis that deepfakes may alter political attitudes even when they do not deceive, through a political affect prime. To test this hypothesis, we exploit recordings from Saturday Night Live, in which candidates Joe Biden and Bernie Sanders were mocked by cast members, who appeared as Biden and Sanders in otherwise realistic settings.[2] In the control, subjects are assigned to the original recording, where the candidate is clearly portrayed by an actor on SNL. In the treatment, subjects are exposed to a deepfake in which the true candidate's face is transfered on to the actor's. The only manipulated component of the clip is the actor's face.

As in experiment 1, subjects are asked questions P1, P2, and P3 before and after the experiment. We estimate the effect of a face transfer deepfake after with the difference in differences.

[2]Specifically, we extract clips from this skit: https://www.youtube.com/watch?v=CBUxNeXgC70

6 Experiment 3: Ability to Identify DeepFakes

After subjects complete either experiment 1 or experiment 2, they participate in a final experiment intended to test ability to distinguish deepfake videos. At the outset of the experiment, subjects are shown the following prompt:

> *We're going show you a series of short videos involving current Democratic candidates for US President.*
>
> *As you may be aware, the technology for manipulating videos has been developing quickly, and it can sometimes be difficult to tell the difference between authentic videos and convincing-looking fakes. We'd like you to tell us which of the following videos you think have been manipulated.*
>
> *The videos that you'll be shown are randomly drawn from a larger pool, so it might be the case that some, all, or none of the videos have been manipulated. After you watch the videos, we will tell you which ones had been manipulated.*

Here, we draw from publicly released deepfakes. For example:

- Original Buzzfeed video
- White House correspondents dinner
- Trump/Obama conversation Fallon
- Coding Elite demo of Trump speech
- Trump interviews self in mirror
- Larry David / Bernie

- Bernie getting shredded
- Hillary/Kate McKinnon crying on SNL
- Obama / Trevor Noah sex talk

For each video, subjects are given a forced choice of deepfake/not deepfake.

7 Debrief

These experiments pose obvious ethical concerns. We will extensively debrief subjects upon completion of the survey, including extensive explanation of deefpakes. If subjects saw a video that misrepresented a candidate's position, they are told the candidate's true position.

It is our hope that participation in the survey educates and informs study participants, in exchange for temporary deceipt.

References

Allcott, Hunt, Matthew Gentzkow, and Chuan Yu. 2019. "Trends in the diffusion of misinformation on social media". *Research & Politics* 6 (2): 2053168019848554.

Ansolabehere, Stephen, and Shanto Iyengar. 1997. *Going negative: How political advertisements shrink and polarize the electorate*. The Free Press,

Guess, Andrew, Brendan Nyhan, and Jason Reifler. 2018. "Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign". *European Research Council* 9.

Mutz, Diana C, and Byron Reeves. 2005. "The new videomalaise: Effects of televised incivility on political trust". *American Political Science Review* 99 (1): 1–15.