**How Do “DeepFakes” Misinform? A Preliminary Research Design**

Soubhik Barari, Christopher Lucas, and Kevin Munger

**1. Motivation**

Online misinformation is a serious problem confronting democracy. The technology for creating photorealistic doctored videos using pre-trained neural networks has just - within the past year - become accessible to tech-savvy propagandists with no more than a powerful personal computer. In this project we ask: What threats do these video “DeepFakes” pose on an evidence-based political information environment and can their effects be circumvented by treatments meant to improve digital literacy?

We conduct the first study of the effects of different fake news videos on political beliefs. We do so by training fake videos of the Democratic primary candidates of the 2020 Presidential Election, using a convenient repository of candidate guide videos, in which we manipulate the informational and affective content of the speech.

**2. Theory**

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 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 (Nyhan, Guess 2018; Gentzkow 2018; Allcott, Gentzkow 2017), popular DeepFakes are explicitly tagged as either satire, entertainment content, or technological demos by the producers themselves [cite examples]. Although many DeepFakes have so far explicitly self-presented as entertainment or satirical media, others do not. 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 statement (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 (e.g., Obama lip-synching to Jordan Peele calling President Trump a ‘complete and utter dipshit’), 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 (Varrari, n.d.) 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 humiliating 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. Research Design**

We recruit 5,000 online survey respondents from Lucid. We measure the following characteristics:

* party id
* age
* gender
* digital literacy

There are two possible experiments at the first stage (issue position manipulation/AARP video *or* affect manipulation/SNL video). This is followed by a choice-task experiment in the second stage.

### **3.1. Issue Position Manipulation Experiment (Stage 1a)**

This experiment will consist of watching an [AARP voter guide video](https://www.youtube.com/watch?v=CNScC1nst5U) 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. In particular we select responses to the question:

“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 respondent to one of **10** treatment conditions (focusing on white male candidates to control for speaker race/gender):

* 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 X, presidential candidate Bernie Sanders responded: X*
  + *When asked by AARP advocacy group what is your position on X, presidential candidate Joe Biden responded: X*
* 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 X, presidential candidate Bernie Sanders responded: X*
  + *When asked by AARP advocacy group what is your position on X, presidential candidate Joe Biden responded: X*
  + *When asked by AARP advocacy group what is your position on X, presidential candidate Joe Biden responded: X*

Before showing them the video, we ask the following baseline:

**B1:** How much do you trust your online news sources?

* Strongly trust
* Moderately trust
* Kinda distrust
* Strongly distrust

**B2a:** How do you feel about candidate X?

* Strongly like
* Moderately like
* Kinda like
* Dislike(?)
* Strongly dislike

**B2b:** Rate how fit you think candidate X is to be president from a scale of Y-Z

**B3:** What do you think candidate X's position on updating Medicare / Social Security is?

**B4:** What are your views on Social Security?

**B5:** What are your views on Medicare?

After the video, we ask the following endline:

**E1:** What are your views on Social Security?

**E2:** What are your views on Medicare?

**E3:** What do you think candidate X's position on updating Medicare / Social Security is?

**E4a:** How do you feel about candidate X?

**E4b:** Rate how fit you think candidate X is to be president from a scale of Y-Z

**E5:** How much do you trust your news sources?

**E6:** Did you think the video that you saw was real?

We follow up with a strong debrief for individuals exposed to the deepfake treatment.

### **3.2. Affect Manipulation Experiment (Stage 1b)**

The idea here is that we show them video clips of Biden and Bernie from SNL: specifically clips that elicit laughter + mock them rather than make them look awesome / appealing [though we do have clips of those also!]. In case, the clips mock them for their age / Biden for touchy-feeliness.

* Biden control video ([URL of specific clip](https://youtu.be/lgA0fjztqaQ?t=325))
* Bernie control video ([URL of specific clip](https://youtu.be/lgA0fjztqaQ?t=180))
* Biden face transfer DeepFake­
* Bernie face transfer DeepFake

### **3.3. Choice Task Experiment Experiment (Stage 2)**

In this second stage, people are given instructions to identify which of X videos are deepfakes and which are real.

Examples:

* Original Buzzfeed video (<https://www.youtube.com/watch?v=cQ54GDm1eL0>)
* White House correspondents dinner (<https://www.youtube.com/watch?v=s34NWArsxVw>)
* Trump/Obama conversation Fallon (<https://youtu.be/rvF5IA7HNKc?t=15> )
* Coding Elite demo of Trump speech (<https://www.youtube.com/watch?v=8PowkVIQI0E>)
* Trump interviews self in mirror (<https://www.youtube.com/watch?v=J2yNZ0s9snU>)
* Larry David / Bernie (<https://www.youtube.com/watch?v=f0lFtwlG4N4>)
* Bernie getting shredded (<https://www.youtube.com/watch?v=Pe0nEej6l28>)
* Hillary/Kate McKinnon crying on SNL (<https://www.youtube.com/watch?v=RWZmLKw7PG8>)
* Obama / Trevor Noah sex talk (<https://youtu.be/nIU_DvjICnY?t=18>)

**4. 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. We present a few hypotheses on the relative magnitude of these effects across important dimensions:

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

- People who recognize the candidate in the AARP video experiment will be less likely to be deceived by the DeepFake itself

- High political knowledge viewers in the AARP video experiment will be less likely to update their beliefs

- Out-partisan viewers will more negatively evaluate a candidate after seeing them in the face transfer condition in the SNL experiment (but they will also negatively evaluate them in baseline)