Pre-Analysis Plan: An Experiment on the Effect of Political Deepfakes on Beliefs and Attitudes*

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

September 18, 2020

Abstract

The significance of political misinformation is widely appreciated and is the focus of much ongoing research. However, recent advances in machine learning present a new threat to the integrity of political information: the ability to digitally alter video footage to depict a political actor making a false or inflammatory statement with extreme realism. These doctored videos, or "deepfakes," are potentially more dangerous than existing sources of misinformation, given that video is generally treated as prima facie evidence. In this experiment, we develop a realistic, synthetic video, which we use to experimentally test whether and to what extent deepfake videos are effective at misinformation, compared to extant modes of political deception. We test a battery of heterogeneous behavioral effects on different 'at-risk' subpopulations in a realistic simulated news feed. Additionally, we test whether simply priming the existence of deepfakes – through benign information interventions or chance recognition 'in the wild' – reduces trust in media. Finally, we provide the first descriptive evidence of the ability of a well-informed population to distinguish deepfake videos from sincere videos. In sum, these experiments serve as the first direct tests of the threat posed by previously unseen deepfake videos, along with an evaluation of our ability to ameliorate these effects through interventions that educate and prime voters.

*We thank the Wiedenbaum Center for generously funding this experiment. For helpful comments, we thank the Political Data Science Lab and the Junior Faculty Reading Group at Washington University in St. Louis; the Imai Research Group; the Enos Research Design Happy Hour; the American Politics Research Workshop at Harvard University; and Andy Guess, Yphtach Lelkes, and Steven Webster for helpful comments. We are especially grateful to Sid Gandhi, Rashi Ranka, and the entire Deepfakeblue team for their collaboration on the production of the videos used in this project.

[†]Ph.D. Candidate, Harvard University, 1737 Cambridge St., Cambridge MA 02138; soubhikbarari.org, sbarari@g.harvard.edu

[‡]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; kevinmunger.com, kmm7999@psu.edu

1 Introduction

Political misinformation is amongst the most pressing concerns in American politics. In 2019, half of the American public viewed factually inaccurate news as a bigger problem than violent crime, climate change, and terrorism (Mitchell et al., 2019). And in the 2016 presidential election, false stories about the election received more engagement than those from credible outlets (Silverman, 2016). While misinformation's effect on the 2016 election outcome is still not well-understood, it is certain that academics, policymakers, journalists, and other stakeholders did not adequately anticipate its magnitude.

We study a new and supposedly looming threat to the integrity of political information: widespread capacity to create false videos of politicians doing and saying that which they never did nor said, colloquially termed *deepfakes*. According to popular media, security experts, and even some Congressmen, deepfakes are a near existential threat to democracy.[1] Testifying at an intelligence committee hearing, Senator Marco Rubio called deepfakes the "next wave of attacks against America and western democracies" (Rubio, 2018).

To determine the actual threat posed by deepfakes, we develop and compile a battery of novel, highly realistic deepfake videos and randomize subject exposure to these previously unseen videos. We measure different behavioral effects of exposure in comparison with existing modes of political communication of (mis)information. Finally, we benchmark the efficacy of common informational interventions thought to reduce the effects of misinformation on participant's ability to detect our novel deepfakes.

2 Theoretical Motivation

Models of electoral accountability emphasize the importance of a well-informed population of voters (Barro, 1973; Holmstrom, 1982; Ferejohn, 1986; Rogoff, 1987; Fearon, 1999; Besley, 2006). Information allows voters to accurately judge candidate attributes such as leadership,

[1]For example, Toews (2020) writes in Forbes, "Deepfakes Are Going To Wreak Havoc On Society. We Are Not Prepared" and Galston (2020) writes in Brookings, "A well-timed forgery could tip an election."

expertise, competence, character, and values in order to make principled decisions at the ballot-box (Pierce, 1993; Caprara et al., 2006; Alexander and Andersen, 1993). Misinformation, then, threatens to impair the electorate's ability to credibly evaluate their public officials (Hollyer, Rosendorff and Vreeland, 2019).

Do deepfakes pose a unique danger to an informed electorate? One view is that deepfakes are largely similar to traditional textual misinformation, for which there is an appreciable volume of research. Extant studies demonstrate that false claims that are consistent with partisan beliefs are more likely to be believed and harder to correct (Weeks, 2015; Nyhan and Reifler, 2010; Kahan, 2012; Thorson, 2016). Other work documents risk factors of susceptibility such as repeated exposure (Pennycook, Cannon and Rand, 2018), emotional state (Weeks, 2015), age (Guess, Nagler and Tucker, 2019), and lazy thinking (Pennycook and Rand, 2019). However, studies largely focus on textual information and remain agnostic about these effects with regard to the medium of information delivery. Moreover, little is known about heterogeneities in behavioral effects across different subgroups. We hypothesize that video footage can more effectively manipulate two outcomes: belief in evidence and affective response towards politicians. We discuss these outcomes and their effect heterogeneities below.

2.1 Belief in evidence

Audiovisual media is widely considered to be *prima facie* evidence. Here, we note why, and clarify how this observation underlies the expectation that deepfakes will deceive at higher rates than existing modes of misinformation.

Most obviously, it is easier to lie with words than with video. For example, while it is trivially easy to assert that one saw aliens at Roswell, it is considerably more difficult to provide video evidence in support of this sighting. More seriously, legal precedent in the United States privileges video evidence over secondhand reporting or eyewitness testimony (Wallace, 2009). Cellphone video recordings, public surveillance feeds, body-worn camera footage, and 'hot mics' regularly capture politicians or bureaucrats engaging in violence, bribery, or other

acts of wrongdoing (Reeves, 2005; Dietrych and Testa, N.d.; Fahrenthold, 2016). Journalists and criminal prosecutors regularly cite such footage to hold public officials legally accountable for their actions (Puglisi and Snyder Jr, 2011). To take a recent example, during the 2016 United States presidential election, *The Washington Post* published an authentic video of then-candidate Donald Trump graphically boasting of sexual misconduct. This quickly became the most viewed online article in the publication's history (Farhi, 2016) and was subsequently presented as evidence in later-President Donald Trump's impeachment proceedings (Baker, 2020). Video-based accountability predates the Internet: footage of former president John F. Kennedy's shooting, though partially obscured and taken from a distance, was widely accepted by broadcast news anchors, criminal investigators, and the vast majority of the public as factual evidence of his assassination (McHoskey, 1995).[2] The choice of video medium itself, thus, may serve as a "peripheral" means of persuasion, separate from the "central" route through the information being conveyed (Petty and Cacioppo, 1986).

Thus, a bad-faith actor may exploit trust in video evidence in order to misinform. Given the gold standard of video-as-evidence, a sufficiently high-quality deepfake is more likely to deceive, defame, and potentially indict a political adversary than an equivalently fabricated written report of improper conduct. This motivates our first hypothesis, which is consistent with the expectations of the popular press, politicians, and policymakers regarding the effect of deepfakes.

H$_1$: Deepfake videos will more successfully deceive subjects that a scandal occurred compared to equivalent information conveyed via audio and text.

Existing work supports the evidentiary hypothesis. Wittenberg et al. (2020) finds support for the evidentiary value of video viz-a-viz the transcript of pre-circulated videos. They do not, however, find much evidence of additional persuasive effect of video. In Section 3, we explore how our design breaks apart different elements of a deepfake video treatment to examine the scope conditions of this minimal result.

[2]Conspiracy theories surrounding J.F.K.'s assassination dispute the identity of his shooter, not the fact of his assassination.

2.2 Affective appeal

Need deepfake videos deceive – that is, persuade the viewer of a scandal that never happened – in order to impact attitudes and beliefs about the target? There is reason to believe not. Besides its evidentiary usage in watchdog journalism and legal testimony, video media is a powerful tool for activating emotion. The primary goal of Hollywood visual effects – for instance, photo-realistically depicting former President John F. Kennedy shaking hands with a fictional character – is not to persuade audiences that fictitious events actually occurred, but rather elicit affective responses through visual storytelling. Similarly, although negative campaign ads cite facts, they more successfully persuade with emotional appeals through affective language, visual frames, and musical cues (Brader, 2006). In some cases, this response demotivates electoral participation (Ansolabehere and Iyengar, 1997). While some work questions whether negative affect manifests as unfavorable attitudes towards target candidates (Lau et al., 1999; Brader, 2006; Lau, Sigelman and Rovner, 2007), others discover that negative ad exposure directly decreases candidate favorability (Fridkin and Kenney, 2011, 2004). A deepfake depicting President Barack Obama slandering a political opponent[3] may have a negative effect on his perceived character even if the video is not literally interpreted, much like an ad.

Additionally, political "infotainment" (e.g., satire, late night talk shows, comedy), the main source of political news for a large swath of Americans (Mitchell et al., 2016), is thought to engage audiences by cultivating both positive and negative emotional attachments to political figures and concepts (Baym and Holbert, N.d.; Boukes et al., 2015). Comedic impersonations that depict caricatured negative traits of politicians effectively prime viewers of those traits and can also influence viewers' electoral support (Esralew and Young, 2012).

In overview, if interpreted as a type of adversarial campaign message or political satire depicting a hypothetical scandal rather than documenting a real one, a deepfake may not deceive, but still solicit an affective response translating into either decreased motivation for

[3]See youtube.com/watch?v=cQ54GDm1eL0

electoral participation and/or lower politician favorability. The effects of deepfakes amongst the subgroup that is successfully deceived should be even higher. This logic motivates our next hypothesis.

H_2: Deepfake videos will have a larger effect on negative politician favorability compared to equivalent information conveyed via text and audio.

We note that we cannot disentangle the direct effect of exposure to a deepfake independent of that which operates through a mediating variable capturing whether or not a subject was deceived. While much recent work develops designs for the identification of such effects (Imai et al., 2011), we leave this question to future researchers.

2.3 Distrust in media

The issue of uncertainty in the democratic process dates back at least to Downs et al. (1957), who argued that "uncertainty arises among citizens because the costs of acquiring accurate information are too high." These costs have varied over time as information technologies have changed the way that citizens access information. One constant, as we have argued above, is that video could be treated as *prima facie* evidence and thus accurate information.[4] Learning about the existence of deepfakes should weaken this perception, leading to increased uncertainty in the accuracy of video information.

Following Vaccari and Chadwick (2020), we expect that information about the existence of deepfakes should decrease trust. One open question is the locus of this decrease in trust. Vaccari and Chadwick (2020) find that uncertainty prompted by deepfake exposure decrease trust "in news on social media," but does not test for effects on "trust in media."

These are conceptually distinct; in the latter, *media* refers to the central democratic institution rather than to individual "media objects" like news broadcasts or Facebook posts. A recent review piece by Schiffrin (2019) defines trust in media as how well people believe

[4]This purported objectivity of images and videos was never in fact the case, as every aspect of the selection, framing and editing of these objects involved human decisions. Without resolving any debates about the *correct* level of certainty in the information conveyed via videos, we maintain that the existence of deepfakes should cause a decrease in that level of certainty.

the media perform their role, including the tasks of selecting news and ensuring its accuracy. There is no clear analogue for "news on social media," as no central actor is playing this role of selecting and vetting news. The role of social media in democracy is evolving rapidly, but trust in news on social media is currently lower than trust in other forms of media (Newman et al., 2019).

Thus, our hypothesis consists of both an attitudinal component and behavioral component of distrust as follows:

H$_3$: Increasing the salience of deepfake videos will decrease trust in authentic video media (**H**$_{3a}$) and increase the false detection of deepfake videos (**H**$_{3b}$), compared to not increasing the salience of deepfake videos.

2.4 Heterogeneities in deepfake effects (single exposure)

We identify a number of moderator variables on the aforementioned deepfake effects on belief and affect upon exposure to a *single* deepfake (as opposed to, counterfactually, an equivalent single text or audio clipping). Specifically, we register that (1) information provision and (2) cognitive resources will ameliorate deepfake effects, while (3) forms of directional motivated reasoning about the deepfake target's identity will exacerbate deepfake effects.[5] Here, we briefly define these heterogeneities and provide theoretical justification for why we might expect effects in different directions. In this section, we are careful to note that only (1) is randomizable, therefore only (1) can be tested for *causal* moderation, while the rest are *non-causal* moderators. Indeed, the strength of the claims between non-causal and causal moderators differ as Bansak (2017) highlights.

Information provision. Though increased information about the proliferation of fake news may decrease media trust, we believe that it should also have its intended effect: decrease the likelihood that viewers believe the content of the deepfake stimulus. Indeed, past large-scale interventions to provide basic information on characteristics of fake news have proved

[5] Although other factors, such as ones described in 2.5, surely moderate single-shot deepfake effects, we register a select number of important variables that we can effectively measure pre-treatment without respondent fatigue.

successful in improving discernment (Pennycook et al., 2019; Guess et al., 2020). Therefore, we predict that exposure to information will decrease susceptibility to deception.

H_4: Deepfake videos' effect on deception will be lower for individuals provided with information about their existence prior to exposure.

Cognitive resources. Another dimension of heterogeneity that has recently been theorized to be relevant specifically to the study of online misinformation is lazy or inattentive thinking. Pennycook and Rand (2019) demonstrate that poor performance on the Cognitive Reflection Task (designed to measure the capacity to ignore initial impressions and take the time to engage higher-level thinking) is highly predictive of willingness to share fake news. Thus, we preregister the following hypothesis.

H_5: Deepfake videos' effect on deception will be smaller amongst subjects with high cognitive reflection.

Directional motivated reasoning. Directional motivated reasoning, or the selective acceptance of information based on consistency with previous beliefs, may powerfully shape how voters respond to deepfakes. We hypothesize two types of prior dispositions that may predict whether individuals are deceived by certain kinds of political deepfakes: partisan identity and sexist attitudes.

A large literature documents how *partisan identity* directs voters' attitudes about events, issues, and candidates even in the light of information that contradicts prior expectations (Kahan, 2012; Druckman and McGrath, 2019; Bolsen, Druckman and Cook, 2014; Leeper and Slothuus, 2014; Enders and Smallpage, 2019). In particular, strong partisans are likely to hold highly negative views of out-party elites and citizens (Abramowitz and Webster, 2018; Iyengar et al., 2019; McCarty, Poole and Rosenthal, 2016). However, motivated reasoning specifically, as Baker (2020) point out, requires a combination of strong partisan identification and high cognitive resources. In the absence of the latter, a partisan may not deepfake target's partisanship, not engaging in motivated reasoning. As we discuss above, performance on the CRT should negatively predict deepfake deception. However, the literature on partisan motivated

reasoning predicts that strong partisan identification combined with cognitive resources should predict greater acceptance of new information with partisan cues. Therefore, when presented with a scandal of an outpartisan politician, we hypothesize:

H_6: Deepfake videos' effect on deception will be higher (H_{6a}) and on negative affect about target will be higher (H_{6b}) amongst subjects with strong out-partisan identification and high cognitive reflection.

Additionally, voters' evaluations of candidates can be driven by negative stereotypes towards groups other than out-partisans, such as *sexist attitudes* towards women (Jamieson, Hall et al., 1995; Teele, Kalla and Rosenbluth, 2017). For example, although large swathes of the American public theoretically support female politicians for office, they evaluate them according to different criteria from men (Bauer, 2020). Moreover, a recent survey finds that, next to partisanship, holding ambivalent sexist views[6] most predicted electoral support for Trump in the 2016 election (Schaffner, MacWilliams and Nteta, 2018). One manifestation of ambivalent sexism is a belief that contemporary women violate 'benevolent' expectations placed on them: for instance, that women nowadays often display poor moral sensibility to men and outwardly display less aggression, offense, or anger (Glick and Fiske, 1996). Theories of motivated reasoning would predict that individuals holding this stereotype would be motivated to accept information that documents, and thus confirms, it. Put together, we hypothesize that documentation of a female politician behaviorally confirming a viewer's sexist stereotype will most potently deceive and affectively trigger when presented as a photo-realistic video:

H_7: Deepfake videos' effect on deception (where the target politician behaves in accordance with a sexist stereotype) will be relatively higher amongst subjects who hold those sexist stereotypes.

We note that it is logistically difficult to generate an exact counterfactual deepfake video where the target politician does not behave in accordance with the stereotype; as such, our counterfactuals are simply the same stereotype-confirming information, but conveyed via text

[6]Ambivalent sexism describes a bundle of both outright hostile (e.g., "women are physically inferior to men") and deceptively benevolent views about women (e.g., "women are objects of desire") (Glick and Fiske, 1996)

or audio. We also note that there is some initial evidence that the effects of deepfakes are larger when they are combined with micro-targeting a particular audience with susceptible prior dispositions (Dobber et al., 2020), supporting both $\mathbf{H}_6$ and $\mathbf{H}_7$.

2.5 Predictors of detection accuracy (multiple exposure)

We now consider the scenario of exposure to *multiple* deepfake stimuli embedded in a real-world video news environment such as that found on Facebook, TikTok, or Instagram. Moreover, we assume now that viewers know of the existence of deepfakes and must now distinguish deepfake videos from sincere videos. Previously mentioned moderator variables – cognitive resources, directional motivated reasoning – are all expected to predict a greater accuracy rate in deepfake detection. In addition, we stipulate two additional variables that predict how well news-seekers can detect deepfakes: accuracy salience and digital literacy.

Accuracy salience. In the context of textual fake news shared on Twitter, Pennycook et al. (2019) note that priming the concept of accuracy reduces intentions to share fake news content. We suggest that deepfake video detection may operate similarly, and thus expect that accuracy priming will increase the rates of successful deepfake detection.

$\mathbf{H}_8$: Deepfake detection accuracy will be larger amongst subjects who are primed to think about accuracy.

Digital literacy. Guess and Munger (2020) overview the concept of digital literacy, arguing that it is a sufficiently important moderator for online media effects now that we have exited the era of minimal effects and entered the era of heterogeneous effects. In the context of deepfakes, the logic by which we expect heterogeneous effects across levels of digital is simply that subjects with high digital literacy are more likely to distinguish deepfakes from video that was never altered.

Much evidence suggests that digital literacy moderates misinformation effects online. For example, Guess, Nagler and Tucker (2019) report that "users over 65 shared nearly 7 times as many articles from fake news domains as the youngest age group." during the 2016 US

Presidential election. Similarly, Barbera (2018) finds that people over 65 shared roughly 4.5 as many fake news stories on Twitter as people 18 to 24. And matching Twitter users to voter files, Osmundsen et al. (2020) find that the oldest age group was 13 times more likely to share fake news than the youngest. Applying this body of research to context of our experiment, we register the following hypothesis.

H$_9$: Deepfake detection accuracy will be larger amongst subjects with high digital literacy.

3 Research Design

We employ a survey experiment fielded to a nationally representative sample on the Lucid survey research platform. To the extent that our sample is still not representative of the American population, we expect any bias to be downward, as the moderating effect of digital literacy is likely smaller for the subset of the population that participates in an online survey pool.

Aronow et al. (2020), posted shortly before we planned to field our study, show rising rates of inattentiveness on Lucid. As a result, we include several attention checks immediately following the consent form and terminate subjects who fail them. We do so primarily because in this case, null results for our primary hypotheses are interesting, given the general popular expectation that deepfakes pose a tremendous threat to democracy. However, if our sample consists largely of inattentive subjects, we may observe null results that are due to lack of attention rather than a true null average treatment effect.

After the attention checks, subjects respond to a standard battery of demographic questions. They then enter the first stage (*exposure*) of the experiment where they are placed in a "news feed" – similar to a feed found on Facebook or Twitter – about the 2020 Democratic primary candidates, in which there may be a deepfake video. They then enter the second stage (*detection*) of the experiment where they are asked to identify deepfake videos – either before or after being debriefed about the presence of deepfake videos. In this section, we describe the experimental conditions associated with these stages. Appendix Section A describes the

creation of the videos used in the *exposure* stage.

3.1 Exposure stage

In the first stage of our experiment, we implement a 2 x 6 factorial design, after which we measure several outcomes. The two factors in our first experiment correspond first to a randomization over treatments that we expect will moderate the evaluation of and response to media, while the second factor corresponds to exposure to one of several media (or to a control condition with no media exposure).

In the first factor, the manipulation is as follows. Subjects are assigned uniformly to either a control condition (no exposure), or a condition regarding information about the existence of misinformation and of the increased technological capacity to manipulate televisual media. This stimulus is as follows:

> *During the 2016 Presidential campaign, many people learned about the risk of "fake" or "zero-credibility news": fabricated news stories posted on websites that imitated traditional news websites. While this is still a problem, there is now also the issue of digitally manipulated videos (sometimes called "deepfakes"). Tech experts are warning everyone not to automatically believe everything they read or watch online.*

In the second factor, we randomize exposure to either a control condition (no media), a fake text condition (fake news presented only as quoted text), deepfake video (the same information as the fake text condition, but with a digitally manipulated video to corroborate the story), "cheapfake" video (the video of the impersonator we hired to use as the base of the deepfake video) fake audio (the same audio as the deepfake and cheapfake conditions, but without the video), or to an attack ad about the subject presented in the previous deepfake videos. Each of these conditions is a media clipping of Senator Elizabeth Warren, a Democratic primary candidate at the time of writing[7].

[7] A key limitation of prior studies (Vaccari and Chadwick, 2020) is that deepfake stimuli used are taken directly from what was (circa 2019, when the study was conducted) the highest-quality and most famous deepfake ever created: the Jordan Peele/Obama deepfake from Buzzfeed (Silverman, N.d.). This both limits the external validity of such work and drastically limits the potential to re-use such stimuli in future studies as this particular video becomes highly recognizable.

We surround the experimentally manipulated media exposure with five media clippings, two before and three after. These reports are all real, presented either in audio, textual, or audiovisual form, and are each about a different Democratic primary candidate. The order and content of these media are fixed, and primarily serve to make mask the main manipulation in a natural "news feed"-like environment, replicating the experience of scrolling on the Facebook News Feed. The six conditions of our manipulation are:

1. (`none`) No additional media about Warren.
2. (`video`) Deepfake video and accompanying text headline of Warren (1 of following 5).
   (a) *LEAK: Elizabeth Warren calls Joe Biden "a piece of sh*t" and a pedophile in call with contributor*
   (b) *LEAK: Elizabeth Warren calls Donald Trump "a piece of sh*t" and a pedophile in call with contributor*
   (c) *LEAK: Elizabeth Warren re-claims Cherokee heritage in call with contributor*
   (d) *LEAK: Elizabeth Warren admits she doesn't "endorse the LGBTQ lifestyle" in call with contributor*
   (e) *LEAK: Elizabeth Warren flips stance on student loan debt in call with contributor*
3. (`audio`) Fake audio and accompanying text headline of Warren (1 of previous 5).
4. (`text`) Fake text headline about Warren (equivalent 1 of previous 5).
5. (`skit`) Cheapfake video (equivalent 1 of previous 5).
6. (`ad`) Negative campaign ad video about Warren.

In conditions 2-5, we randomize over multiple stories (a-e) to reduce the possibility that our results are local to a single story. The five treatment conditions are meant to simulate, though not exhaustively, different possible defamation strategies for a bad-faith actor: (a) depict incivility toward an in-party member (b) depict incivility towards an out-party member (c) prime a past controversy (d) depict a novel controversy (e) depict political insincerity. We do not register any hypotheses about heterogeneous effects across these stories within condition, but we will conduct exploratory analysis.

In the `skit` conditions, subjects are exposed to the videos used in the creation of the deepfake video, *prior* to the application of the neural network. That is, this condition displays the unaltered video of the paid actress hired to impersonate Elizabeth Warren, where the exact title of the conditions 2a-e are displayed with *"Leak"* replaced with *"Spot-On Impersonation"*. This condition represents the most conservative test of the hypothesis that deepfake videos uniquely deceive, since it is exactly like the deepfake condition, except without the "deepfaking," that is the computer-assisted falsification of a real politician from the actress performing the on-screen action. If we observe a difference between the `audio` and `text` conditions when compared to the deepfake condition, but not between the deepfake and the `skit` condition, it suggests that the mechanism is the video and not the falsification.

Finally, in the `ad` condition, subjects are exposed to a negative campaign ad titled, *"Tell Senator Warren: No Faux Casino, Pocahontas!"*, which highlights Senator Warren's supposedly illicit support for federally funding a local casino owned by an Indian tribe, despite her previous opposition to such legislation and her false claims of Cherokee heritage. Although the ad frames Warren as politically insincere, similar to condition (e) and primes the viewer of her Cherokee heritage controversy, similar to condition (c), it stylistically and informationally differs in many other ways, and thus is not an exact ad counterfactual of our deepfake. Instead, the ad simply serves as a benchmark comparison for a deepfake's affective effect, since it is an actual campaign stimuli used in the primary election to activate negative emotions towards Warren.

Denote a subject's particular first-stage condition as $\mathsf{Exposed}_i \in \{\texttt{none}, \texttt{video}, \texttt{audio}, \texttt{text}, \texttt{skit}, \texttt{ad}\}$. After exposure to each of these manipulations, we measure our primary outcomes of interest. Specifically, we measure the three following outcomes for each subject i:

- $\mathsf{Believe}_i$: Whether or not the subject believes the media they were exposed were sincere or fake/doctored [1-5],
- Favor_i: General favorability toward Senator Warren as a politician [1-100],
- $\mathsf{Distrust}_i$: To what degree the subject trusts the credibility of their media environment [1-5].

Appendix Section ?? notes the full survey text and all conditions, including the surrounding media, the questions measuring the outcomes denoted above, and distraction questions that we ask about the media which was not part of our experimental manipulation. Note that, following Huber and Arceneaux (2007) and (Brader, 2006) we measure candidate favorability using both a feeling thermometer question and a candidate vote choice question. We measure the doubt and distrust outcomes with a single question, to avoid raising suspicion about the presence of misinformation.

3.2 Detection stage

After completing the battery of questions in which we measure our primary outcomes of interest and ask another attention check question, the subjects begin a subsequent experimental task that measures ability to discriminate between real and fake videos.

Before this task, half of the subjects (in addition to all of the subjects not taking part in this task) will be debriefed from the experimental condition. The other half will be debriefed after this final task. This randomization allows us to test for the effect of the debrief itself.

Here, we employ videos created by Agarwal et al. (2019), which are of lower quality (that is, detecting the video manipulation is easier) than the ones we created, in addition to using the deepfakes we created for this subsequent manipulation. We mix these videos with similar videos taken from YouTube, and expose subjects evenly to one of three conditions: no fake videos, only low-quality fake videos, mix of quality of fake videos. Appendix Section ?? displays screenshots of each of these videos.

After completing this portion of the experiment, subjects are either debriefed regarding their condition in the exposure stage or the survey immediately concludes, depending on assignment.

4 Analysis

We now lay out the exact operationalizations, specifications and associated statistical tests for each hypothesis.[8] Unless otherwise denoted, each respondent's vector of control covariates for our outcomes of interest are given as

$$\mathbf{X}_i = \Big(\mathsf{Age}_i, \mathsf{Race}_i, \mathsf{Educ}_i, \mathsf{Gender}_i, \mathsf{PID}_i, \mathsf{DigLit}_i, \mathsf{PolKnow}_i\Big). \tag{1}$$

We index theoretically relevant parameters such as the treatment effect (τ) for each hypothesis test. For convenience, we do not index theoretically irrelevant parameters such as each error term (ϵ) or coefficient vector for the controls (β). To reduce the model dependence of our results, we expect to run additional specifications of the models stated, with the same directional hypotheses.

$\mathbf{H}_1$ (deepfake video effect on deception). First, we test a simple difference in means belief (deception) between respondents assigned to a deepfake video vs. respondents assigned to equivalent skit or audio or text clips in the first stage. We expect that these effects will be statistically greater than 0. That is, we test the alternative hypothesis that

$$\tau_{1a} = \mathbb{E}\Big[\mathsf{Believe}_i(\mathsf{Exposed}_i = \texttt{video})\Big] - \mathbb{E}\Big[\mathsf{Believe}_i(\mathsf{Exposed}_i = \texttt{text})\Big] > 0, \tag{2}$$

and

$$\tau_{1b} = \mathbb{E}\Big[\mathsf{Believe}_i(\mathsf{Exposed}_i = \texttt{video})\Big] - \mathbb{E}\Big[\mathsf{Believe}_i(\mathsf{Exposed}_i = \texttt{audio})\Big] > 0, \tag{3}$$

and

$$\tau_{1c} = \mathbb{E}\Big[\mathsf{Believe}_i(\mathsf{Exposed}_i = \texttt{video})\Big] - \mathbb{E}\Big[\mathsf{Believe}_i(\mathsf{Exposed}_i = \texttt{skit})\Big] > 0. \tag{4}$$

Additionally, we perform a parametric test adjusting for the aforementioned control covariates via the following linear model estimated via

$$\mathsf{Believe}_i = \boldsymbol{\tau}_1 \mathbf{1}_{\mathsf{Exposed}_i} + \beta \mathbf{X}_i + \epsilon_i, \tag{5}$$

[8] For hypotheses with multiple tests, we will adjust our p-values via the Benjamini-Hochberg Procedure.

where $\mathbf{1}_{\mathsf{Exposed}_i}$ is a vector of dummy variables of length 4 indicating which Warren media condition relative to `video` as the reference category that subject i is assigned.

$\mathbf{H}_2$ (deepfake video effect on affect). Equivalent to that for $\mathbf{H}_1$, except with Favor_i as the outcome.

$\mathbf{H}_{3a}$ (deepfake salience effect on media distrust). After the first stage exposure, we query our respondents about their level of trust in the media. Prior to measuring this outcome, in the context of our experiment, we argue that the idea of deepfakes can be made salient in three ways:

(I) By receiving an information prompt about deepfakes before the first stage, $\mathsf{InfoAware}_i = 1$.

(II) By recognizing that the stimulus is a deepfake in the first stage, $\mathbf{1}\{\mathsf{Exposed}_i = \texttt{video}\} \times \mathsf{Belief}_i$.

In (I) and (II), we expect this increased salience to increase the likelihood of the respondent reporting distrust in the media. As such, we perform the corresponding two tests,

$$\mathsf{Distrust}_i = \boldsymbol{\tau}_{3a_{\mathrm{I}}}\mathsf{InfoAware}_i + \beta\mathbf{X}_i + \epsilon_i, \tag{6}$$

$$\mathsf{Distrust}_i = \boldsymbol{\tau}_{3a_{\mathrm{II}}}(\mathbf{1}\{\mathsf{Exposed}_i = \texttt{video}\} \times \mathsf{Belief}_i) + \beta\mathbf{X}_i + \epsilon_i, \tag{7}$$

and register that $\boldsymbol{\tau}_{3a_{\mathrm{I}}}$ and $\boldsymbol{\tau}_{3a_{\mathrm{II}}}$ will be negative.

$\mathbf{H}_{3b}$ (deepfake salience effect on false detection). We expect that increased salience of deepfakes will increase the false detection rate of deepfakes in the detection stage of our experiment. In addition to the ways stipulated above in $\mathbf{H}_{3a}$ that deepfakes can be primed before the *exposure* stage, there are two additional ways deepfakes can be primed ahead of the *detection* stage:

(III) By being debriefed that the stimulus in the first stage was a deepfake before entering the second stage rather than at the end of the experiment, $\mathsf{DebriefBefore}_i = 1$.

(IV) By receiving an accuracy prompt directing the respondent's attention on fake news content, $\mathsf{InfoAcc}_i = 1$.

Taken together, these different ways of raising salience of deepfakes imply a series of multiplicative linear models:

$$\mathsf{DetectFPR}_i = \tau_{3b_{\mathrm{I}}}\mathsf{InfoAware}_i + \beta\mathbf{X}_i + \epsilon_i, \tag{8}$$

$$\mathsf{DetectFPR}_i = \boldsymbol{\tau}_{3b_{\mathrm{II}}}(\mathbf{1}\{\mathsf{Exposed}_i = \texttt{video}\} \times \mathsf{Belief}_i) + \beta\mathbf{X}_i + \epsilon_i, \tag{9}$$

$$\mathsf{DetectFPR}_i = \tau_{3b_{\mathrm{III}}}\mathsf{DebriefBefore}_i + \beta\mathbf{X}_i + \epsilon_i, \tag{10}$$

$$\mathsf{DetectFPR}_i = \tau_{3b_{\mathrm{IV}}}\mathsf{InfoAcc}_i + \beta\mathbf{X}_i + \epsilon_i. \tag{11}$$

We register that $\tau_{3b_{\mathrm{I}}}, \boldsymbol{\tau}_{3b_{\mathrm{II}}}, \tau_{3b_{\mathrm{III}}}, \tau_{3b_{\mathrm{IV}}}$ will all be negative.

$\mathbf{H}_4$ (heterogeneity in deception effect by information provision). Random provision of information about deepfakes ($\mathsf{InfoAware}_i = 1$) will decrease the treatment effect of deepfaking on deception. We test this via the following multiplicative model:

$$\mathsf{Believe}_i = \boldsymbol{\tau}_4^{(1)}(\mathbf{1}_{\mathsf{Exposed}_i} \times \mathsf{InfoAware}_i) + \boldsymbol{\tau}_4^{(2)}\mathbf{1}_{\mathsf{Exposed}_i} + \boldsymbol{\tau}_4^{(3)}\mathsf{InfoAware}_i + \beta_{5a}\mathbf{X}_i + \epsilon_i. \tag{12}$$

We register that that $\boldsymbol{\tau}_4^{(1)}$ will be negative. Note that since information is provided in a randomized way, we can interpret $\mathsf{InfoAware}_i$ as a causal moderator.

$\mathbf{H}_5$ (heterogeneity in deception effect by cognitive resources). We operationalize cognitive resources as a respondent's performance on the CRT (CR_i), measured prior to the exposure stage. We test the moderating effect of cognitive resources on video deepfake deception by using a multiplicative interactive linear model:

$$\mathsf{Believe}_i = \boldsymbol{\tau}_5^{(1)}(\mathbf{1}_{\mathsf{Exposed}_i} \times \mathsf{CR}_i) + \boldsymbol{\tau}_5^{(2)}\mathbf{1}_{\mathsf{Exposed}_i} + \boldsymbol{\tau}_5^{(3)}\mathsf{CR}_i + \beta\mathbf{X}_i + \epsilon_i. \tag{13}$$

Accordingly, we hypothesize that $\boldsymbol{\tau}_5^{(1)}$ will be negative.

$\mathbf{H}_{6a}$ (heterogeneity in deception effect by partisan motivated reasoning). The specification for testing partisan motivated reasoning – a combination of strong out-partisan (in this case, Republican) identity and high cognitive resources – is given as a multiplicative

interaction binary regression:

$$\begin{aligned}\mathsf{Believe}_i = &\boldsymbol{\tau}_{6a}^{(1)}(\mathbf{1}_{\mathsf{Exposed}_i} \times \mathsf{PID}_i \times \mathsf{CR}_i) + && (14)\\ &\boldsymbol{\tau}_{6a}^{(2)}(\mathbf{1}_{\mathsf{Exposed}_i} \times \mathsf{PID}_i) + \boldsymbol{\tau}_{6a}^{(3)}(\mathbf{1}_{\mathsf{Exposed}_i} \times \mathsf{CR}_i) + \boldsymbol{\tau}_{6a}^{(4)}(\mathsf{PID}_i \times \mathsf{CR}_i) + && (15)\\ &\boldsymbol{\tau}_{6a}^{(5)}\mathbf{1}_{\mathsf{Exposed}_i} + \boldsymbol{\tau}_{6a}^{(6)}\mathsf{PID}_i + \boldsymbol{\tau}_{6a}^{(7)}\mathsf{CR}_i + \beta_{6a}\mathbf{X}_i + \epsilon_i && (16)\end{aligned}$$

where $\mathbf{X}_i$ is the same as before but not does not include PID_i. $\boldsymbol{\tau}_{6a}^{(1)}$ is the moderating effect of partisan motivated reasoning on deepfake deception, which we hypothesize to be positive. Note that $\boldsymbol{\tau}_{6a}^{(1)}$ cannot be interpreted as a causal moderator.

$\mathbf{H}_{6b}$ (heterogeneity in favorability effect by partisan motivated reasoning). As above, except with Favor_i as the outcome.

$\mathbf{H}_7$ (heterogeneity in favorability effect by sexist motivated reasoning). As $\mathbf{H}_{6a}$, except with $\mathsf{AmbivalentSexism}_i$, instead of $\mathsf{PID}_i \times \mathsf{CR}_i$ as the moderator, a pre-treatment measure from 1-5 of a respondent's ambivalent sexism – modified for brevity from Glick and Fiske (1996) to minimize survey fatigue as the outcome and priming.

$\mathbf{H}_8$ (positive effect of accuracy salience on detection accuracy). We test via the specification:

$$\mathsf{DetectAcc}_i = \tau_8\mathsf{InfoAcc}_i + \beta\mathbf{X}_i + \epsilon_i, \tag{17}$$

and hypothesize that τ_8 will be positive.

$\mathbf{H}_9$ (positive effect of digital literacy on detection accuracy). Here, we conceptualize digital literacy as knowledge of digital technologies and applications such as social media sites and mobile devices. We ask a series of questions about such technologies prior to respondents being entered into the detection stage and grade their digital literacy as DigLit_i [0-10]. We test our hypothesis via the specification:

$$\mathsf{DetectAcc}_i = \tau_9\mathsf{DigLit}_i + \beta\mathbf{X}_i + \epsilon_i, \tag{18}$$

and register that τ_9 will be positive.

5 Ethics

We highlight the ethical considerations pursuant to a study that uses stimuli which we expect to be uniquely deceptive.

First, in addition to the subjects randomly assigned to a debrief in the middle of the survey, we extensively debrief all subjects at the completion of the survey. This debrief goes beyond the standard description of study procedures. We require respondents to type out the following phrase, depending on which experimental arm they were assigned to:

> *"The [video/audio/text] about Elizabeth Warren is false."*

Second, to minimize the risk of influencing the proximate election, we opted to make a deepfake of high-profile 2020 Democratic Presidential candidate who was not ultimately selected as the nominee. Elizabeth Warren is a salient politician, making our experiment more ecologically valid than one with a low-profile or hypothetical politician, but she is slated for re-election until 2024. We selected a female candidate because women are more likely to be the targets of non-political deepfakes, and we specifically test for whether pre-existing prejudice against women among subjects changes the effect of the deepfake. Two of the treatments do refer to Presidential nominees Trump and Biden, but since they are otherwise identical, any effects they produce would be offset.

Third, we carefully weigh the risks to subjects against the potential risks that may be averted with the knowledge gained through our experiment. The potential long-term consequences of exposure to a single piece of media are minimal. That is, participants are unlikely to change their political behavior as a response to treatment, given our extensive debrief. Given that we have no experimental evidence either way, it is at least as likely that our experiment will *benefit* subjects as cause harm. The experiment gives subjects experience detecting fake media, followed up by the debrief which contains feedback and information about how the deepfake process works. Given the importance and seeming inevitability of more deepfakes in

the future, and the uncertainty around their effects, we argue that academics in fact have an "obligation to experiment" (Ko, Mou and Matias, 2016). We believe that improved understanding of how deepfakes function and evidence from our low-cost interventions will in fact serve to prevent real-world harms from deepfakes in the future.

Finally, a similar argument applies to the knowledge we generate from the perspective of policy-makers, journalists, and election administrators (Agarwal et al., 2019). More specifically, our study can inform future legislation or platform policies designed to minize the threat posed by this technology.[9]

References

Abadi, Martín, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. pp. 265–283.

Abramowitz, Alan I and Steven W Webster. 2018. "Negative partisanship: Why Americans dislike parties but behave like rabid partisans." *Political Psychology* 39:119–135.

Agarwal, Shruti, Hany Farid, Yuming Gu, Mingming He, Koki Nagano and Hao Li. 2019. Protecting world leaders against deep fakes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. pp. 38–45.

Alexander, Deborah and Kristi Andersen. 1993. "Gender as a Factor in the Attribution of Leadership Traits." *Political Research Quarterly* 46(3):527–545.

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

Aronow, Peter M., Josh Kalla, Lilla Orr and John Ternovsk. 2020. "Evidence of Rising Rates of Inattentiveness on Lucid in 2020.".
URL: *https://osf.io/preprints/socarxiv/8sbe4/*

Baker, Peter. 2020. "Now Testifying for the Prosecution: President Trump." *New York Times* .

Bansak, Kirk. 2017. "A Generalized Framework for the Estimation of Causal Moderation Effects with Randomized Treatments and Non-Randomized Moderators." *arXiv preprint arXiv:1710.02954* .

[9] See `SB 6513` introduced in the WA state legislature at the time of writing, intended to restrict the use of deepfake audio or visual media in campaigns for elective office.

Barbera, Pablo. 2018. Explaining the spread of misinformation on social media: Evidence from the 2016 US presidential election. In *Symposium: Fake News and the PoliticsofMisinformation. APSA*.

Barro, Robert J. 1973. "The control of politicians: an economic model." *Public choice* pp. 19–42.

Bauer, Nichole M. 2020. "Shifting standards: How voters evaluate the qualifications of female and male candidates." *The Journal of Politics* 82(1):1–12.

Baym, Geoffrey and R Lance Holbert. N.d. Beyond Infotainment. In *The Oxford Handbook of Electoral Persuasion*.

Besley, Timothy. 2006. *Principled agents?: The political economy of good government*. Oxford University Press on Demand.

Bolsen, Toby, James N Druckman and Fay Lomax Cook. 2014. "The influence of partisan motivated reasoning on public opinion." *Political Behavior* 36(2):235–262.

Boukes, Mark, Hajo G Boomgaarden, Marjolein Moorman and Claes H De Vreese. 2015. "At odds: Laughing and thinking? The appreciation, processing, and persuasiveness of political satire." *Journal of Communication* 65(5):721–744.

Brader, Ted. 2006. *Campaigning for hearts and minds: How emotional appeals in political ads work*. University of Chicago Press.

Caprara, Gian Vittorio, Shalom Schwartz, Cristina Capanna, Michele Vecchione and Claudio Barbaranelli. 2006. "Personality and politics: Values, traits, and political choice." *Political psychology* 27(1):1–28.

Dietrych, Bryce and Paul Testa. N.d. "Seeing is Believing: How video of police action affects criminal justice beliefs." *Working Paper*. Forthcoming.

Dobber, Tom, Nadia Metoui, Damian Trilling, Natali Helberger and Claes de Vreese. 2020. "Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?" *The International Journal of Press/Politics* p. 1940161220944364.

Downs, Anthony et al. 1957. "An economic theory of democracy.".

Druckman, James N and Mary C McGrath. 2019. "The evidence for motivated reasoning in climate change preference formation." *Nature Climate Change* 9(2):111–119.

Enders, Adam M and Steven M Smallpage. 2019. "Informational Cues, Partisan-Motivated Reasoning, and the Manipulation of Conspiracy Beliefs." *Political Communication* 36(1):83–102.

Esralew, Sarah and Dannagal Goldthwaite Young. 2012. "The influence of parodies on men-

tal models: Exploring the Tina Fey–Sarah Palin phenomenon." *Communication Quarterly* 60(3):338–352.

Fahrenthold, David A. 2016. "Trump recorded having extremely lewd conversation about women in 2005." *The Washington Post* .

Farhi, Paul. 2016. "A caller had a lewd tape of Donald Trump. Then the race to break the story was on." *The Washington Post* .

Fearon, James D. 1999. "Electoral accountability and the control of politicians: selecting good types versus sanctioning poor performance." *Democracy, accountability, and representation* 55:61.

Ferejohn, John. 1986. "Incumbent performance and electoral control." *Public choice* pp. 5–25.

Fridkin, Kim L and Patrick Kenney. 2011. "Variability in citizens' reactions to different types of negative campaigns." *American Journal of Political Science* 55(2):307–325.

Fridkin, Kim Leslie and Patrick J Kenney. 2004. "Do negative messages work? The impact of negativity on citizens' evaluations of candidates." *American Politics Research* 32(5):570–605.

Galston, William A. 2020. "Is Seeing Still Believing? The Deepfake Challenge to Truth in Politics." *Brookings. January* 8.

Glick, Peter and Susan T Fiske. 1996. "The Ambivalent Sexism Inventory: Differentiating Hostile and Benevolent Sexism." *Journal of personality and social psychology* 70(3):491.

Guess, Andrew, Jonathan Nagler and Joshua Tucker. 2019. "Less than you think: Prevalence and predictors of fake news dissemination on Facebook." *Science advances* 5(1):eaau4586.

Guess, Andrew and Kevin Munger. 2020. "To See What's in Front of One's Screen: Digital Literacy and Online Political Behavior." *OSF preprint* .

Guess, Andrew M., Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler and Neelanjan Sircar. 2020. "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India." *Proceedings of the National Academy of Sciences* 117(27):15536–15545. Publisher: National Academy of Sciences ˙eprint: https://www.pnas.org/content/117/27/15536.full.pdf.
URL: *https://www.pnas.org/content/117/27/15536*

Hollyer, James R, B Peter Rosendorff and James Raymond Vreeland. 2019. "Transparency, protest and democratic stability." *British Journal of Political Science* 49(4):1251–1277.

Holmstrom, Bengt. 1982. "Moral hazard in teams." *The Bell Journal of Economics* pp. 324–340.

Huber, Gregory A and Kevin Arceneaux. 2007. "Identifying the persuasive effects of presidential advertising." *American Journal of Political Science* 51(4):957–977.

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppei Yamamoto. 2011. "Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies." *American Political Science Review* pp. 765–789.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra and Sean J Westwood. 2019. "The origins and consequences of affective polarization in the United States." *Annual Review of Political Science* 22:129–146.

Jamieson, Kathleen Hall, Kathleen Hall et al. 1995. *Beyond the double bind: Women and leadership*. Oxford University Press on Demand.

Kahan, Dan M. 2012. "Ideology, motivated reasoning, and cognitive reflection: An experimental study." *Judgment and Decision making* 8:407–24.

Ko, Allan, Merry Mou and Nathan Matias. 2016. "The Obligation To Experiment." *Medium* .

Lau, Richard R, Lee Sigelman, Caroline Heldman and Paul Babbitt. 1999. "The effects of negative political advertisements: A meta-analytic assessment." *American Political Science Review* 93(4):851–875.

Lau, Richard R, Lee Sigelman and Ivy Brown Rovner. 2007. "The effects of negative political campaigns: a meta-analytic reassessment." *Journal of Politics* 69(4):1176–1209.

Leeper, Thomas J and Rune Slothuus. 2014. "Political parties, motivated reasoning, and public opinion formation." *Political Psychology* 35:129–156.

Makhzani, Alireza, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow and Brendan Frey. 2015. "Adversarial autoencoders." *arXiv preprint arXiv:1511.05644* .

McCarty, Nolan, Keith T Poole and Howard Rosenthal. 2016. *Polarized America: The dance of ideology and unequal riches*. mit Press.

McHoskey, John W. 1995. "Case closed? On the John F. Kennedy assassination: Biased assimilation of evidence and attitude polarization." *Basic and Applied Social Psychology* 17(3):395–409.

Mitchell, Amy, Elisa Shearer, Jeffrey Gottfried and Michael Barthel. 2016. "Where Americans Are Getting News About the 2016 Presidential Election." *Pew Research Center* .
URL: *http://www.journalism.org/2016/02/04/the-2016-presidential-campaign-a-news*

Mitchell, Amy, Jeffrey Gottfried, Sophia Fedeli, Galen Stocking and Mason Walker. 2019. "Many Americans say made-up news is a critical problem that needs to be fixed." *Pew Research Center. June* 5:2019.

Newman, Nic, Richard Fletcher, Antonis Kalogeropoulos and Rasmus Nielsen. 2019. *Reuters institute digital news report 2019*. Vol. 2019 Reuters Institute for the Study of Journalism.

Nyhan, Brendan and Jason Reifler. 2010. "When corrections fail: The persistence of political misperceptions." *Political Behavior* 32(2):303–330.

Osmundsen, Mathias, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann and Michael Bang Petersen. 2020. "Partisan polarization is the primary psychological motivation behind "fake news" sharing on Twitter.".

Pennycook, Gordon and David G Rand. 2019. "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning." *Cognition* 188:39–50.

Pennycook, Gordon, Tyrone D Cannon and David G Rand. 2018. "Prior exposure increases perceived accuracy of fake news." *Journal of experimental psychology: general* 147(12):1865.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles and David G Rand. 2019. "Understanding and reducing the spread of misinformation online.".

Petty, Richard E and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*. Springer pp. 1–24.

Pierce, Patrick A. 1993. "Political sophistication and the use of candidate traits in candidate evaluation." *Political psychology* pp. 21–35.

Puglisi, Riccardo and James M Snyder Jr. 2011. "Newspaper coverage of political scandals." *The Journal of Politics* 73(3):931–950.

Reeves, Philip. 2005. "Media 'Stings' Create Scandal in India." *NPR* .
URL: *https://www.npr.org/templates/story/story.php?storyId=5050300*

Rogoff, Kenneth S. 1987. "Equilibrium political budget cycles.".

Rubio, Marco. 2018. "At Intelligence Committee Hearing, Rubio Raises Threat Chinese Telecommunications Firms Pose to U.S. National Security.".
URL: *https://www.youtube.com/watch?v=DTFNAH7NTWQ*

Schaffner, Brian F, Matthew MacWilliams and Tatishe Nteta. 2018. "Understanding white polarization in the 2016 vote for president: The sobering role of racism and sexism." *Political Science Quarterly* 133(1):9–34.

Schiffrin, Anya. 2019. Credibility and Trust in Journalism. In *Oxford Research Encyclopedia of Communication*.

Silverman, Craig. 2016. "This analysis shows how viral fake election news stories outperformed real news on Facebook." *BuzzFeed news* 16.

Silverman, Craig. N.d. "How to spot a deepfake like the Barack Obama–Jordan Peele video." *BuzzFeed*. Forthcoming.
URL: https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-d

Teele, Dawn, Joshua Kalla and Frances McCall Rosenbluth. 2017. "The ties that double bind: social roles and women's underrepresentation in politics." *Available at SSRN 2971732* .

Thorson, Emily. 2016. "Belief echoes: The persistent effects of corrected misinformation." *Political Communication* 33(3):460–480.

Toews, Rob. 2020. "Deepfakes Are Going To Wreak Havoc On Society. We Are Not Prepared.".
URL: https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared

Vaccari, Cristian and Andrew Chadwick. 2020. "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news." *Social Media+ Society* 6(1):2056305120903408.

Wallace, Sue. 2009. "Watchdog or witness? The emerging forms and practices of videojournalism." *Journalism* 10(5):684–701.

Weeks, Brian E. 2015. "Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation." *Journal of communication* 65(4):699–719.

Wittenberg, Chloe, Jonathan Zong, David Rand et al. 2020. "The (Minimal) Persuasive Advantage of Political Video over Text.".

A Developing deepfakes

Deepfakes that swap the face of a target (e.g., President Barack Obama) with an actor (e.g., Hollywood actor Jordan Peele) are synthesized via a particular class of artificial neural networks called Adversarial Autoencoders (Makhzani et al., 2015).

The deepfaker's task is to train two autoencoders to accurately represent (encode) the two respective faces in a latent space and accurately reconstruct (decode) them as images. Let $\mathbf{X}_{\text{target}}$ denote a set of facial images of the target and $\mathbf{X}_{\text{actor}}$ denote a set of facial images of the actor. Denoting $\mathcal{G}_{\text{target}}$ as the function for the target autoencoder and $\mathcal{G}_{\text{actor}}$ as the function for the actor autoencoder, the networks are structured as $\mathcal{G}_{\text{target}}(x) = \delta_{\text{target}}\{\pi(x)\}$ and $\mathcal{G}_{\text{actor}}(x') = \delta_{\text{actor}}\{\pi(x')\}$ where π is an encoder subnetwork, δ_{target} and δ_{actor} are the decoder subnetworks for the target and actor respectively, and $x \in \mathbf{X}_{\text{target}}, x' \in \mathbf{X}_{\text{actor}}$. Both autoencoders share

an encoder function π which discover a common latent representation for the targets' and actors' faces; separate decoders are charged with realistically reconstructing the input faces. The objective function to be optimized is:

$$\min_{\substack{\pi, \\ \delta_{\text{target}}, \\ \delta_{\text{actor}}}} \mathbb{E}_{x \sim \mathbf{X}_{\text{target}}} \left[||\delta_{\text{target}}\{\pi(x)\} - x||^2 \right] + \mathbb{E}_{x' \sim \mathbf{X}_{\text{actor}}} \left[||\delta_{\text{actor}}\{\pi(x')\} - x'||^2 \right] \tag{19}$$

To produce a deepfake given a audiovisual performance of the actor with respective facial image frames $\mathbf{Y}_{\text{actor}} = \left[y_1, \ldots, y_N \right]$, we input the frames into the trained target autoencoder which outputs $\mathbf{Y}_{\text{actor}} = \left[\delta_{\text{target}}\{\pi(y_1)\}, \ldots, \delta_{\text{target}}\{\pi(y_N)\} \right]$ that can be recombined with the audio of the actor's performance.

To maximize the realism of outputs created from actor inputs fed to the target autoencoder, we train a third discriminator neural network $\mathcal{D}$ which aims to accurately classify the latent representations of images as belonging to either the target or actor. The final adversial objective is given as:

$$\begin{aligned} \max_{\mathcal{D}} \min_{\substack{\pi, \\ \delta_{\text{target}}, \\ \delta_{\text{actor}}}} & \mathbb{E}_{x \sim \mathbf{X}_{\text{target}}} \left[||\delta_{\text{target}}\{\pi(x)\} - x||^2 \right] + \mathbb{E}_{x' \sim \mathbf{X}_{\text{actor}}} \left[||\delta_{\text{actor}}\{\pi(x')\} - x'||^2 \right] \\ & + \quad \mathbb{E}_{x'' \sim \mathbf{X}} \left[||\mathcal{D}\{\pi(x'')\} - \mathbf{1}\{x'' \in \mathbf{X}_{\text{actor}}\}||^2 \right] \end{aligned} \tag{20}$$

Optimization of this objective function can be performed via alternating iterative updating of the two networks' weights using stochastic gradient descent. After sufficient rounds of training, the target autoencoder can accurately reproduce the target's face using images of only the actor's face and is thus able to effectively 'fool' the discriminator. Figure 1 graphically illustrates the resulting procedure for producing deepfake face-swap videos.

In practice, this workflow for deepfake synthesis is implemented using the `TensorFlow` library (Abadi et al., 2016). Deepfake producers utilize code from several popular public code repositories which implement variants of this base framework – including multiple discriminators and autoencoders, regularization schemes, and particular network architecture choices.

In collaboration with an industry partner[10], we produced a series of deepfake videos using target footage of 2020 presidential candidate Elizabeth Warren and actor performances of a professional Elizabeth Warren impersonator. We describe the content of the performances which are used in the first stage of our experiment in the next section.

[10] See https://dfblue.com/

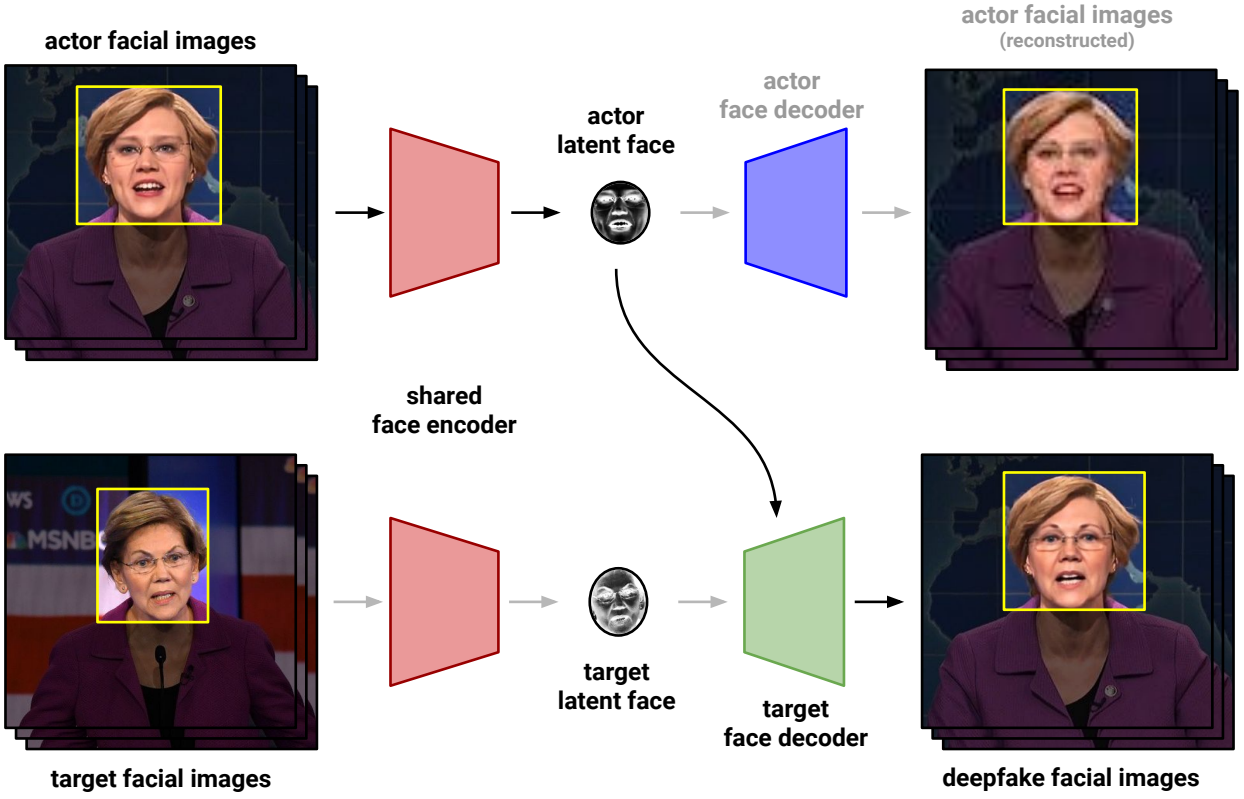

Figure 1: **Deepfake Image Frame Generation via Face-swap.** This graphic illustrates the process of generating the individual image frames of a deepfake video via face-swap. First, an autoencoder network $\mathcal{G}_{\text{actor}}$ is trained to accurately represent/encode and reconstruct/decode facial images of an actor (top) and a network $\mathcal{G}_{\text{target}}$ is trained to do the same for a target (bottom). Then, to execute the face-swap, the latent representation of each frame of the actor's performance are decoded using the target network's decoder (green), rather than the actor network's decoder (blue). Shown in the lower right are the resulting "deepfake" facial images (yellow box) seamlessly edited back onto the actor's background. Finally, the deepfake images are combined with the actor's original vocal performance to create the deepfake video.