Political Deepfake Videos Misinform the Public, But No More than Other Fake Media*

Soubhik Barari † Christopher Lucas ‡ Kevin Munger §

January 27, 2021

Abstract

We demonstrate that political misinformation in the form of videos synthesized by deep learning ("deepfakes") can convince the American public of scandals that never occurred at alarming rates – nearly 50% of a representative sample – but no more so than equivalent misinformation conveyed through existing news formats like textual headlines or audio recordings. Similarly, we confirm that motivated reasoning about the deepfake target's identity (e.g., partisanship or gender) plays a key role in facilitating persuasion, but, again, no more so than via existing news formats. In fact, when asked to discern real videos from deepfakes, partisan motivated reasoning explains a massive gap in viewers' detection accuracy, but only for real videos, not deepfakes. Our findings come from a nationally representative sample of 5,750 subjects' participation in two news feed experiments with exposure to a novel collection of realistic deepfakes created in collaboration with industry partners. Finally, a series of randomized interventions reveal that brief but specific informational treatments about deepfakes only sometimes attenuate deepfakes' effects and in relatively small scale. Above all else, broad literacy in politics and digital technology most strongly increases discernment between deepfakes and authentic videos of political elites.

*For excellent research assistance, we thank Jordan Duffin Wong. 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, Connor Huff, Yphtach Lelkes, and Steven Webster for helpful comments. We thank Hany Farid for sharing video clips used in this project. We are especially grateful to Sid Gandhi, Rashi Ranka, and the entire Deepfakeblue team for their collaboration on the production of videos used in this project. All replication data and code is publicly available here.

†Ph.D. Candidate, Harvard University; soubhikbarari.org, sbarari@g.harvard.edu

‡Assistant Professor, Washington University in St. Louis; christopherlucas.org, christopher.lucas@wustl.edu

§Assistant Professor, Pennsylvania State University; kevinmunger.com, kmm7999@psu.edu

1 Introduction

Studies of democratic politics have long emphasized the importance of a well-informed electorate for bolstering democratic accountability (Lippmann, 1922; Berelson, Lazarsfeld and McPhee, 1954; Downs, 1957; Snyder Jr and Strömberg, 2010; Herman and Chomsky, 2010). Information allows voters to accurately judge attributes of electoral candidates such as leadership, expertise, competence, character, and values in order to make principled decisions at the ballot-box (Popkin, 1991; Pierce, 1993; Alexander and Andersen, 1993; Alvarez, 1998; Strömberg, 2004; Caprara et al., 2006). Political misinformation, then, threatens the electorate's ability to credibly evaluate their public officials (Carpini and Keeter, 1996; Kuklinski et al., 2000; Hollyer, Rosendorff and Vreeland, 2019; Aral and Eckles, 2019; Jerit and Zhao, 2020).

Recent concerns about misinformation have centered on open-source deep learning technology capable of synthesizing realistic false videos of politicians making statements that they never said, colloquially termed *deepfakes*. Unlike previously available video manipulation tools, contemporary deepfake tools are free of cost, unlicensed, unregulated, and can be harnessed by hobbyists with only basic computer skills Government Accountability Office: Science and Analytics (2020). Figure 1 graphically summarises the two major technologies for the production of deepfakes, which, by many counts, are responsible for the production of the vast majority of political deepfakes at the time of writing (Lewis, 2018; Davis, 2020; Ajder et al., 2019). Since the advent of popular deepfake applications[1], political elites around the world have been adversarially targeted in scandalous deepfake videos. Notable examples include a 2018 deepfake of Gabon president Ali Bongo, which triggered a coup attempt, and a viral deepfake reporting on a sex scandal involving a Malaysian cabinet minister Harwell (2019).

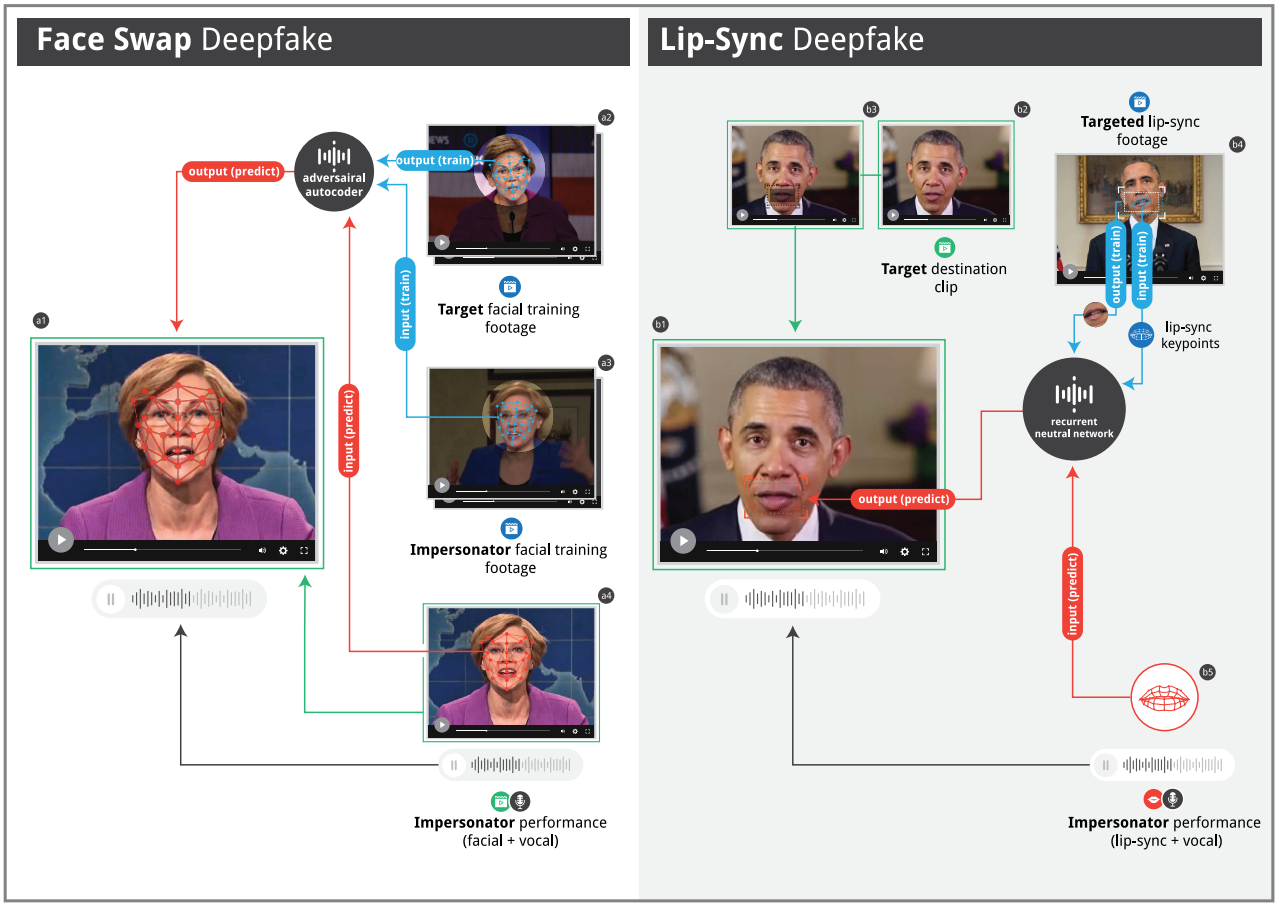
In the United States, deepfakes have been the subject of Congressional testimony from both sides of the political aisle. For example, Senator Marco Rubio (R-FL) raised the issue during a Senate Intelligence Committee hearing in May 2018 *Press Releases* (2018) and subsequently authored letters with Senator Mark Warner (D-VA) to Facebook, Twitter, YouTube, and other social media companies urging a stronger response to the potential threat from deepfakes Gazis and Becket (2019). At the same time, Senator Ben Sasse (R-NE) penned a widely-circulated opinion piece in *The Washington Post* expressing concern with the potential for deepfake technology to undermine the democratic process Sasse (2018), then later introduced legislation aimed at curtailing deepfakes Brown (2019). And in the U.S. House, Rep. Yvette Clarke (D-NY) proposed similar legislation Galston (2020). Even former President Barack

[1] According to many reports (Lewis, 2018; Davis, 2020; Ajder et al., 2019), the earliest deep learning face-swap tool to receive popular press was `Face2Face` in 2016; see also Suwajanakorn, Seitz and Kemelmacher-Shlizerman (2017) in 2017, `FakeApp` in 2018, `Faceswap` and `DeepFaceLab` in 2019.

Obama has expressed concern over the looming threat posed by deepfakes Lum (2019).

In addition to elected representatives, news outlets Harwell (2019); Parkin (2019); Frum (2020); Hwang and Watts (2020); Schick (2020); Toews (2020) and civil society research (Lewis, 2018; Davis, 2020; Ajder et al., 2019; Bateman, 2020) alike continue to emphasize the potential harm that deepfakes may cause to democracy, and legislation exists in more than a dozen states to regulate the production and dissemination of deepfake videos Prochaska, Grass and West (2020). Meanwhile, philosophers warn about the possibility that deepfake technology removes our ability to verify testimony, undercutting the credibility of both honest and dishonest claims (Rini, 2020).

Figure 1: **How Deepfake Videos are Produced**



Notes: Shown are two major methods of producing deepfakes. The left illustrates the production of a *face-swap deepfake* which requires: a full clip featuring the impersonator's performance including the audio (**black**) and the background context for the clip (**green**) where the facial features are swapped (**red**) via a trained (**blue**) deep learning model called an autoencoder. The right illustrates a *lip-sync deepfake* which requires a destination clip of the target (**green**) and a vocal impersonator's performance including their audio (**black**) and lip sync keypoints (**red**); these keypoints are transferred into a matching synthetic lip-sync video of the target via a deep convolutional neural network model trained on the target (**blue**).

Evaluating whether or not these concerns are warranted first requires answering a fun-

damental research question, which is the primary contribution of this article: can deepfakes more powerfully persuade the public of non-existent scandals for real public officials than comparable media formats such as textual headlines or audio recordings? Here, we decompose persuasion into two possible effects: *deception* into believing in an event that never occurred and increased negative *affect* towards the target elite. Research on audiovisual media effects on these two outcomes spans the fields of law, journalism, psychology, neuroscience, and political science, producing somewhat conflicting answers, which we briefly summarize here.

One body of research suggests that audiovisual information is the prima facie format for communication: that, compared to textual information, it enhances short- and long-term recall (Graber, 1990; Witten and Knudsen, 2005; Stenberg, 2006; Prior, 2014), particularly for emotionally charged events (Christianson and Loftus, 1987; Kassin and Garfield, 1991), and is more persuasive in a variety of contexts including courtroom testimony (Kassin and Garfield, 1991), presidential election campaigns (Grabe and Bucy, 2009), support for medical aid (Yadav et al., 2011), and belief in climate change (Goldberg et al., 2019). One reason for its relative persuasive power may be humans' perceived "reliability" of visual signals relative to audio-only information (Witten and Knudsen, 2005). Because of this cognitive heuristic, false images of fabricated political events are more likely to deceive receivers than equivalent verbal misinformation (Frenda et al., 2013). Additionally, audiovisual media may facilitate stronger persuasion than textual media because of its enhanced capabilities of affective appeal. In the political sphere, audiovisual stimuli such as campaign advertisements, televised partisan media, and 'soft' political news elicit such emotional responses through music, uncivil language, humor, facial competence, and visual symbols (Ansolabehere and Iyengar, 1997; Brader, 2006; Atkinson et al., 2009; Boukes et al., 2015; Mutz, 2016; Baym and Holbert, 2020). This literature predicts that a political scandal presented in a deepfake video is more likely to deceive and elicit a negative response against its target than if presented as a news headline or audio recording.

A contrary body of scholarship predicts that a deepfake scandal may be minimally, or not at all, more persuasive than an identical scandal presented in any other communication medium. Although sensational 'blood and guts' video evidence may move jurors' attitudes, many other psychological studies find that information from printed media is better retained and applied in the typical courtroom setting (Furnham and Gunter, 1989; Fishfader et al., 1996; Pezdek, Avila-Mora and Sperry, 2010). Ecologically, the highest-profile misinformation environment—fake news in the 2016 election—was entirely text- and image-based (Vosoughi, Roy and Aral, 2018). Video- and audio-based misinformation is effective in creating alternative media ecosystems like YouTube's (Lewis, 2018; Munger and Phillips, 2020), but has not had impact at the scale of the 2016 phenomenon. Moreover, many deepfakes are produced by

essentially swapping the faces of political elites with their impersonators' in comedic sketches, including the example depicted in Figure 1 with Senator Elizabeth Warren inserted into Saturday Night Live actor Kate McKinnon's performance. Given this methodological quirk, it is possible that such deepfakes are interpreted no differently as digitally enhanced satirical impersonations; nevertheless, this type of 'soft news' is found to have affective and persuasive appeal for less poltiically engaged viewers on some, but not all issues (Gray, Jones and Thompson, 2009; Boukes et al., 2015; Esralew and Young, 2012). Lastly, a growing collection of studies find that advertising and news media exert small persuasive effects on attitudes, with little heterogeneity by medium or any other qualities (Furnham and Gunter, 1989; Lau, Sigelman and Rovner, 2007; Bennett and Iyengar, 2008; Coppock, Hill and Vavreck, 2020). Inconsistent with popular expectations, this body of research predicts that the marginal deceptive and affective effects of a deepfaked political scandal are likely to be small or null relative to other closely related media.

In addition to heterogeneity based on medium, our research investigates heterogeneity by viewer traits. That is, we ask: what immutable or intervenable characteristics predict heterogeneities in a viewer's susceptibility to deepfakes' misinformation? In Table 1, we register a series of hypothesized "at-risk" subgroups that may be differentially susceptible to deepfakes.[2] Each of the stable categories of media consumers, though not mutually exclusive, have been shown to process political information in ways that make them less likely to hold true beliefs. Our first category draws on the observation that older adults, 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. Next, we believe that directional motivated reasoning, or the selective acceptance of information based on consistency with previous beliefs, may powerfully shape how voters respond to deepfakes in at least two ways. A large literature documents how *partisan motivated reasoning* directs voters' attitudes about events, issues, and candidates even in the light of information that contradicts prior expectations (Kahan, 2012; Druckman and McGrath, 2019; Leeper and Slothuus, 2014; Enders and Smallpage, 2019). Moreover, 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). A recent survey finds that, next to partisanship, holding ambivalent sexist views[3] most strongly predicted support for Donald Trump in the 2016 election (Schaffner, MacWilliams and Nteta, 2018). Another set of

[2]On all outcomes of interest, pre-registered analyses in Appendix F estimate the effects of additional control characteristics such as internet usage, device platform (i.e. mobile vs. desktop), gender, and education.

[3]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)

Table 1: **Subgroups with Hypothesized Susceptibility to Deepfake Deception**

Susceptible Subgroup	Mechanism(s) of Deception
Older adults ($\geq$ 65 y.o.)	Inability to evaluate accuracy of digital information (Guess, Nagler and Tucker, 2019; B
Partisans (w/out-partisan deepfake target)	Motivated reasoning about target (Kahan, 2012; Druckman and McGrath, 2019; Leeper
Sexists (w/female deepfake target)	Motivated reasoning about target (Glick and Fiske, 1996; Schaffner, MacWilliams and N
Low cognitive reflection	Overreliance on intuition when making judgments (Pennycook and Rand, 2019; Pennyc
Low political knowledge	(1) Inability to evaluate plausibility of political events (2) Inability to recognize real facial features of target (Brenton et al., 2005; Mori, MacD
Low digital literacy	(1) Inability to evaluate accuracy of digital information (2) Limited/no awareness of deepfake technology (Guess et al., 2020; Munger et al., 202
Low accuracy salience	Limited/no attention to factual accuracy of media (Pennycook et al., 2020, 2019)
Uninformed about deepfakes	Limited/no awareness of deepfake technology

subgroups may be especially susceptible to deepfakes due to constraints on cognitive resources or knowledge. Performance in cognitive reflection tasks measures reliance on "gut" intuition which may preclude careful examination of video evidence. Similarly, those with little political knowledge may have little prior exposure to the targetted political figure, rendering them unable to discern "uncanny" deepfake artifacts that resemble, but do not perfectly replicate their intended facial features (Mori, MacDorman and Kageki, 2012; Brenton et al., 2005). Finally, the last two categories describe traits that we can intervene on via information provision and priming, each of which we expect to reduce deepfakes' deceptive potential.

The few existing studies of "deepfake effects" support the latter hypothesis of minimal effects. However, they either recycle well-recognized deepfakes which underestimates deception rates (Vaccari and Chadwick, 2020; Wittenberg et al., 2020) or lack sufficient power to benchmark against the equivalent text-only and audio-only misinformation or comparable reference stimuli such as ads or digitally un-altered impersonations of elites (Dobber et al., 2020). To provide an enriched answer to the inquiries posed in this article, we conduct two experiments with a battery of customized deepfakes and a sufficient sample size to facilitate the estimation of interaction effects across a number of comparable media.

2 Materials and Methods

We employ two survey experiments fielded to a nationally representative sample of 5,750 respondents on the Lucid survey research platform.[4] The first experiment (exposure) shows respondents a news feed – similar to a feed found on Facebook or Twitter – with posts about candidates in the 2020 Democratic presidential primary, in which a single deepfake video may or may not be embedded. The second experiment (detection) asks the same respondents to scroll through a feed of eight news videos and discern deepfakes from unmanipulated news

[4]At the time of fielding, Aronow et al. (2020) note systematic trends in inattentive survey respondents on Lucid. We describe the battery of attention checks we employ to maintain a high-quality sample in Appendix E.

clips.

This dual experimental design affords internal and external validity by allowing us to observe deceptive effects (and affective effects, but only in the first experiment) of deepfakes in two settings that differ in at least four important ways. First, the settings vary in their context. The former surrounds the deepfake stimulus with authentic campaign news across textual, audio, and video formats replicating a natural news-browsing experience, while the latter simulates an adversarial experience with an explicit normative goal. As such, the first setting encourages users to engage with news with their baseline motivational state while the second setting encourages users to engage with a goal motivation. Second, the two settings allow for different behavioral measures. The exposure setting isolates a measurement of deception and affect from a single clip while the detection setting can produce a within-individual deception rate or 'grade' across eight clips. Third, the settings facilitate different comparison groups. In the exposure setting, we can compare the attitudinal effects of a single deepfake to its related textual, audio, and un-deepfaked video counterparts, whereas in the detection setting, we can observe deception rates across different aggregate news feeds. Fourth, and most important for external validity, the settings differ in the type, quality, and targets of the deepfake video clips used. The exposure experiment's deepfakes are high quality face-swap videos of a single elite depicted in a number of slightly different scandals that we have produced to be maximally deceptive (details in Appendix A), while the detection experiment's deepfakes span a representative mixture of styles, setting, qualities, and targeted elites from the existing population of political deepfakes accessible on the Internet (details in Appendix B).

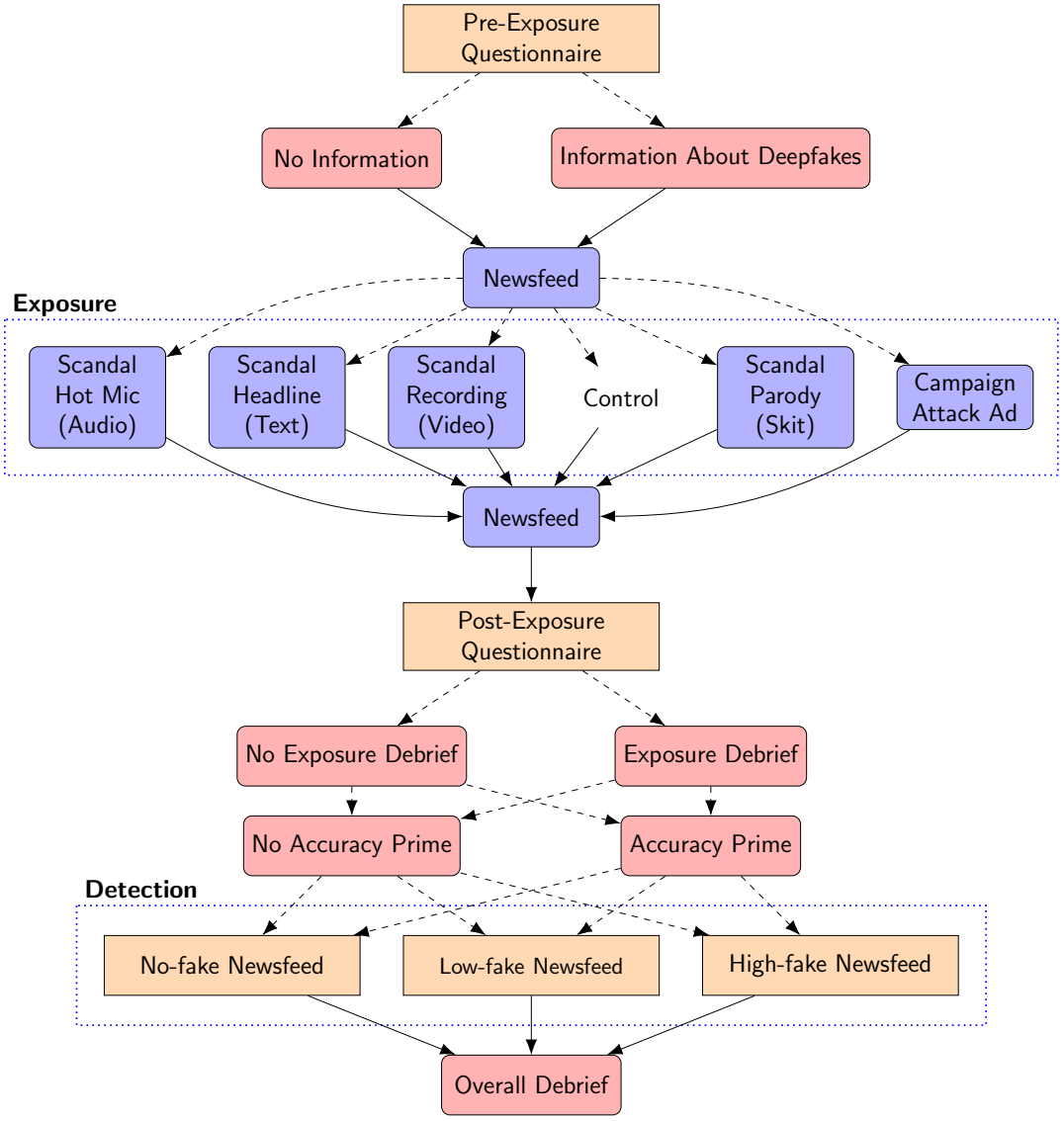
To adjust for observable demographic skews in this sample, all analyses are replicated using post-stratification weights estimated from the U.S. Census in Appendix F. Details of this post-stratification and other characteristics of the sample are given in Appendix E. Figure 2 provides a graphical summary of our experiments.

2.1 Exposure experiment

In the first experiment, we implement a 2 x 6 factorial design pairing a randomized informational message about deepfakes with randomization into one of six conditions – a deepfake `video`, or alternatively `audio`, `text`, or `skit` presentation of a political scandal involving a 2020 Democratic primary candidate Elizabeth Warren, a campaign attack `ad` against Warren, or a `control` condition of no clip at all – after which we measure several outcomes.

To mimic a natural environment for media consumption, we surround the experimentally manipulated media exposure with five media clips, two before and three after. These reports are all real coverage of different Democratic primary candidates, presented either in audio, textual, or video form. The order and content of these media are fixed, and primarily serve

Figure 2: **Summary of Experimental Flow**

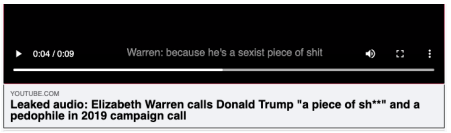
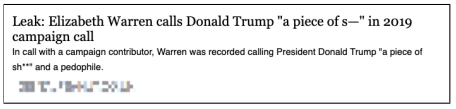


Notes: In the **audio**, **text**, **skit**, and **video** exposure cells, respondents are further randomized to one of the 5 clippings in Table A3. Subjects who do not receive an exposure debrief prior to the detection task receive it immediately after in overall debrief.

to 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 (`video`, `audio`, `text`, `skit`, `ad`, `control`) and their exact differences from each other are shown in Table 2, where `video` is the group assigned to the deepfake.

Participants in the `video`, `audio`, and `skit` conditions are randomly exposed to one of five different scandal events to reduce the possibility that our results are being driven by a single story. Each scandal is entirely fictitious, and the media associated with it was created in collaboration with a professional actor and tech industry partners. Specifically, the `audio` condition consists of the audio recording of the actor making a scandalous statement. The

Table 2: **Experimental Conditions in Exposure Experiment**

	Condition	Description of Variation	Example Clip
Scandal Clips (Script Held Constant)	**Video** ($n = 872$)	Face-swap performed on video in skit condition; title and video edited to resemble leaked video footage.	
	Audio ($n = 954$)	Visuals stripped from video condition; title edited to resemble leaked hot mic.	
	Text ($n = 950$)	Visuals and sound stripped from video video condition; title describes scandal as a leak; subtitle describes event captured on video.	
	Skit ($n = 956$)	Filmed impersonator portraying a campaign scandal event.	
Reference Stimuli	**Ad** ($n = 935$)	Campaign attack advertisement describing real scandal event.	
	Control ($n = 916$)	No stimulus presented.	N/A

`skit` condition includes the same audio recording, but with the accompanying video of the actor speaking in a realistic "hot-mic" setting. Finally, the `video` condition employs a deepfake constructed from the footage used in the `skit` condition.

The five unique, fictitious scandals are meant to simulate five possible defamation strategies by 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 conduct exploratory analyses which show minimal differences across conditions (Appendix G). Details on the creation of these stimuli are provided in Appendix A and each of the five scripts are provided in Table A3.

Importantly, participants in the `skit` condition are exposed to the *original* videos used in the creation of the deepfake video, *prior* to the modifications made by the neural network algorithm. That is, this condition displays the unaltered video of the paid actress hired to impersonate Elizabeth Warren which is clearly framed as a skit: the title of corresponding deepfake in the `video` condition is shown, but "Leak" is replaced with "Spot-On Impersonation". This condition represents a conservative test of the hypothesis that deepfake videos uniquely deceive relative to even their un-deepfaked seed footage: it is identical to the deepfake condition, but without the computer-assisted falsification of the real politician from the actress performing an impersonation. If we observe a difference between the `audio` and `text` conditions when compared to the deepfake `video` condition, but not between the `video` and the `skit` condition, it suggests that the mechanism is the video depiction of the scandal, whether or not it is true or clearly fictionalized.

Finally, in the `ad` condition, subjects are exposed to a real negative campaign ad titled, "Tell Senator Warren: No Faux Casino, Pocahontas!", which criticizes 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 disputed claims of Cherokee heritage. Although the ad frames Warren as politically insincere, similar to script (e) and primes the viewer of her Cherokee heritage controversy, similar to script (c), it stylistically and informationally differs in many other ways, and thus is not an exact ad counterfactual of our deepfake. Nevertheless, the ad serves as a benchmark comparison for a deepfake's affective effect, since it is an actual campaign stimulus used in the primary election to activate negative emotions towards Warren.

2.2 Detection experiment

After completing the battery of questions in which we measure our primary outcomes of interest and ask another attention check question, the subjects begin the second 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) are debriefed about whether or not they were exposed to a deepfake in the first experiment. The other half are debriefed after this final task. This randomization allows us to test for the effect of the debrief itself. Additionally, half of all respondents are provided an accuracy prime – an intervention designed to increase the salience of information accuracy

(Pennycook and Rand, 2019).

Here, we employ videos created by (Agarwal et al., 2019) and a mixture of other publicly available deepfake videos of both lip-sync and face-swap varieties. To the extent that respondents have previously viewed these videos, we should expect detection performance to be biased upwards, although no respondent indicated as such in open feedback. Subjects were randomly assigned to one of three environmental conditions: the percentage of deepfakes in their video feed was either 75% (high-fake), 25% (low-fake) or 0% (no-fake). Appendix B displays screenshots and descriptions of each of these videos.

2.3 Ethical considerations

Creating deepfakes raises important ethical concerns, which we aimed to address at every stage of our research design. First, given the risk of deepfakes disrupting elections, understanding their effects is of the utmost importance: this research has the potential to improve the resilience of democratic politics to this technological threat by better informing policy and consumer behavior. Second, we created deepfakes of a candidate who was not currently running for office to insure that our experiment could not plausibly influence the outcome of an election. Third, we designed "active debriefs" that required subjects to affirm in writing whether they were exposed to false media. Finally, deepfakes are increasingly part of the standard media environment, so our study only exposes subjects to things they should be prepared to encounter online. We discuss these points in more detail in Appendix D.

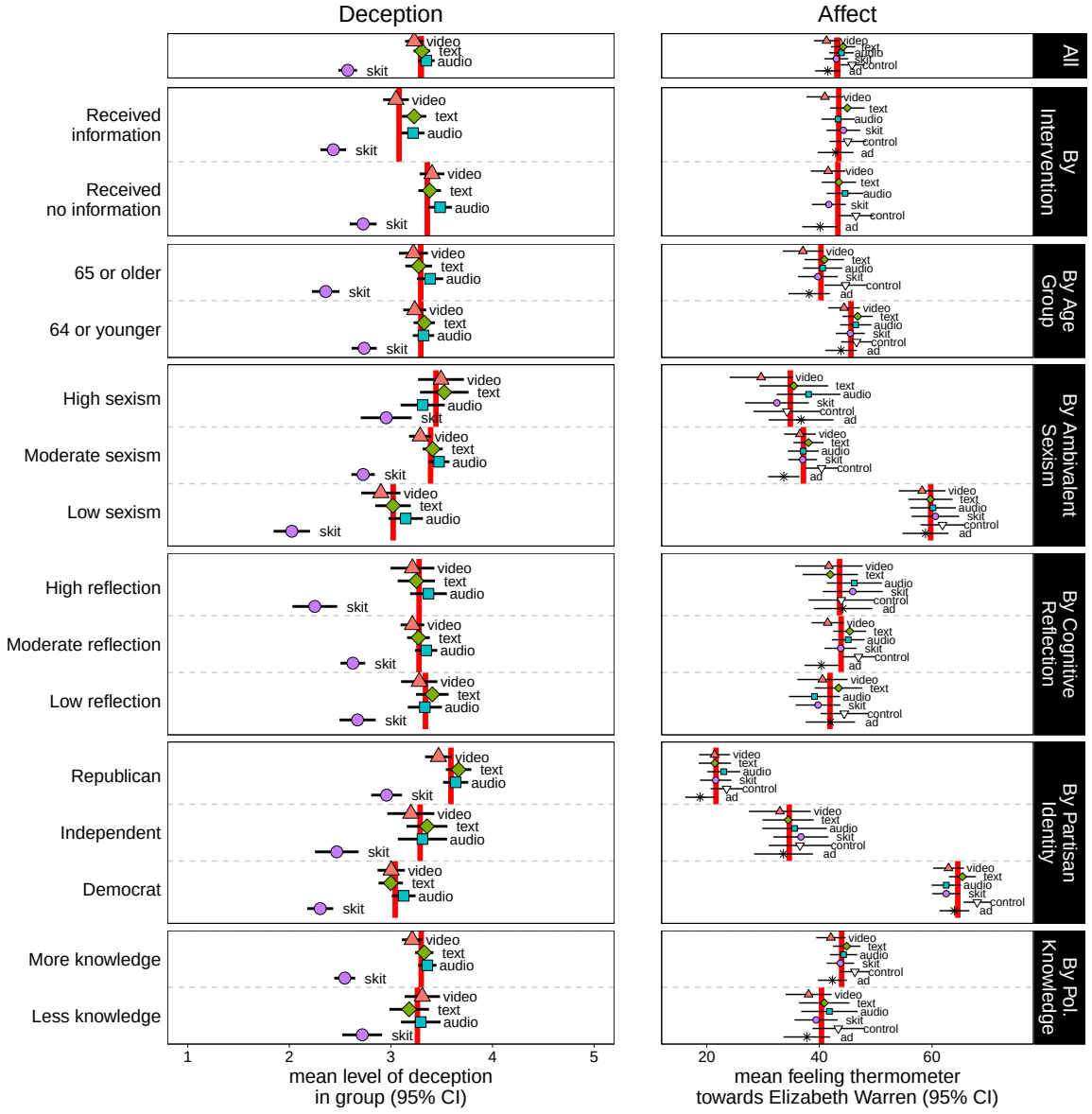
3 Results

3.1 Exposure experiment results

Our deepfake videos – with an average deception level of 3.22 out of 5 or 47% of respondents who are more confident the clip is real rather than not – were statistically no better at deceiving subjects than the same information presented in audio (3.34 or 48%) or text (3.30 or 43%), shown in the top-left cell in Figure 3.[5] The inclusion of the skit of our Warren impersonator doing her best performance without deepfaking provides a baseline to compare these numbers against (2.57 or 30%). There are clear and significant differences (denoted henceforth as Δ) in deception levels between the video and the skit conditions as measured by a two-sample t-test ($\Delta = -4.53, t = 8.47, p < 0.01$), though none with the text ($\Delta = -0.07, t = -1.21, p = 0.22$) and likely none with the audio ($\Delta = -4.53, t = -1.92, p = 0.05$). In absolute terms, just under half of subjects were deceived by the best-performing stimuli. Table F5 and Table F6

[5]Later manipulation checks that pair belief in clip authenticity with confidence in the depicted event shows that respondents who believe the video is fake also believe the event did not occur (Figure G29 and Figure G30 in Appendix G). This holds but to a much lesser degree for highly implausible or controversial clips such as Trump publically resigning.

Figure 3: **Marginal Means in Exposure Experiment Outcomes**



Notes: The outcome on the left (deception) is the inverted categorical response to the question *"To what extent do you think that the clipping of [scandal description] is fake or doctored?"* from respondent subgroups exposed to a scandal-presenting clip (n=3,732). The outcome on the right (affect) is the response from 1-100 to *"How would you rate your feeling towards Elizabeth Warren?"* from all respondent subgroups (n=5,750). The discrete categories for ambivalent sexism, cognitive reflection, and political knowledge are constructed as equal-sized percentiles from their observed integer values. The subgroup-specific vertical **red** lines in both columns indicate the subgroup averages for those assigned to the `video`, `text`, and `audio` conditions.

show that these differences are robust to a variety of model-based adjustments.

Relative to no exposure, videos do increase negative affect towards Elizabeth Warren as measured by the 0-100 feeling thermometer ($\Delta = -4.53, t = -2.79, p < 0.01$). However, there are demonstrably null effects of the deepfake video on affect relative to text ($\Delta = -2.94, t = -1.84, p = 0.06$) and audio ($\Delta = -2.64, t = -2.64, p = 0.09$), as seen in the top-right cell in

Figure 3. Deepfake videos are also at least as affectively triggering as negative advertisement videos, a decades-old technology, of the same target ($\Delta = -0.23, t = -0.14, p = 0.89$). Table F8 robustly estimates this same null with models.

Receiving a brief informational warning about the existence of deepfakes reduces the credibility of all fake news clippings ($\Delta = -0.28, t = -6.84, p < 0.01$) with the largest drop for video recipients ($\Delta = -0.35, t = -3.91, p < 0.01$), and there are no effects on affect towards the target (second row in Figure 3). This offers a degree of optimism about a simple method to quickly inoculate citizens against the deceptive effects of deepfakes, although the effects are much smaller in magnitude compared to swapping the deepfake with an identical skit (model-based results in Table F7).

We next investigate whether our topline null results mask any deepfake effect heterogeneity for subgroups specified in Table 1 (rows 3-7 in Figure 3): in short, we find none. First, we see that although the elderly are more likely to be affectively triggered by the Warren fake media clippings relative to those below 65 ($\Delta = -5.42, t = -5.73, p < 0.01$), there is no detectable difference in deception nor across media (row 3). Second, we find moving from the lowest to the highest level of ambivalent sexism produces increases deception across the board ($\Delta = 0.55, t = -7.98, p < 0.01$), and decreases favorability ($\Delta = -25.96, t = 17.78, p < 0.01$), but nothing statistically differentiates videos from other media (row 4). Third, variations in cognitive reflection do not predict more or less deception or affective response between or within any of the media conditions (row 5). Fourth, there are detectable differences in deception between Democrats and Republicans ($\Delta = 0.61, t = -13.77, p < 0.01$) and enormously large differences in target affect ($\Delta = -42.58, t = 53.53, p < 0.01$); again, there are no further distinctions between the three types of fake media (row 6). Finally, somewhat surprisingly, the cohort with higher political knowledge shows no more skepticism toward any of the three fake media ($\Delta = 0.00, t = 0.11, p = 0.92$), and slightly more favorability towards Warren ($\Delta = 3.85, t = -3.65, p < 0.01$). Parametric model adjustments in the Appendix support these findings (Tables F12–F19).

Altogether, the exposure experiment furnishes little evidence that deepfake videos have a unique ability to fool voters or to shift their perceptions of politicians. In fact, the `audio` condition had the largest average deceptive effect, though the difference is statistically insignificant relative to `video` in all but two subgroups; some models that estimate adjusted marginal coefficients of deception for each group find this difference to be significant.[6] Appendix Figure G24 and Figure G25 highlight that there is slight heterogeneity by the particular scandal presented (in-party civility appears to be a less credible scandal), but this is neither robustly significant nor large. On affective responses, `video` seems to induce a small, but statistically

[6] In Appendix F, see Table F5, Table F8, and Table F17.

insignificantly greater affective response across all respondents and subgroups in Figure 3.

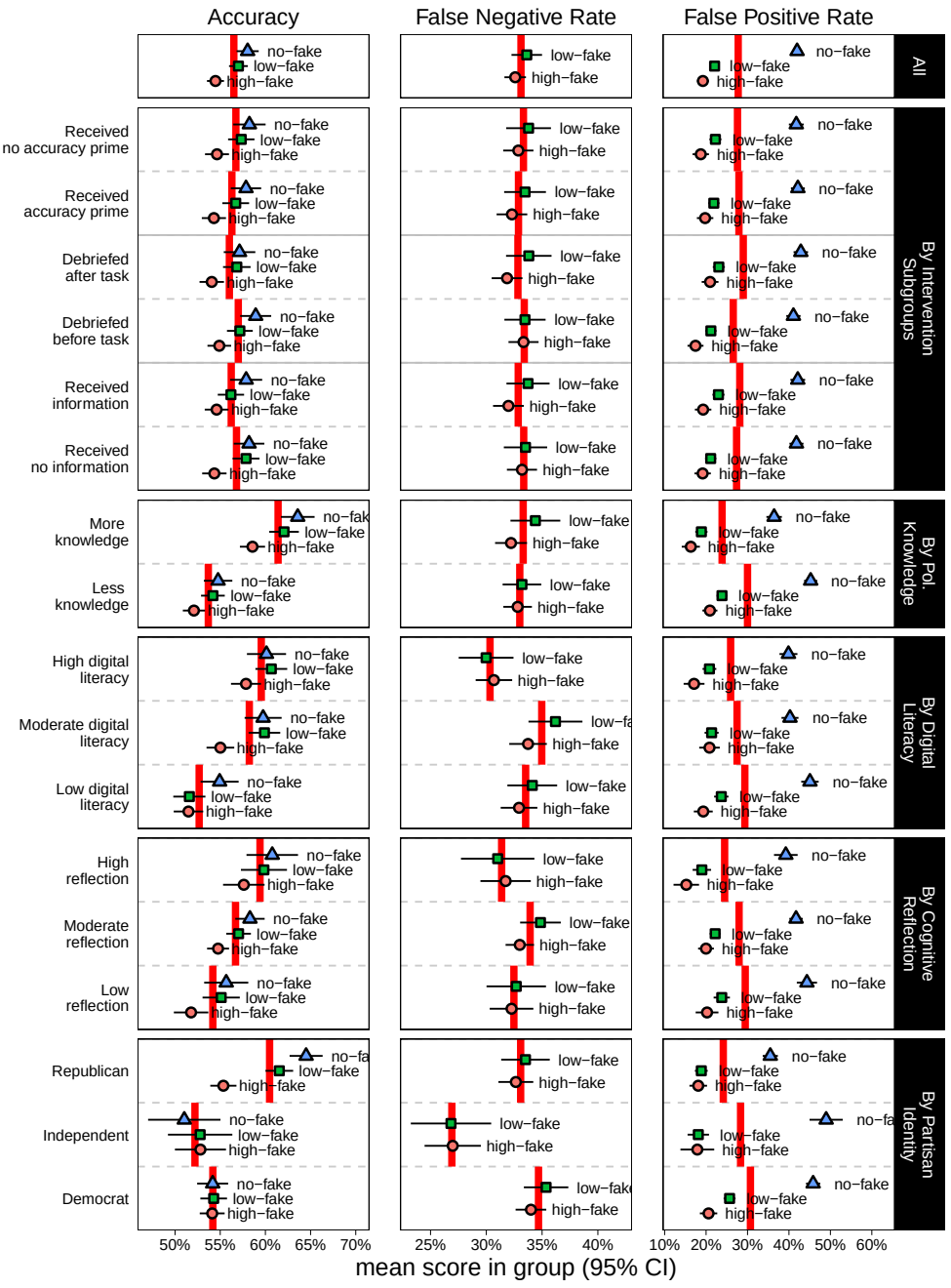
3.2 Detection experiment results

Figure 4 summarises the overall and subgroup accuracy, false positive and false negative rates in the detection experiment. Unlike in the exposure experiment, the cognitively reflective and politically knowledgeable are deceived far less than their peers. Out of all of our subgroups, the politically knowledgeable achieve the best detection accuracy, on average discerning 63% or 5 out of the 8 videos in their news feed. However, those with high digital literacy experience the largest *marginal* gains in accuracy relative to the least digitally literate – multiple regression results adjusting for all other subgroup variables (Figure 5) finds that a unit increase in literacy produces a roughly 21% increase in accuracy across environments. Remarkably, though fake news is disproportionately shared by Republicans (Guess, Nagler and Tucker, 2019), Republicans out-perform Democrats in discernment; this appears to be driven by Democrats' increased propensity for false positives, particularly in news feeds with no deepfakes at all. Nevertheless, Figure 5 shows that the above two literacy traits largely outweigh the effects of partisanship on performance.

There are no statistically significant differences between fake and real clips or clips with source logos and clips without (Appendix Figure G27). However, it is worth summarizing the detection rates of several noteworthy clips. The least correctly identified clip (21% correct) is a short deepfake where Hillary Clinton appears to make a poignant but uncontroversial point about her opponent's tax plan in a presidential debate while the most correctly identified clip (89% correct) is a deepfake where President Donald Trump publicly announces his resignation before the election. Digital literacy, political knowledge, and cognitive reflection bolster correct detections roughly evenly for all clips. However, the striking differential in Democrats' and Republicans' performances is explained by Figure G28: partisans fare much worse in correctly identifying real, but not deepfaked, video clips portraying their own party's elites in a scandal. 58% of Republicans believed that real leaked footage of Obama caught insinuating a post-election deal with the Russian president was authentic compared to 22% of Democrats, a highly significant differential according to a simple Chi-squared test ($\chi^2 = 333.34, p < 0.01$). The numbers are flipped for the clip of Donald Trump's public misnaming of Apple CEO Tim Cook which was correctly identified by 87% of Democrats, but only 51% of Republicans ($\chi^2 = 75.15, p < 0.01$). Perhaps most striking is that for an authentic clip from a presidential address of Trump urging Americans to take cautions around the COVID-19 pandemic, the finding holds in the opposite direction: it is a positive portrayal[7], at least for Democrats who by

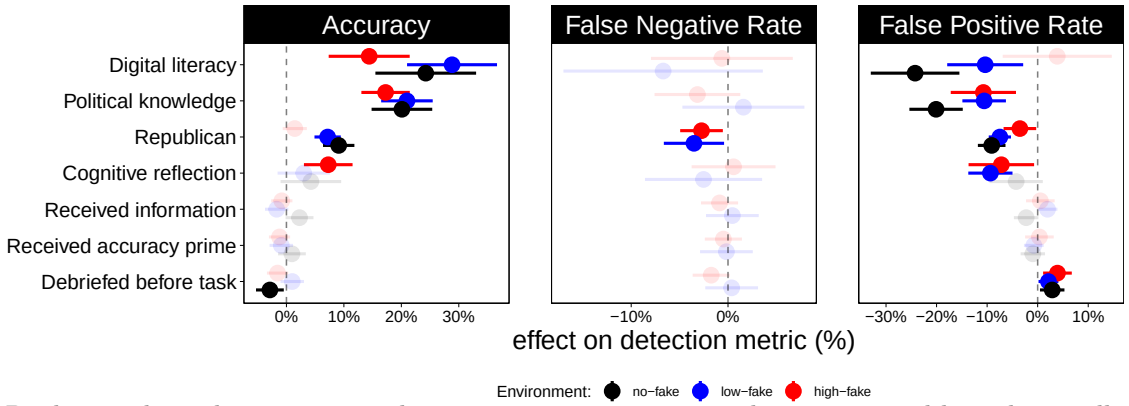
[7]Positive portrayal, here, means depiction of positive valence traits or characteristics that, all else equal, Democratic voters would unanimously prefer more of rather than less of (Bartels, 2002).

Figure 4: **Marginal Means in Detection Experiment Outcomes**



Notes: Shown are three different measures for n=5,497 (99%) of respondents who provide a response to at least one video in the detection experiment task; coefficient estimates are given in Appendix F and are robust to the choice of missing-ness threshold. Accuracy is measured as the % of all videos correctly classified as either fake or real. False negative rate is measured as the % of deepfakes in the task incorrectly classified as real (as such, this quantity is degenerate in the no-fake condition). False positive rate is measured as the % of real videos in the task incorrectly classified as false. The discrete categories for ambivalent sexism, digital literacy, cognitive reflection, and political knowledge are constructed as equal-sized percentiles from their observed integer values. The subgroup-specific vertical **red** lines in all three columns indicate the subgroup averages across the displayed conditions.

Figure 5: **Predictors of Detection Experiment Performance by Environment**



Notes: Predictors shown here are respondent traits or interventions with pre-registered hypotheses, all commonly re-scaled to the [0,1] range. Results are for 5497 (99%) of respondents who provide a response to at least one video in the detection experiment. Appendix Figure F23 shows that the choice of non-response threshold does not change the substantive interpretation of detection experiment results.

and large hold similarly cautionary attitudes towards COVID-19 (Clinton et al., 2020), yet only 60% of Democratic viewers flagged it as authentic whereas fully 82% of Republicans believed it to be real ($\chi^2 = 169.96, p < 0.01$). Partisan motivated reasoning about the video target's capacity for scandal does explain video misdetection, but this operates through discrediting real videos than rather believing deepfake videos. This particularly striking given that, at the time, these real events and their clips made headlines with left- and right-leaning news outlets and, in some cases, became viral on social media, which we would expect to universally boost their recognizability.

Furthermore, results show that brief textual interventions – information provision in the first experiment, debriefing about the first experiment, and accuracy priming – have no detectable effects on performance in the detection experiment. In fact, receiving a debrief actually marginally increases the false positive rate by a small amount (Figure 5). None of these interventions significantly improves the likelihood that the subject correctly identifies a deepfake as such.

Finally, examining the heterogeneities by environment we see that the biggest gains in the detection experiment come from improved performance in the no-fake condition by high-political knowledge and high-digital literacy subjects.

4 Conclusion

The aphorism "seeing is believing" has largely survived the digital age; indeed, the sentiment has been expanded in the affirmative, to "pics or it didn't happen." The incipient ubiquity of deepfakes challenges this previously indisputable belief that we may trust that which we

see. And while Hollywood blockbusters and doctored photos alike problematize this epistemic assumption, deepfake technology provides millions of people the opportunity to produce convincing video evidence of events that did not take place. As such, popular concerns over this technology's potential to scalably damage the reputations of political elites and the credibility of the broader political news environment are widespread and extreme.

With these expectations in mind, our results are somewhat encouraging. We have demonstrated that deepfakes, even when professionally produced and adversarially designed to defame a prominent politician, are not uniquely powerful at deception or affective manipulation: they are no more effective than the same misinformation presented as text or audio or the same target attacked via a campaign ad or mocked in a satirical skit. The most deceptive video in our detection task, tricking nearly eight out of every ten respondents, was the seemingly least controversial, while the least credible video – of a historically controversial pre-election resignation – fooled roughly one out of every ten subjects. Moreover, partisan motivated reasoning appears to facilitate misinformation, but for authentic video news more than for deepfakes. Altogether, this hints at a hopeful conclusion that should be corroborated in future research: that as the subjective level of controversy in a deepfake-depicted event increases, the empirical credibility of the event decreases, diminishing its potential to cause political scandal to begin with. If true, policy-makers should devote more time and resources to bolster the credibility of real news videos and curb the development and spread of deepfake videos that perpetrate psychological or social damages against their targets. A recent count of deepfakes on the Internet finds that most are non-consensual pornographic clips of women, suggesting that perhaps the greater, more novel harm of deepfakes is the harassment of its targets, not misinformation of its viewers Harris (2018).

Finally, also relevant for a future policy agenda, we show that several cognitive characteristics are essential components of how citizens process political information. In particular, the respondents with the highest levels of general knowledge about politics, literacy in digital technology, and propensity for cognitive reflection performed the best in the detection experiment. These skills will only grow in importance as digital video technology approaches the limit of realism. While we encourage technological solutions to constrain the spread of manipulated video, there will never be a substitute for an informed, digitally literate, and reflective public for the practice of democracy.

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.

Abram, Cleo. 2020. "The most urgent threat of deepfakes isn't politics. It's porn.".
URL: *https://www.vox.com/2020/6/8/21284005/urgent-threat-deepfakes-politics-porn-kristen-bell*

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.

Ajder, Henry, Giorgio Patrini, Francesco Cavalli and Laurence Cullen. 2019. "The State of Deepfakes: Landscape, Threats, and Impact." *Policy Brief* .
URL: *http://regmedia.co.uk/2019/10/08/deepfake_report.pdf*

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

Alvarez, R Michael. 1998. *Information and Elections*. University of Michigan Press.

Ansolabehere, Stephen and Shanto Iyengar. 1997. *Going Negative: How Political Advertisements Shrink and Polarize the Electorate*. The Free Press.

Aral, Sinan and Dean Eckles. 2019. "Protecting elections from social media manipulation." *Science* 365(6456):858–861.

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

Atkinson, Matthew D, Ryan D Enos, Seth J Hill et al. 2009. "Candidate faces and election outcomes: Is the face-vote correlation caused by candidate selection." *Quarterly Journal of Political Science* 4(3):229–249.

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

Bartels, Larry M. 2002. "The impact of candidate traits in American presidential elections." *Leaders' Personalities and the Outcomes of Democratic Elections* pp. 44–69.

Bateman, Jon. 2020. "Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios.".

Baym, Geoffrey and R. Lance Holbert. 2020. "Beyond Infotainment." *The Oxford Handbook of Electoral Persuasion* p. 455.

Bennett, W Lance and Shanto Iyengar. 2008. "A new era of minimal effects? The changing foundations of political communication." *Journal of Communication* 58(4):707–731.

Berelson, Bernard R, Paul F Lazarsfeld and William N McPhee. 1954. *Voting: A Study of Opinion Formation in a Presidential Campaign*. University of Chicago Press.

Blum, Jeremy. 2020. "Trump Rant About 'Anarchist' Protesters Wielding Deadly 'Cans Of Soup' Goes Viral.".
URL: *https://www.huffpost.com/entry/trump-deadly-cans-of-soup˙n˙5f4fbcc6c5b69eb5c0379f01*

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.

Brenton, Harry, Marco Gillies, Daniel Ballin and David Chatting. 2005. The Uncanny Valley: Does it Exist? In *Proceedings of the Conference of Human Computer Interaction*.

Brown, Nina Iacono. 2019. "Congress Wants to Solve Deepfakes by 2020. That Should Worry Us.".
URL: *https://slate.com/technology/2019/07/congress-deepfake-regulation-230-2020.html*

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.

Carpini, Michael X Delli and Scott Keeter. 1996. *What Americans Know about Politics and Why It Matters*. Yale University Press.

Christianson, Sven-åke and Elizabeth F Loftus. 1987. "Memory for traumatic events." *Applied Cognitive Psychology* 1(4):225–239.

Clinton, J., J. Cohen, J. Lapinski and M. Trussler. 2020. "Partisan pandemic: How partisanship and public health concerns affect individuals' social mobility during COVID-19." *Science Advances* .

Coppock, Alexander, Seth J Hill and Lynn Vavreck. 2020. "The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments." *Science Advances* 6(36).

Davis, Raina. 2020. "Technology Factsheet: Deepfakes." *Policy Brief* .
URL: *https://www.belfercenter.org/publication/technology-factsheet-deepfakes*

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* .

Downs, Anthony. 1957. *An Economic Theory of Democracy.* New York.

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

Fishfader, Vicki L, Gary N Howells, Roger C Katz and Pamela S Teresi. 1996. "Evidential and extralegal factors in juror decisions: Presentation mode, retention, and level of emotionality." *Law and Human Behavior* 20(5):565–572.

Frenda, Steven J, Eric D Knowles, William Saletan and Elizabeth F Loftus. 2013. "False memories of fabricated political events." *Journal of Experimental Social Psychology* 49(2):280–286.

Frum, David. 2020. "The Very Real Threat of Trump's Deepfake.".
URL: *https://www.theatlantic.com/ideas/archive/2020/04/trumps-first-deepfake/610750/*

Furnham, Adrian and Barrie Gunter. 1989. "The primacy of print: Immediate cued recall of news as a function of the channel of communication." *The Journal of General Psychology* 116(3):305–310.

Galston, William A. 2020. "Is seeing still believing? The deepfake challenge to truth in politics.".
URL: *https://www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/*

Gazis, Olivia and Stefan Becket. 2019. "Senators Pressure Social Media Giants to Crack Down on "Deepfakes".".
URL: *https://www.cbsnews.com/news/deepfakes-mark-warner-marco-rubio-pressure-social-media-giants-to-crack-down/*

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.

Goldberg, Matthew H, Sander van der Linden, Matthew T Ballew, Seth A Rosenthal, Abel Gustafson and Anthony Leiserowitz. 2019. "The experience of consensus: video as an effective medium to communicate scientific agreement on climate change." *Science Communication* 41(5):659–673.

Goodman, J. David. 2012. "Microphone Catches a Candid Obama.".

URL: *https://www.nytimes.com/2012/03/27/us/politics/obama-caught-on-microphone-telling-medvedev-of-flexibility.html*

Government Accountability Office: Science, Technology Assessment and Analytics. 2020. "Science and Tech Spotlight: Deepfakes.".
URL: *https://www.gao.gov/assets/710/704774.pdf*

Grabe, Maria Elizabeth and Erik Page Bucy. 2009. *Image Bite Politics: News and the Visual Framing of Elections*. Oxford University Press.

Graber, Doris A. 1990. "Seeing is remembering: How visuals contribute to learning from television news." *Journal of Communication* 40(3):134–155.

Gray, Jonathan, Jeffrey P Jones and Ethan Thompson. 2009. *Satire TV: Politics and Comedy in the Post-Network Era*. New York University Press.

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).

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.
URL: *https://www.pnas.org/content/117/27/15536*

Harris, Douglas. 2018. "Deepfakes: False pornography is here and the law cannot protect you." *Duke L. & Tech. Rev.* 17:99.

Harwell, Drew. 2019. "Top AI Researchers Race to Detect 'Deepfake' Videos: 'We are outgunned'.".
URL: *https://www.washingtonpost.com/technology/2019/06/12/top-ai-researchers-race-detect-deepfake-videos-we-are-outgunned/*

Herman, Edward S and Noam Chomsky. 2010. *Manufacturing Consent: The Political Economy of the Mass Media*. Random House.

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.

Hwang, Tim and Clint Watts. 2020. "Opinion: Deepfakes are coming for American democracy. Here's how we can prepare.".
URL: *https://www.washingtonpost.com/opinions/2020/09/10/deepfakes-are-coming-american-democracy-heres-how-we-can-prepare/*

Jamieson, Kathleen Hall, Kathleen Hall et al. 1995. *Beyond the Double Bind: Women and Leadership*. Oxford University Press.

Jerit, Jennifer and Yangzi Zhao. 2020. "Political misinformation." *Annual Review of Political Science* 23:77–94.

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

Kassin, Saul M and David A Garfield. 1991. "Blood and guts: General and trial-specific effects of videotaped crime scenes on mock jurors." *Journal of Applied Social Psychology* 21(18):1459–1472.

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

Krook, Mona Lena and Juliana Restrepo Sanín. 2020. "The cost of doing politics? Analyzing violence and harassment against female politicians." *Perspectives on Politics* 18(3):740–755.

Kuklinski, James H, Paul J Quirk, Jennifer Jerit, David Schwieder and Robert F Rich. 2000. "Misinformation and the currency of democratic citizenship." *Journal of Politics* 62(3):790–816.

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.

Lewis, Rebecca. 2018. "Alternative influence: Broadcasting the reactionary right on YouTube." *Data & Society* 18.

Lippmann, Walter. 1922. *Public Opinion*. New Cork.

Lum, Zi-Ann. 2019. "Obama Tells Canadian Crowd He's Worried About 'Deepfake' Videos.". **URL:** *https://www.huffingtonpost.ca/entry/obama-deepfake-video˙ca˙5cf29aafe4b0e8085e3ad233*

Lupia, Arthur. 2016. *Uninformed: Why People Know So Little About Politics and What We Can Do About It*. Oxford University Press.

Makhzani, Alireza, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow and Brendan Frey. 2015. "Adversarial Autoencoders." *Working Paper* .

Mori, Masahiro, Karl F MacDorman and Norri Kageki. 2012. "The Uncanny Valley." *IEEE Robotics & Automation Magazine* 19(2):98–100.

Munger, Kevin and Joseph Phillips. 2020. "Right-Wing YouTube: A Supply and Demand Perspective." *The International Journal of Press/Politics* .

Munger, Kevin, Mario Luca, Jonathan Nagler and Joshua Tucker. 2020. "The (null) effects of clickbait headlines on polarization, trust, and learning." *Public Opinion Quarterly* .

Mutz, Diana C. 2016. *In-your-face Politics: The Consequences of Uncivil Media*. Princeton University Press.

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.".

Parkin, Simon. 2019. "The Rise of the Deepfake and the Threat to Democracy.".
URL: *https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy*

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, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu and David G Rand. 2020. "Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention." *Psychological Science* 31(7):770–780.

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

Pezdek, Kathy, Elizabeth Avila-Mora and Kathryn Sperry. 2010. "Does trial presentation medium matter in jury simulation research? Evaluating the effectiveness of eyewitness expert testimony." *Applied Cognitive Psychology* 24(5):673–690.

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

Popkin, Samuel L. 1991. *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*. University of Chicago Press.

Press Releases. 2018.
URL: *https://www.rubio.senate.gov/public/index.cfm?p=Press-Releases*

Prior, Markus. 2014. "Visual political knowledge: A different road to competence?" *The Journal of Politics* 76(1):41–57.

Prochaska, Stephen, Michael Grass and Jevin West. 2020. "Deepfakes in the 2020 Election and Beyond: Lesseons From the 2020 Workshop Series." *Center for an Informed Republic* .
URL: *https://cpb-us-e1.wpmucdn.com/sites.uw.edu/dist/6/4560/files/2020/10/CIP˙Deepfakes˙Report*

Rini, Regina. 2020. "Deepfakes and the epistemic backstop." *Philosopher's Imprint* 20(24).

Rupar, Aaron. 2019. "Trump's bizarre "Tim/Apple" tweet is a reminder the president refuses to own tiny mistakes.".
URL: *https://www.vox.com/2019/3/11/18259996/trump-tim-cook-apple-tweet-time-and-words*

Sasse, Ben. 2018. "Opinion: This new technology could send American politics into a tailspin.".
URL: *https://www.washingtonpost.com/opinions/the-real-scary-news-about-deepfakes/2018/10/19/6238c3ce-d176-11e8-83d6-291fcead2ab1˙story.html*

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.

Schick, Nina. 2020. "Deepfakes are jumping from porn to politics. It's time to fight back.".
URL: *https://www.wired.co.uk/article/deepfakes-porn-politics*

Snyder Jr, James M and David Strömberg. 2010. "Press coverage and political accountability." *Journal of Political Economy* 118(2):355–408.

Stenberg, Georg. 2006. "Conceptual and perceptual factors in the picture superiority effect." *European Journal of Cognitive Psychology* 18(6):813–847.

Strömberg, David. 2004. "Mass media competition, political competition, and public policy." *The Review of Economic Studies* 71(1):265–284.

Suwajanakorn, Supasorn, Steven M Seitz and Ira Kemelmacher-Shlizerman. 2017. "Synthesizing obama: learning lip sync from audio." *ACM Transactions on Graphics (TOG)* 36(4):1–13.

Teele, Dawn, Joshua Kalla and Frances McCall Rosenbluth. 2017. "The ties that double bind: social roles and women's underrepresentation in politics." *American Political Science Review* .

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/?sh=7fd212087494*

Tucker, Joshua A, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal and Brendan Nyhan. 2018. "Social media, political polarization, and political disinformation: A review of the scientific literature." *Hewlett Foundation* .

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).

Vosoughi, Soroush, Deb Roy and Sinan Aral. 2018. "The spread of true and false news online." *Science* 359(6380):1146–1151.

Witten, Ilana B and Eric I Knudsen. 2005. "Why seeing is believing: merging auditory and visual worlds." *Neuron* 48(3):489–496.

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

Yadav, Aman, Michael M Phillips, Mary A Lundeberg, Matthew J Koehler, Katherine Hilden and Kathryn H Dirkin. 2011. "If a picture is worth a thousand words is video worth a million? Differences in affective and cognitive processing of video and text cases." *Journal of Computing in Higher Education* 23(1):15–37.

Zimmer, Ben. 2019. "Elizabeth Warren and the Down-to-Earth Trap.".
URL: *https://www.theatlantic.com/entertainment/archive/2019/01/why-elizabeth-warrens-beer-moment-fell-flat/579544/*

Political Deepfake Videos Misinform the Public, But No More than Other Fake Media

Supplementary Information

Contents

A Stimuli in Exposure Experiment

A.1 Production details

We discuss the ethical reasoning behind our research design in more detail in Appendix D, but we first highlight here our selection of Elizabeth Warren for on both ethical and practical grounds. Senator Elizabeth Warren is a salient politician, making our experiment more ecologically valid than one with a low-profile or hypothetical politician (nearly all political deepfakes target high-profile politicians), but she is not slated for re-election until 2024. We selected a female candidate because women are more likely to be the targets of non-political deepfakes (Ajder et al., 2019; Abram, 2020) and harassment more broadly (Krook and Sanín, 2020) and we specifically test whether pre-existing prejudice against women among subjects changes the effect of the deepfake.

Prior to production, we consulted *Buzzfeed* CEO Jonah Peretti, who produced the first viral deepfake video in 2018 of Barack Obama telling the world that "President Trump is a complete and utter dipsh*t". In the correspondence below, he explained how the deepfake, created via a professional actor's expert impersonation and synthesized via face-swap, came to exist, emphasizing the need for high-quality impersonator and post-production:

> "The idea was shaped by Jordan's ability to do a good Obama impersonation - so that part isn't fancy tech. Jordan is just better at impressions than other people making deep fakes and he did Obama as a character on Key & Peele.
>
> Then we worked with Jared who used a combination of deep fake software downloaded from Reddit and Adobe products we use to do video effects and post production work. It wasn't straightforward and required a combination of approaches and Jared's prodigious talents."

In collaboration with an industry partner and following the lessons from our correspondence with *Buzzfeed*, we produced a series of deepfake videos using target footage of 2020 presidential candidate and senator Elizabeth Warren and performances of a professional Elizabeth Warren impersonator. Warren's campaign disseminated a series of campaign video recordings of the senator in her home kitchen making personal thank-you calls to campaign donors and, in some cases, discussing policy matters and events during the campaign. We produced a series of videos performances of the impersonator in a similar kitchen performing several different sketches that each represented a potential "scandal" for Warren. To script these scandals, we carefully studied past controversial hot mic scandals of Democratic politicians as well as exact statements made by Warren in these campaign videos. We then scripted statements in Warren's natural tone and affliction that appeared plausible in the universe of political hot mic statements. As such, these statements are not meant to invoke extreme disbelief or incredulity, though testing the credulity threshold of deepfake scandals in a principled manner could the subject of future research. Table A3 describes the content of the final performances selected for our experiment. We used the audio from these sketches for the audio condition

Table A3: **Descriptions and Scripts of Scandal Performances**

Scandal Description	Title	Script
In-Party Incivility	LEAK: Elizabeth Warren calls Joe Biden "a piece of sh*t" and a pedophile in 2019 campaign call	"Why shouldn't you vote for Joe Biden in 2020? Well, I'll tell you why: because he's a sexist piece of shit who likes to grope young girls, that's why."
Out-Party Incivility	LEAK: Elizabeth Warren calls Donald Trump "a piece of sh*t" and a pedophile in 2019 campaign call	"Why shouldn't you vote for Donald Trump in 2020? Well, I'll tell you why: because he's a sexist piece of shit who likes to grope young girls, that's why."
Past Controversy (racialized comment)	LEAK: Elizabeth Warren re-claims Cherokee heritage in 2019 campaign call	"Well, you know, as someone who has Cherokee ancestry, who's proud of their Native heritage, I deeply identify with other indigenous people and people of color in this country and I will do everything I can to fight for you in Washington."
Novel Controversy (homophobic comment)	LEAK: Elizabeth Warren admits she doesn't "endorse the LGBTQ lifestyle" in 2019 campaign call	"Well, as a Christian woman of faith, I don't personally support the LGBTQ lifestyle, but I will try to do what I can for marriage equality in Washington."
Political Insincerity	LEAK: Elizabeth Warren flips stance on student loan debt in 2019 campaign call	"Well, I know I've said that before, but I don't really think that eliminating student loan debt for anyone is fair or realistic."

and the video plus audio for the parody skit. We then performed the procedure to create a face-swap deepfake to produce the final deepfake video treatments, one for each selected scandal performance.

A.2 Face-swap algorithm

Deepfakes that swap the face of a target (e.g., President Barack Obama) with an actor (e.g., Hollywood actor Jordan Peele) – dubbed face-swaps in Figure 1 – 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:

$$\min_{\substack{\pi, \\ \delta_{\text{target}}, \\ \delta_{\text{actor}}}} \mathbb{E}_{x \sim \mathbf{X}_{\text{target}}}\left[||\delta_{\text{target}}\{\pi(x)\} - x||^2\right] + \mathbb{E}_{x' \sim \mathbf{X}_{\text{actor}}}\left[||\delta_{\text{actor}}\{\pi(x')\} - x'||^2\right] \tag{1}$$

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:

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

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.

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.

B Stimuli in Detection Experiment

This section provides sceenshots of the videos used in the detection experiment. All subject are assigned a mix of videos in which there are either no deepfakes (i.e., all displayed videos are of real media), a low proportion of deepfakes, or a high proportion of deepfakes. Each of these three conditions employs eight videos. While the order in which videos are presented varies within these conditions, the videos within each condition are fixed across subjects.

This section shows screenshots of each of these videos. Section B.1 shows screenshots

and descriptions of all the authentic videos, while section Section B.2 shows screenshots and descriptions of the deepfakes employed in this experiment.

Subjects assigned to the no-fake condition saw real videos B6 through B13. Subjects in the low-fake condition saw fake videos B15 and B16, and real videos B8, B9, B10, B12, B13, and B14. Subjects in the high-fake condition saw fake videos B15, B16, B17, B18, B19, B20 and real videos B7 and B12.

Heterogeneity in detection performance at the clip level (both for the entire pool and across subgroups) can be found in Section G.

B.1 Real videos used

Figure B6: **Donald Trump ("soup" press conference gaffe)**. Following national demonstrations in the summer of 2020, President Donald Trump decries protestors weaponizing cans of soup against police officers in a soon-to-be viral press conference clip (Blum, 2020).

Figure B7: **Joe Biden (town hall 'push-up contest' gaffe)**. After a heated exchange, Democratic presidential candidate Joe Biden challenges a combative voter at a town hall to a push-up contest.

Figure B8: **Joe Biden (stutter gaffe)**. A video compilation of Joe Biden stuttering in various campaign appearances.

Figure B9: **Donald Trump (COVID-19 precautions announcement)**. In a public address from the White House, President Trump urges Americans to take personal precautions to avoid COVID-19.

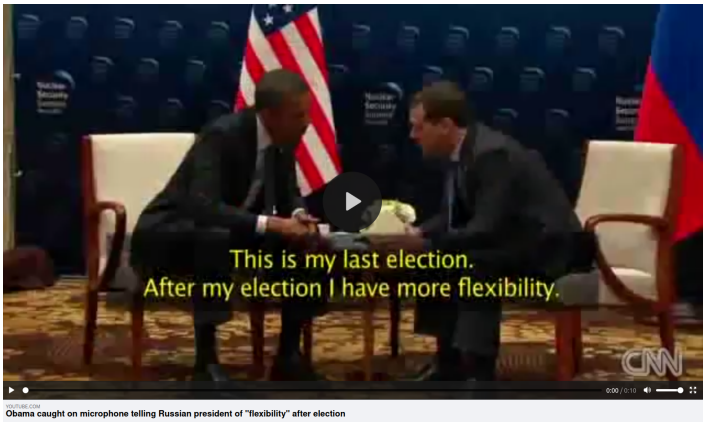


Figure B10: **Barack Obama (Russian president hot mic).** President Barack Obama is caught on a hot mic telling Russian President Dmitry Medvedev of "more flexibility" following his "last election" to negotiate on the issue of missile defense; an exchange that critics suggested revealed a lack of concern about re-election and lack of diplomatic transparency criticized (Goodman, 2012).

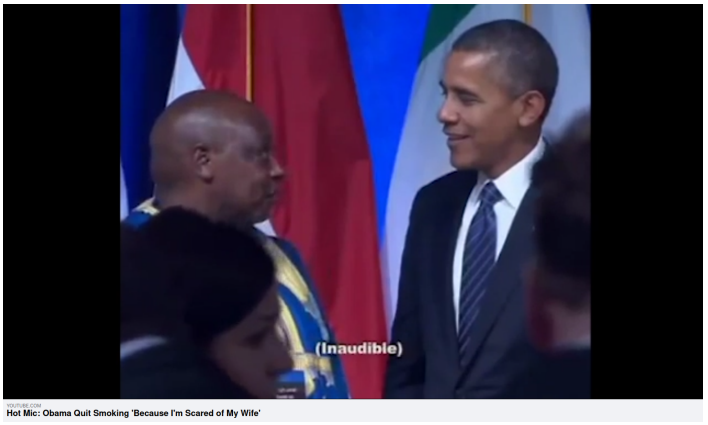


Figure B11: **Barack Obama (smoking hot mic).** President Barack Obama is caught on a hot mic to a U.N. National Assembly attendee saying that he quit smoking because "[he's] scared of [his] wife".

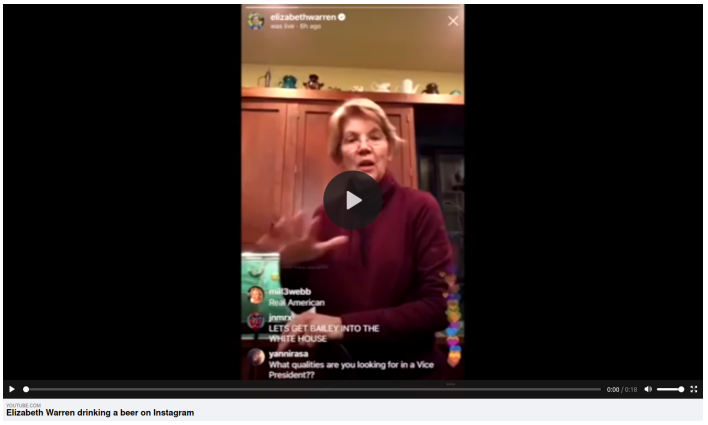


Figure B12: **Elizabeth Warren (Instagram beer gaffe).** Democratic primary candidate Elizabeth Warren furnishes a beer on an livestream video broadcasted on Instagram, a moment criticized as inauthentic and pandering by news media (Zimmer, 2019).

Figure B13: **Elizabeth Warren (post-debate hot mic).** Democratic primary candidate Elizabeth Warren confronts fellow candidate Bernie Sanders on live television for "calling [her] a liar on national TV".

Figure B14: **Donald Trump (Apple press conference gaffe).** During an on-camera White House event, President Donald Trump mistakenly calls Apple CEO Tim Cook "Tim Apple" in a clip to go viral soon after (Rupar, 2019).

B.2 Deepfake videos used

Figure B15: **Donald Trump (fake AIDS cure announcement).** In a campaign rally speech, President Donald Trump announces that under his administration, scientists have found a cure to AIDS.

Figure B16: **Barack Obama (fake news announcement).** In a White House address, President Barack Obama stresses the importance of relying on trusted news sources.

Figure B17: **Bernie Sanders (fake debate).** In a televised presidential town hall event, Democratic primary candidate Bernie Sanders recalls marching for civil rights in Selma, Alabama.

Figure B18: **Boris Johnson (fake Brexit announcement).** Sitting Prime Minister Boris Johnson announces that in order to "rise above the divide [on Brexit]", he will endorse opposition party leader Jeremy Corbyn in the upcoming U.K. general election.

Figure B19: **Donald Trump (fake resignation announcement).** In a White House address, President Donald Trump notes the American public's disappointment in his leadership and announces his resignation before the 2020 election, citing a need to "put the interests of America first".

Figure B20: **Hillary Clinton (fake debate).** In a televised debate, 2016 Democratic presidential candidate Hillary Clinton labels opponent Donald Trump's tax plan as only benefiting the 1%.

C Pre-Analysis Plan

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

$$\mathbf{X}_i = \Big(\mathsf{Age}_i, \mathsf{Educ}_i, \mathsf{Gender}_i, \mathsf{PID}_i, \mathsf{DigLit}_i, \mathsf{Internet}_i, \mathsf{PolKnow}_i\Big). \tag{3}$$

where DigLit_i measures digital literacy, $\mathsf{PolKnow}_i$ measures political knowledge according to a standard inventory of questions, $\mathsf{Internet}_i$ measures internet usage (1-7).

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 figures, 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

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

and

$$\tau_{1b} = \mathbb{E}\Big[\mathsf{Believe}_i(\mathsf{Exposed}_i = \texttt{video})\Big] - \mathbb{E}\Big[\mathsf{Believe}_i(\mathsf{Exposed}_i = \texttt{audio})\Big] > 0, \tag{5}$$

and

$$\tau_{1c} = \mathbb{E}\Big[\mathsf{Believe}_i(\mathsf{Exposed}_i = \texttt{video})\Big] - \mathbb{E}\Big[\mathsf{Believe}_i(\mathsf{Exposed}_i = \texttt{skit})\Big] > 0. \tag{6}$$

Additionally, we perform a parametric test adjusting for the aforementioned control covariates

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

via the following linear model estimated via

$$\mathsf{Believe}_i = \boldsymbol{\tau}_1 \mathbf{1}_{\mathsf{Exposed}_i} + \beta \mathbf{X}_i + \epsilon_i, \tag{7}$$

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,

$$\mathsf{Distrust}_i = \boldsymbol{\tau}_{3a_{\mathrm{I}}} \mathsf{InfoAware}_i + \beta \mathbf{X}_i + \epsilon_i, \tag{8}$$

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

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 to fake news content, $\mathsf{InfoAcc}_i = 1$.

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

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

$$\mathsf{DetectFPR}_i = \boldsymbol{\tau}_{3b_{\mathrm{II}}} (\mathbf{1}\{\mathsf{Exposed}_i = \texttt{video}\} \times \mathsf{Belief}_i) + \beta \mathbf{X}_i + \epsilon_i, \tag{11}$$

$$\mathsf{DetectFPR}_i = \tau_{3b_{\mathrm{III}}} \mathsf{DebriefBefore}_i + \beta \mathbf{X}_i + \epsilon_i, \tag{12}$$

$$\mathsf{DetectFPR}_i = \tau_{3b_{\mathrm{IV}}} \mathsf{InfoAcc}_i + \beta \mathbf{X}_i + \epsilon_i. \tag{13}$$

We register that $\tau_{3b_{\text{I}}}, \boldsymbol{\tau}_{3b_{\text{II}}}, \tau_{3b_{\text{III}}}, \tau_{3b_{\text{IV}}}$ 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:

$$\mathsf{Believe}_i = \boldsymbol{\tau}_4^{(1)}(\mathbf{1}_{\mathsf{Exposed}_i} \times \mathsf{InfoAware}_i) + \boldsymbol{\tau}_4^{(2)}\mathbf{1}_{\mathsf{Exposed}_i} + \boldsymbol{\tau}_4^{(3)}\mathsf{InfoAware}_i + \beta_{5a}\mathbf{X}_i + \epsilon_i. \tag{14}$$

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:

$$\mathsf{Believe}_i = \boldsymbol{\tau}_5^{(1)}(\mathbf{1}_{\mathsf{Exposed}_i} \times \mathsf{CR}_i) + \boldsymbol{\tau}_5^{(2)}\mathbf{1}_{\mathsf{Exposed}_i} + \boldsymbol{\tau}_5^{(3)}\mathsf{CR}_i + \beta\mathbf{X}_i + \epsilon_i. \tag{15}$$

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:

$$\begin{aligned}
\mathsf{Believe}_i =& \boldsymbol{\tau}_{6a}^{(1)}(\mathbf{1}_{\mathsf{Exposed}_i} \times \mathsf{PID}_i \times \mathsf{CR}_i) + && (16)\\
& \boldsymbol{\tau}_{6a}^{(2)}(\mathbf{1}_{\mathsf{Exposed}_i} \times \mathsf{PID}_i) + \boldsymbol{\tau}_{6a}^{(3)}(\mathbf{1}_{\mathsf{Exposed}_i} \times \mathsf{CR}_i) + \boldsymbol{\tau}_{6a}^{(4)}(\mathsf{PID}_i \times \mathsf{CR}_i) + && (17)\\
& \boldsymbol{\tau}_{6a}^{(5)}\mathbf{1}_{\mathsf{Exposed}_i} + \boldsymbol{\tau}_{6a}^{(6)}\mathsf{PID}_i + \boldsymbol{\tau}_{6a}^{(7)}\mathsf{CR}_i + \beta_{6a}\mathbf{X}_i + \epsilon_i && (18)
\end{aligned}$$

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:

$$\mathsf{DetectAcc}_i = \tau_8\mathsf{InfoAcc}_i + \beta\mathbf{X}_i + \epsilon_i, \tag{19}$$

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:

$$\mathsf{DetectAcc}_i = \tau_9 \mathsf{DigLit}_i + \beta \mathbf{X}_i + \epsilon_i, \tag{20}$$

and register that τ_9 will be positive.

D Ethical Considerations

We highlight the ethical considerations pursuant to a study that uses stimuli which are expected 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 specif-

ically, our study can inform future legislation or platform policies designed to minimize the threat posed by this technology.[9]

E Sample Description

Our survey experiment was fielded to a nationally representative sample on the Lucid survey research platform to a total of 17501 subjects launched in two waves between September 29th 2020 and October 29th 2020. Of this 17501, only 5750 subjects successfully completed the survey experiment or passed a series of quality checks. One of these quality checks was a battery of randomly dispersed attention checks in response to a recently-publicized issue with in-attention among survey respondents during this period as documented in Aronow et al. (2020). Additionally, we imposed a series of "technology checks," namely that the subjects be able to watch and listen to a video. In addition, 629 respondents failed front-end pre-treatment attention checks: namely, they entered gender or age values that did not match up (or come to close to matching up) with respondent demographic characteristics provided by Lucid. We coded these respondents as "low-quality" respondents which we drop in our analyses as a robustness measure. As expected by Aronow et al. (2020), results largely hold across the two cohorts, but nearly all coefficient estimates are slightly diminished for the low-quality cohort.

Table E4 compares our sample's demographic traits to the demographic traits in the most recent Current Population Survey (CPS) – in particular, traits like education, age, and household income that are hypothesized to have correlations with deepfake deception and affective appeal (by their correlation with digital literacy, internet usage, and political knowledge) as well race, gender and ethnicity which are correlates of partisanship, another predictor of our measured behavioral responses. To adjust for remaining disrepancies, we generate post-stratification weights via raking to match the CPS marginal population totals. We perform weighted regression in our analyses as a robustness measure to guard against measurement error from possible demographic skews.

F Pre-Registered Analyses

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

Table E4: **Sample Demographics and Representativeness after Post-stratification**

		CPS	Unweighted Sample	Weighted Sample
Education	<High school	10.95%	1.1%	2.87%
	High school	47.14%	29.88%	45.52%
	College	30.3%	47.17%	35.98%
	Postgraduate	11.61%	21.86%	15.63%
Age	18-24	10.42%	6.33%	8.9%
	25-34	13.88%	12.66%	14.91%
	35-44	12.58%	16.97%	16.91%
	45-64	25.76%	31.25%	34.17%
	65+	15.81%	32.77%	25.11%
Household Income	<$25k	19.11%	30.16%	22.63%
	$100k-$150k	14.95%	6.1%	10.67%
	>$150k	15.47%	4.77%	10.44%
	$25k-$49k	20.79%	21.74%	23.17%
	$50k-$74k	17.2%	15.63%	18.6%
	$75k-$99k	12.48%	19.27%	14.5%
Gender	Male	48.75%	33.58%	44.19%
	Female	51.25%	66.14%	55.81%
Race	Asian	5.42%	3.95%	4.4%
	Black	10.28%	5.76%	8.41%
	Other	4.18%	3.81%	3.79%
	White	80.12%	85.79%	83.4%
Hispanic	Yes	14.66%	5.18%	8.59%
	No	85.34%	94.1%	91.41%

Notes: Weights are constructed via Iterative Proportional Fitting to match sample marginal totals to CPS marginal totals on displayed demographic traits. Weights in the final column used for all analyses in paper.

Table F5: **Models of Belief in Exposure-Stage (News Feed) Scandal Clipping**

	Extent of belief that clipping was not fake or doctored [1-5]						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Audio	0.08	0.14	0.10	0.10	0.12	0.12	0.17*
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
Text	−0.02	−0.11	0.03	−0.01	−0.15*	0.02	−0.06
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
Skit	−0.68***	−0.61***	−0.65***	−0.68***	−0.66***	−0.67***	−0.60***
	(0.10)	(0.10)	(0.11)	(0.10)	(0.10)	(0.11)	(0.10)
On Mobile				0.05	0.09	0.17*	0.28***
				(0.08)	(0.08)	(0.08)	(0.09)
Age 65+				0.05	0.05	0.09	0.13*
				(0.07)	(0.07)	(0.07)	(0.08)
High School				−0.04	−0.26	−0.25	−0.29
				(0.33)	(0.20)	(0.34)	(0.20)
College				−0.03	−0.24	−0.21	−0.21
				(0.33)	(0.20)	(0.34)	(0.20)
Postgrad				−0.21	−0.36*	−0.39	−0.37*
				(0.34)	(0.21)	(0.34)	(0.22)
Independent PID					0.26**	0.11	0.23**
					(0.10)	(0.11)	(0.11)
Republican PID					0.60***	0.56***	0.61***
					(0.07)	(0.08)	(0.08)
CRT				−0.15	−0.15	−0.03	0.08
				(0.14)	(0.14)	(0.15)	(0.15)
Male				−0.01	0.03	0.03	0.06
				(0.07)	(0.07)	(0.07)	(0.07)
Political Knowledge				0.06	0.23	0.09	0.27*
				(0.15)	(0.15)	(0.15)	(0.15)
Internet Usage				0.05	0.09*	0.07	0.09*
				(0.05)	(0.05)	(0.05)	(0.05)
Ambivalent Sexism				0.16***	0.04	0.05	0.07
				(0.04)	(0.04)	(0.04)	(0.04)
Constant	3.40***	3.44***	3.39***	2.68***	2.55***	2.67***	2.28***
	(0.06)	(0.06)	(0.07)	(0.50)	(0.42)	(0.52)	(0.44)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
N	1,619	1,619	1,445	1,619	1,619	1,445	1,445
R^2	0.04	0.04	0.04	0.05	0.09	0.09	0.09
Adjusted R^2	0.04	0.03	0.04	0.04	0.08	0.08	0.08

*p < .1; **p < .05; ***p < .01

Notes: Reference category for medium is Video. CRT is scaled 0-1, political knowledge and ambivalent sexism are 0-1, internet usage is 1-7. Sample did not receive information in the first stage.

Table F6: **Models of Binarized Belief in Exposure-Stage (News Feed) Scandal Clipping**

	Belief that clipping was not fake or doctored [y/n]						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Audio	0.01	0.06*	0.02	0.02	0.05	0.03	0.06*
	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
Text	−0.04	−0.05	−0.02	−0.05	−0.07**	−0.03	−0.04
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Skit	−0.17***	−0.14***	−0.16***	−0.18***	−0.16***	−0.17***	−0.14***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
On Mobile				0.07**	0.09***	0.10***	0.15***
				(0.03)	(0.03)	(0.03)	(0.03)
Age 65+				0.03	0.03	0.04	0.06*
				(0.03)	(0.03)	(0.03)	(0.03)
High School				−0.08	−0.13*	−0.12	−0.14*
				(0.12)	(0.08)	(0.13)	(0.08)
College				−0.07	−0.11	−0.10	−0.11
				(0.12)	(0.08)	(0.13)	(0.08)
Postgrad				−0.10	−0.11	−0.14	−0.11
				(0.12)	(0.08)	(0.13)	(0.08)
Independent PID				0.02	0.04	0.004	0.03
				(0.04)	(0.04)	(0.04)	(0.04)
Republican PID				0.20***	0.21***	0.20***	0.21***
				(0.03)	(0.03)	(0.03)	(0.03)
CRT				−0.06	−0.06	−0.03	0.02
				(0.05)	(0.05)	(0.06)	(0.06)
Male				0.03	0.03	0.03	0.03
				(0.03)	(0.03)	(0.03)	(0.03)
Political Knowledge				0.12**	0.21***	0.14**	0.25***
				(0.06)	(0.06)	(0.06)	(0.06)
Internet Usage				0.03*	0.04*	0.03*	0.04*
				(0.02)	(0.02)	(0.02)	(0.02)
Ambivalent Sexism				0.02	0.03*	0.02	0.03*
				(0.02)	(0.02)	(0.02)	(0.02)
Constant	0.47***	0.47***	0.46***	0.09	0.02	0.07	−0.08
	(0.02)	(0.02)	(0.03)	(0.19)	(0.16)	(0.20)	(0.16)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
N	1,619	1,619	1,445	1,619	1,619	1,445	1,445
R^2	0.02	0.02	0.02	0.07	0.08	0.07	0.09
Adjusted R^2	0.01	0.02	0.01	0.06	0.07	0.06	0.08

*p < .1; **p < .05; ***p < .01

Notes: Reference category for medium is Video. CRT is scaled 0-1, political knowledge and ambivalent sexism are 0-1, internet usage is 1-7. Sample did not receive information in the first stage.

Table F7: **Interactive Models of Information Provision and Belief in Clipping**

	Belief that clipping was not fake or doctored [y/n]						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Info Provided	−0.11***	−0.09***	−0.10***	−0.12***	−0.11***	−0.12***	−0.09**
	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.04)
Audio	0.01	0.06*	0.02	0.02	0.06*	0.03	0.07**
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Text	−0.04	−0.05	−0.02	−0.05	−0.06*	−0.03	−0.03
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Skit	−0.17***	−0.14***	−0.16***	−0.18***	−0.16***	−0.17***	−0.14***
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Info x Audio	0.02	−0.04	0.01	0.03	−0.02	0.02	−0.04
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Info x Text	0.09**	0.08*	0.07	0.11**	0.12***	0.10**	0.09*
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Info x Skit	0.04	0.05	0.01	0.06	0.09*	0.05	0.05
	(0.05)	(0.05)	(0.06)	(0.05)	(0.05)	(0.06)	(0.06)
Constant	0.47***	0.47***	0.46***	0.15	0.24**	0.13	0.20
	(0.02)	(0.02)	(0.02)	(0.13)	(0.11)	(0.14)	(0.12)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	3,267	3,267	2,907	3,267	3,267	2,907	2,907
R^2	0.02	0.02	0.02	0.06	0.06	0.06	0.06
Adjusted R^2	0.02	0.02	0.02	0.06	0.05	0.06	0.06

*p < .1; **p < .05; ***p < .01

Notes: Respondents subset to those exposed to a clipping. Reference category for medium is Video.

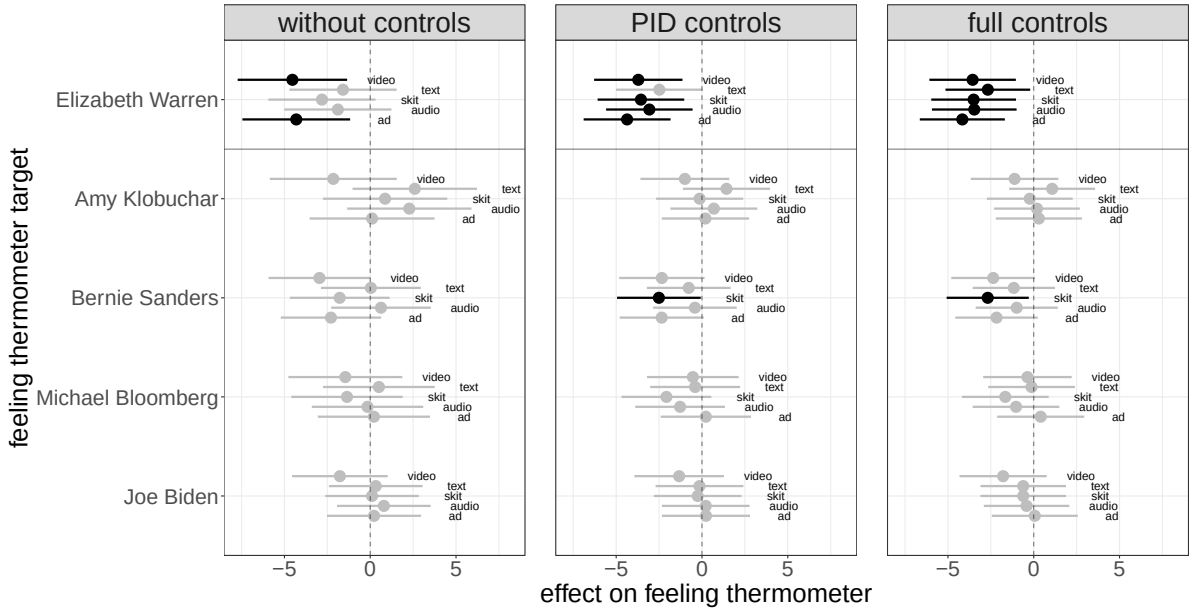
Table F8: **Models of Scandal Target Affect**

	Elizabeth Warren Feeling Thermometer						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Video	−4.54***	−4.08**	−2.86	−3.62***	−4.23***	−2.83**	−3.69***
	(1.63)	(1.62)	(1.75)	(1.29)	(1.31)	(1.37)	(1.39)
Audio	−1.89	−3.09*	−0.68	−3.49***	−4.64***	−2.91**	−4.10***
	(1.59)	(1.59)	(1.70)	(1.26)	(1.28)	(1.33)	(1.37)
Text	−1.59	−0.75	−1.04	−2.66**	−1.32	−2.42*	−1.77
	(1.59)	(1.59)	(1.70)	(1.26)	(1.28)	(1.33)	(1.35)
Skit	−2.81*	−3.12**	−2.00	−3.50***	−3.72***	−3.13**	−4.17***
	(1.59)	(1.58)	(1.70)	(1.26)	(1.28)	(1.33)	(1.35)
Attack Ad	−4.31***	−4.58***	−2.85*	−4.12***	−3.89***	−3.59***	−3.93***
	(1.60)	(1.57)	(1.71)	(1.27)	(1.27)	(1.34)	(1.34)
Info Provided				0.69	0.60	0.66	0.72
				(0.73)	(0.74)	(0.77)	(0.78)
On Mobile				−1.62*	−0.78	−2.73***	−2.62**
				(0.97)	(0.99)	(1.02)	(1.04)
Age 65+				−4.75***	−5.43***	−4.17***	−5.31***
				(0.82)	(0.86)	(0.86)	(0.89)
High School				−0.98	−1.38	−1.92	−1.75
				(3.60)	(2.30)	(3.72)	(2.33)
College				1.18	1.81	−0.74	0.37
				(3.59)	(2.34)	(3.72)	(2.39)
Postgrad				11.05***	14.29***	8.77**	12.71***
				(3.65)	(2.48)	(3.79)	(2.55)
Independent PID				−26.92***	−26.96***	−26.68***	−27.49***
				(1.21)	(1.22)	(1.26)	(1.27)
Republican PID				−39.88***	−37.88***	−40.48***	−39.10***
				(0.83)	(0.84)	(0.88)	(0.89)
CRT				−1.59	−1.98	−0.69	−0.76
				(1.62)	(1.65)	(1.72)	(1.75)
Male				0.65	−0.45	0.41	−0.21
				(0.82)	(0.80)	(0.87)	(0.85)
Political Knowledge				1.18	0.92	1.22	0.80
				(1.68)	(1.67)	(1.77)	(1.76)
Internet Usage				0.99*	0.93	1.15*	1.30**
				(0.57)	(0.58)	(0.61)	(0.62)
Ambivalent Sexism				−3.76***	−3.13***	−4.43***	−3.56***
				(0.47)	(0.47)	(0.50)	(0.50)
Constant	45.81***	45.11***	44.17***	69.17***	67.74***	70.75***	67.56***
	(1.14)	(1.12)	(1.23)	(5.51)	(4.79)	(5.84)	(5.12)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
N	5,524	5,524	4,895	5,523	5,523	4,894	4,894
R^2	0.002	0.002	0.001	0.38	0.35	0.39	0.37
Adjusted R^2	0.001	0.002	−0.0000	0.38	0.35	0.39	0.37

*p < .1; **p < .05; ***p < .01

Notes: Reference category for medium is Control.

Figure F21: **Effects of Clip Modality on Affect Towards Placebo Targets**



Notes: Shown are other politicians who ran in the 2020 Democratic primary. Conservatively, we should see a treatment effect of a Elizabeth Warren deepfake on affective responses to these placebos to the extent attitudes about Elizabeth Warren are correlated with attitudes about other primary candidates.

Table F9: **Models of Deepfake Salience (via Information Provision) and Media Trust Across Sources**

	Trust in...							
	Offline Media		Online Media		Social Media		Combined Index	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Info Provided	0.01	0.01	−0.02	−0.02	−0.02	−0.02	−0.01	−0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	2.65***	2.78***	2.34***	1.95***	1.91***	1.92***	2.30***	2.22***
	(0.02)	(0.15)	(0.01)	(0.14)	(0.02)	(0.15)	(0.01)	(0.12)
Controls?		✓		✓		✓		✓
Observations	5,533	5,533	5,533	5,533	5,532	5,532	5,533	5,533
R^2	0.0000	0.17	0.0002	0.15	0.0002	0.18	0.0001	0.21
Adjusted R^2	−0.0002	0.17	0.0000	0.15	−0.0000	0.18	−0.0001	0.20

Note: Reference categories for medium and age are Video and 18-30 respectively.

Table F10: **Models of Deepfake Salience (via Recognition) and Media Trust Across Sources**

	Trust in Media (Combined Index)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Believed	−0.09***	−0.06***	−0.10***	−0.07***	−0.06***	−0.07***	−0.14***
	(0.01)	(0.01)	(0.03)	(0.01)	(0.01)	(0.01)	(0.03)
Video	−0.23**	−0.22**	−0.08*	−0.41***	−0.18*	−0.34***	−0.19***
	(0.10)	(0.09)	(0.04)	(0.09)	(0.10)	(0.10)	(0.05)
Believed x Video	0.06**	0.05**	0.07	0.09***	0.04	0.07***	0.18***
	(0.03)	(0.03)	(0.07)	(0.03)	(0.03)	(0.03)	(0.07)
Constant	2.59***	2.48***	2.34***	2.28***	2.52***	2.37***	2.14***
	(0.04)	(0.20)	(0.20)	(0.17)	(0.21)	(0.18)	(0.17)
Weighted?				✓		✓	✓
Low-Quality Dropped?					✓	✓	✓
Controls?		✓	✓	✓	✓	✓	✓
Belief Binarized?			✓				✓
N	2,073	2,073	2,073	2,073	1,852	1,852	2,073
R^2	0.03	0.22	0.21	0.26	0.21	0.24	0.25
Adjusted R^2	0.02	0.21	0.20	0.25	0.20	0.23	0.25

Notes: Respondents subset to those exposed to a clipping and not provided with info before exposure. Estimates are larger but similar in magnitude when excluding controls.

Table F11: **Models of Deepfake Salience (via Recognition) and Media Trust Across Sources**

	Trust in...							
	Offline Media		Online Media		Social Media		Combined Index	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Believed	−0.08***	−0.15***	−0.05***	−0.09**	−0.05***	−0.06	−0.06***	−0.10***
	(0.01)	(0.04)	(0.01)	(0.04)	(0.02)	(0.04)	(0.01)	(0.03)
Video	−0.21*	−0.07	−0.23**	−0.09*	−0.23**	−0.09	−0.22**	−0.08*
	(0.12)	(0.06)	(0.11)	(0.05)	(0.12)	(0.06)	(0.09)	(0.04)
Believed x Video	0.05	0.07	0.06**	0.12	0.04	0.03	0.05**	0.07
	(0.03)	(0.08)	(0.03)	(0.08)	(0.03)	(0.09)	(0.03)	(0.07)
Constant	2.93***	2.75***	2.12***	2.00***	2.39***	2.28***	2.48***	2.34***
	(0.26)	(0.25)	(0.24)	(0.24)	(0.26)	(0.26)	(0.20)	(0.20)
Controls?	✓	✓	✓	✓	✓	✓	✓	✓
Belief Binarized?		✓		✓		✓		✓
Observations	2,073	2,073	2,073	2,073	2,072	2,072	2,073	2,073
R^2	0.19	0.19	0.16	0.15	0.18	0.18	0.22	0.21
Adjusted R^2	0.18	0.18	0.15	0.15	0.18	0.17	0.21	0.20

Note: Respondents subset to those exposed to a clipping and not provided with info before exposure. Estimates are larger but similar in magnitude when excluding controls.

Table F12: **Interactive Models of Cognitive Reflection and Belief in Clipping**

	Extent of belief that clipping was not fake or doctored [1-5]						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Audio	0.08	0.05	0.11	0.10	0.09	0.15	0.18
	(0.11)	(0.11)	(0.11)	(0.10)	(0.11)	(0.11)	(0.11)
Text	0.11	−0.04	0.14	0.11	−0.04	0.14	−0.02
	(0.10)	(0.11)	(0.11)	(0.10)	(0.10)	(0.11)	(0.11)
Skit	−0.51***	−0.43***	−0.53***	−0.50***	−0.42***	−0.51***	−0.46***
	(0.12)	(0.12)	(0.13)	(0.12)	(0.12)	(0.12)	(0.12)
CRT	−0.09	−0.12	−0.03	−0.09	−0.10	−0.01	−0.04
	(0.20)	(0.20)	(0.22)	(0.20)	(0.20)	(0.21)	(0.21)
CRT x Audio	0.12	0.13	0.07	0.17	0.10	0.10	−0.03
	(0.28)	(0.28)	(0.30)	(0.27)	(0.28)	(0.29)	(0.30)
CRT x Text	−0.11	0.11	−0.07	0.01	0.16	0.03	0.24
	(0.27)	(0.28)	(0.29)	(0.26)	(0.27)	(0.28)	(0.29)
CRT x Skit	−0.47	−0.47	−0.46	−0.38	−0.42	−0.38	−0.26
	(0.32)	(0.32)	(0.34)	(0.31)	(0.31)	(0.33)	(0.33)
Constant	3.26***	3.33***	3.23***	2.85***	3.17***	2.88***	3.06***
	(0.08)	(0.08)	(0.08)	(0.35)	(0.30)	(0.38)	(0.32)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	3,267	3,267	2,907	3,267	3,267	2,907	2,907
R^2	0.04	0.03	0.05	0.10	0.08	0.10	0.09
Adjusted R^2	0.04	0.03	0.04	0.09	0.07	0.10	0.08

*p < .1; **p < .05; ***p < .01

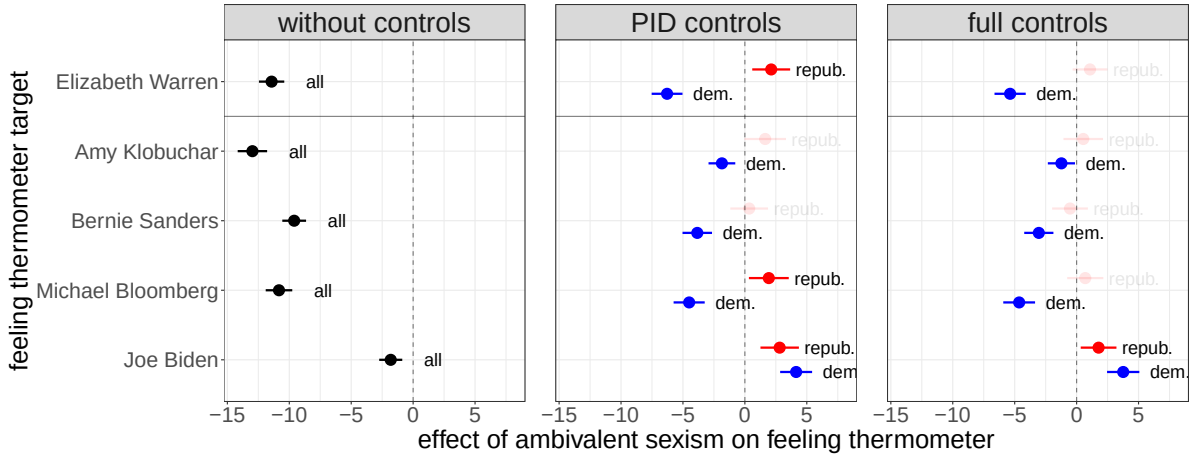
Notes: Reference category for medium is Video.

Table F13: **Interactive Models of Cognitive Reflection and Binarized Belief in Clipping**

	Belief that clipping was not fake or doctored [y/n]						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Audio	−0.01	−0.01	−0.01	−0.01	0.01	0.003	0.03
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Text	−0.001	−0.03	0.01	−0.001	−0.03	0.01	−0.02
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Skit	−0.13***	−0.11**	−0.13***	−0.13***	−0.10**	−0.13***	−0.11**
	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)	(0.05)
CRT	−0.09	−0.12*	−0.08	−0.11	−0.13*	−0.09	−0.10
	(0.07)	(0.07)	(0.08)	(0.07)	(0.07)	(0.08)	(0.08)
CRT x Audio	0.11	0.13	0.11	0.12	0.11	0.11	0.08
	(0.10)	(0.10)	(0.11)	(0.10)	(0.10)	(0.11)	(0.11)
CRT x Text	0.004	0.08	0.01	0.04	0.10	0.04	0.09
	(0.10)	(0.10)	(0.11)	(0.10)	(0.10)	(0.10)	(0.11)
CRT x Skit	−0.07	−0.05	−0.08	−0.05	−0.03	−0.05	−0.01
	(0.12)	(0.12)	(0.12)	(0.11)	(0.11)	(0.12)	(0.12)
Constant	0.45***	0.47***	0.44***	0.14	0.24**	0.14	0.21*
	(0.03)	(0.03)	(0.03)	(0.13)	(0.11)	(0.14)	(0.12)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	3,267	3,267	2,907	3,267	3,267	2,907	2,907
R^2	0.02	0.01	0.02	0.06	0.06	0.06	0.06
Adjusted R^2	0.02	0.01	0.02	0.05	0.05	0.06	0.05

*p < .1; **p < .05; ***p < .01
Notes: Reference category for medium is Video.

Figure F22: **Ambivalent Sexism and Affect Towards Placebo Targets**



Notes: Shown are other politicians who ran in the 2020 Democratic primary. Conservatively, we would expect that ambivalent sexism would predict the strongest negative affective response towards other female candidates.

Table F14: **Interactive Models of Partisan Motivated Reasoning and Belief in Clipping**

	Extent of belief that clipping was not fake or doctored [1-5]						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Repub PID	0.49***	0.29*	0.44***	0.48***	0.31**	0.43***	0.26
	(0.15)	(0.15)	(0.16)	(0.15)	(0.15)	(0.16)	(0.16)
Audio	0.13	0.08	0.15	0.15	0.13	0.18	0.21
	(0.14)	(0.14)	(0.15)	(0.14)	(0.14)	(0.15)	(0.15)
Text	0.09	−0.17	0.06	0.11	−0.13	0.10	−0.16
	(0.13)	(0.14)	(0.14)	(0.13)	(0.14)	(0.14)	(0.14)
Skit	−0.54***	−0.58***	−0.58***	−0.52***	−0.57***	−0.57***	−0.67***
	(0.16)	(0.16)	(0.17)	(0.15)	(0.16)	(0.16)	(0.16)
CRT	−0.005	−0.20	0.02	0.03	−0.13	0.04	−0.11
	(0.26)	(0.26)	(0.27)	(0.26)	(0.26)	(0.27)	(0.27)
Repub x Audio	−0.10	−0.05	−0.05	−0.14	−0.10	−0.10	−0.08
	(0.21)	(0.22)	(0.23)	(0.21)	(0.21)	(0.23)	(0.23)
Repub x Text	0.05	0.31	0.19	−0.02	0.23	0.10	0.36
	(0.21)	(0.21)	(0.22)	(0.21)	(0.21)	(0.22)	(0.22)
Repub x Skit	0.10	0.38	0.17	0.06	0.36	0.13	0.50**
	(0.24)	(0.24)	(0.25)	(0.24)	(0.24)	(0.25)	(0.25)
CRT x Repub	−0.22	0.14	−0.10	−0.26	0.08	−0.14	0.18
	(0.40)	(0.40)	(0.44)	(0.40)	(0.39)	(0.43)	(0.42)
Audio x CRT	−0.05	−0.06	−0.11	−0.03	−0.11	−0.08	−0.22
	(0.35)	(0.37)	(0.38)	(0.35)	(0.37)	(0.38)	(0.39)
Text x CRT	−0.17	0.19	0.02	−0.15	0.12	0.03	0.37
	(0.34)	(0.36)	(0.36)	(0.34)	(0.36)	(0.36)	(0.38)
Skit x CRT	−0.56	−0.36	−0.46	−0.51	−0.35	−0.38	−0.01
	(0.41)	(0.41)	(0.44)	(0.41)	(0.41)	(0.43)	(0.43)
Repub x Audio x CRT	0.48	0.47	0.44	0.51	0.52	0.47	0.48
	(0.55)	(0.56)	(0.60)	(0.55)	(0.56)	(0.60)	(0.60)
Repub x Text x CRT	0.37	−0.06	−0.10	0.47	0.11	0.02	−0.28
	(0.55)	(0.55)	(0.59)	(0.55)	(0.55)	(0.59)	(0.58)
Repub x Skit x CRT	0.31	−0.12	0.07	0.28	−0.17	0.02	−0.56
	(0.63)	(0.63)	(0.68)	(0.62)	(0.63)	(0.68)	(0.67)
Constant	3.04***	3.21***	3.04***	2.92***	3.37***	2.94***	3.25***
	(0.10)	(0.10)	(0.11)	(0.36)	(0.31)	(0.38)	(0.33)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	3,267	3,267	2,907	3,267	3,267	2,907	2,907
R^2	0.08	0.07	0.08	0.10	0.08	0.10	0.09
Adjusted R^2	0.07	0.06	0.08	0.09	0.07	0.09	0.08

*p < .1; **p < .05; ***p < .01

Notes: PID is pooled to Republican/Not Republican for brevity; reference category for medium is Video.

Table F15: **Interactive Models of Partisan Motivated Reasoning and Binarized Belief in Clipping**

	Belief that clipping was not fake or doctored [y/n]						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Repub PID	0.15***	0.10*	0.12**	0.15***	0.11**	0.13**	0.09
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Audio	−0.01	0.02	−0.01	−0.01	0.04	−0.01	0.05
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Text	−0.03	−0.08	−0.04	−0.03	−0.07	−0.03	−0.09*
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Skit	−0.15***	−0.16***	−0.16***	−0.15***	−0.16***	−0.16***	−0.19***
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
CRT	−0.05	−0.11	−0.06	−0.05	−0.10	−0.07	−0.10
	(0.09)	(0.10)	(0.10)	(0.09)	(0.10)	(0.10)	(0.10)
Repub x Audio	0.01	−0.06	0.03	0.01	−0.07	0.02	−0.06
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
Repub x Text	0.07	0.12	0.12	0.05	0.11	0.10	0.17**
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
Repub x Skit	0.05	0.13	0.08	0.04	0.13	0.07	0.19**
	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
CRT x Repub	−0.12	−0.04	−0.04	−0.13	−0.08	−0.06	−0.004
	(0.15)	(0.15)	(0.16)	(0.15)	(0.14)	(0.16)	(0.16)
Audio x CRT	0.05	−0.01	0.07	0.05	−0.04	0.07	−0.05
	(0.13)	(0.14)	(0.14)	(0.13)	(0.13)	(0.14)	(0.14)
Text x CRT	−0.03	0.06	0.04	−0.03	0.05	0.04	0.14
	(0.12)	(0.13)	(0.13)	(0.12)	(0.13)	(0.13)	(0.14)
Skit x CRT	−0.08	−0.03	−0.03	−0.07	−0.03	−0.01	0.08
	(0.15)	(0.15)	(0.16)	(0.15)	(0.15)	(0.16)	(0.16)
Repub x Audio x CRT	0.15	0.33	0.09	0.16	0.35*	0.11	0.29
	(0.20)	(0.21)	(0.22)	(0.20)	(0.21)	(0.22)	(0.22)
Repub x Text x CRT	0.18	0.08	−0.02	0.20	0.10	0.01	−0.10
	(0.20)	(0.20)	(0.22)	(0.20)	(0.20)	(0.22)	(0.21)
Repub x Skit x CRT	0.05	−0.01	−0.08	0.05	−0.01	−0.09	−0.21
	(0.23)	(0.23)	(0.25)	(0.23)	(0.23)	(0.25)	(0.25)
Constant	0.38***	0.42***	0.39***	0.16	0.28**	0.16	0.25**
	(0.04)	(0.04)	(0.04)	(0.13)	(0.11)	(0.14)	(0.12)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	3,267	3,267	2,907	3,267	3,267	2,907	2,907
R^2	0.05	0.04	0.05	0.06	0.06	0.06	0.07
Adjusted R^2	0.04	0.04	0.04	0.06	0.05	0.06	0.06

*p < .1; **p < .05; ***p < .01

Notes: PID is pooled to Republican/Not Republican for brevity; reference category for medium is Video.

Table F16: **Interactive Models of Partisan Motivated Reasoning and Target Affect**

	Elizabeth Warren Feeling Thermometer						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Repub PID	−34.09***	−30.24***	−36.46***	−32.22***	−28.84***	−34.87***	−32.25***
	(3.27)	(3.33)	(3.45)	(3.18)	(3.23)	(3.35)	(3.43)
Video	−7.36**	−9.29***	−6.70**	−8.30***	−9.01***	−7.90**	−8.35**
	(3.09)	(3.22)	(3.28)	(3.00)	(3.12)	(3.19)	(3.31)
Audio	−8.39***	−7.86**	−9.03***	−9.40***	−7.41**	−10.50***	−8.18**
	(3.08)	(3.19)	(3.25)	(2.99)	(3.10)	(3.16)	(3.30)
Text	−1.25	0.70	−2.03	−2.37	0.94	−3.81	−1.27
	(2.97)	(3.15)	(3.13)	(2.89)	(3.06)	(3.04)	(3.24)
Skit	−6.52**	−7.53**	−6.10*	−7.48**	−7.43**	−7.23**	−7.52**
	(3.00)	(3.11)	(3.18)	(2.92)	(3.02)	(3.10)	(3.21)
Attack Ad	−4.63	−4.34	−4.77	−5.52*	−4.69	−6.04*	−3.18
	(3.09)	(3.20)	(3.26)	(3.00)	(3.10)	(3.17)	(3.28)
CRT	1.55	−1.07	−0.34	−4.49	−4.25	−7.52	−4.67
	(5.76)	(5.92)	(6.21)	(5.61)	(5.75)	(6.05)	(6.27)
Repub x Video	2.61	3.80	3.54	3.26	3.34	4.66	5.63
	(4.68)	(4.77)	(4.96)	(4.55)	(4.62)	(4.82)	(4.91)
Repub x Audio	4.36	4.19	5.29	4.22	3.41	6.06	4.39
	(4.66)	(4.78)	(4.91)	(4.52)	(4.63)	(4.76)	(4.91)
Repub x Text	−3.19	−7.09	−1.35	−1.87	−5.38	1.02	−1.15
	(4.56)	(4.68)	(4.75)	(4.43)	(4.54)	(4.62)	(4.76)
Repub x Skit	−0.85	−0.01	−0.15	−0.33	−0.54	0.80	2.08
	(4.53)	(4.64)	(4.76)	(4.40)	(4.50)	(4.63)	(4.74)
Repub x Ad	0.60	−3.82	1.50	2.39	−1.34	3.94	−2.10
	(4.62)	(4.62)	(4.86)	(4.49)	(4.48)	(4.72)	(4.71)
CRT x Repub	−15.10*	−12.87	−9.78	−13.87	−12.43	−6.38	−4.96
	(8.79)	(8.79)	(9.37)	(8.54)	(8.52)	(9.10)	(9.18)
Video x CRT	7.09	15.92*	7.56	9.14	13.80*	10.61	13.09
	(8.08)	(8.33)	(8.64)	(7.85)	(8.08)	(8.40)	(8.62)
Audio x CRT	12.26	6.28	17.75**	12.80*	3.46	19.54**	8.29
	(7.89)	(8.31)	(8.49)	(7.67)	(8.06)	(8.25)	(8.74)
Text x CRT	−6.50	−5.18	−3.59	−5.09	−6.13	−0.62	−2.96
	(7.59)	(8.04)	(8.11)	(7.38)	(7.81)	(7.89)	(8.42)
Skit x CRT	5.74	5.84	5.88	7.30	5.11	8.03	2.13
	(7.88)	(8.05)	(8.51)	(7.66)	(7.81)	(8.27)	(8.54)
Ad x CRT	2.00	5.36	4.59	3.77	7.82	7.27	3.00
	(7.88)	(8.11)	(8.38)	(7.65)	(7.87)	(8.14)	(8.38)
Repub x Video x CRT	2.34	−7.90	0.16	3.15	−4.27	−0.97	−9.96
	(12.40)	(12.44)	(13.31)	(12.04)	(12.05)	(12.94)	(12.93)
Repub x Audio x CRT	0.13	0.70	−6.16	1.88	2.63	−7.65	−3.91
	(12.16)	(12.56)	(13.00)	(11.80)	(12.16)	(12.62)	(13.02)
Repub x Text x CRT	14.77	15.01	9.28	12.82	11.05	4.38	0.97
	(12.15)	(12.26)	(12.82)	(11.79)	(11.90)	(12.44)	(12.64)
Repub x Skit x CRT	13.11	13.30	10.68	13.70	17.12	10.29	12.97
	(12.14)	(12.26)	(12.94)	(11.79)	(11.89)	(12.58)	(12.75)
Repub x Ad x CRT	−4.18	−4.13	−9.36	−6.29	−10.21	−12.46	−9.09
	(12.30)	(12.28)	(13.05)	(11.94)	(11.90)	(12.67)	(12.63)
Constant	61.52***	60.20***	61.27***	60.45***	59.53***	63.32***	59.80***
	(2.20)	(2.31)	(2.34)	(6.02)	(5.36)	(6.37)	(5.72)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	5,524	5,524	4,895	5,523	5,523	4,894	4,894
R^2	0.28	0.25	0.30	0.33	0.30	0.34	0.31
Adjusted R^2	0.28	0.25	0.29	0.32	0.30	0.34	0.31

*p < .1; **p < .05; ***p < .01

Notes: PID is pooled to Republican/Not Republican for brevity; reference category for medium is Video.

Table F17: **Interactive Models of Ambivalent Sexism and Belief in Clipping**

	Extent of belief that clipping was not fake or doctored [1-5]						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ambivalent Sexism	0.24***	0.15***	0.24***	0.14***	0.07	0.15***	0.09
	(0.05)	(0.06)	(0.06)	(0.05)	(0.06)	(0.06)	(0.06)
Audio	0.50**	0.23	0.47**	0.56***	0.29	0.55**	0.33
	(0.22)	(0.23)	(0.23)	(0.21)	(0.22)	(0.23)	(0.24)
Text	0.14	−0.11	0.15	0.26	−0.03	0.27	−0.05
	(0.22)	(0.23)	(0.23)	(0.22)	(0.23)	(0.23)	(0.24)
Skit	−1.01***	−0.83***	−1.00***	−0.93***	−0.72***	−0.92***	−0.81***
	(0.25)	(0.25)	(0.27)	(0.25)	(0.25)	(0.27)	(0.27)
A.S. x Audio	−0.13*	−0.05	−0.12	−0.14**	−0.06	−0.13*	−0.06
	(0.07)	(0.08)	(0.08)	(0.07)	(0.07)	(0.08)	(0.08)
A.S. x Text	−0.02	0.04	−0.01	−0.05	0.01	−0.04	0.04
	(0.07)	(0.08)	(0.08)	(0.07)	(0.08)	(0.08)	(0.08)
A.S. x Skit	0.13	0.09	0.12	0.11	0.06	0.11	0.10
	(0.09)	(0.09)	(0.09)	(0.08)	(0.08)	(0.09)	(0.09)
Constant	2.55***	2.86***	2.54***	2.73***	3.21***	2.77***	3.10***
	(0.16)	(0.17)	(0.17)	(0.37)	(0.32)	(0.40)	(0.35)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	3,267	3,267	2,907	3,267	3,267	2,907	2,907
R^2	0.06	0.04	0.07	0.10	0.08	0.10	0.09
Adjusted R^2	0.06	0.04	0.06	0.09	0.07	0.10	0.08

*p < .1; **p < .05; ***p < .01

Notes: Reference category for medium is Video.

Table F18: **Interactive Models of Ambivalent Sexism and Binarized Belief in Clipping**

	Belief that clipping was not fake or doctored [y/n]						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ambivalent Sexism	0.06***	0.04**	0.06***	0.02	0.01	0.03	0.03
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Audio	0.08	0.08	0.12	0.09	0.08	0.12	0.15*
	(0.08)	(0.08)	(0.09)	(0.08)	(0.08)	(0.08)	(0.09)
Text	−0.03	−0.07	−0.02	−0.01	−0.06	0.01	−0.05
	(0.08)	(0.08)	(0.09)	(0.08)	(0.08)	(0.09)	(0.09)
Skit	−0.14	−0.07	−0.11	−0.14	−0.06	−0.11	−0.04
	(0.09)	(0.09)	(0.10)	(0.09)	(0.09)	(0.10)	(0.10)
A.S. x Audio	−0.02	−0.02	−0.03	−0.02	−0.01	−0.03	−0.03
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
A.S. x Text	0.01	0.02	0.01	0.01	0.02	0.01	0.02
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
A.S. x Skit	−0.004	−0.02	−0.01	−0.002	−0.02	−0.01	−0.02
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Constant	0.25***	0.30***	0.24***	0.12	0.23**	0.11	0.18
	(0.06)	(0.06)	(0.06)	(0.14)	(0.12)	(0.15)	(0.13)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	3,267	3,267	2,907	3,267	3,267	2,907	2,907
R^2	0.02	0.02	0.03	0.06	0.06	0.06	0.06
Adjusted R^2	0.02	0.01	0.02	0.05	0.05	0.06	0.06

*p < .1; **p < .05; ***p < .01

Notes: Reference category for medium is Video.

Table F19: **Interactive Models of Ambivalent Sexism and Target Affect**

	Elizabeth Warren Feeling Thermometer						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ambivalent Sexism	−12.49***	−10.44***	−13.12***	−5.36***	−5.27***	−5.62***	−5.79***
	(1.30)	(1.29)	(1.43)	(1.09)	(1.10)	(1.19)	(1.20)
Video	−5.95	−7.72	−2.15	−4.53	−8.02*	−2.55	−8.06
	(5.44)	(5.59)	(5.91)	(4.49)	(4.67)	(4.87)	(5.04)
Audio	−9.93*	−8.77	−9.18	−11.34***	−10.08**	−11.08**	−11.23**
	(5.30)	(5.42)	(5.74)	(4.39)	(4.54)	(4.74)	(4.91)
Text	−2.90	0.65	0.18	−7.71*	−3.74	−5.41	−4.17
	(5.34)	(5.44)	(5.75)	(4.41)	(4.55)	(4.74)	(4.86)
Skit	−3.46	−10.15*	−2.24	−6.04	−14.21***	−5.19	−16.91***
	(5.32)	(5.40)	(5.76)	(4.40)	(4.52)	(4.75)	(4.85)
A.S. x Video	0.66	1.30	−0.20	0.35	1.35	−0.07	1.58
	(1.84)	(1.87)	(2.02)	(1.52)	(1.57)	(1.66)	(1.70)
A.S. x Audio	2.88	2.10	2.95	2.81*	1.93	2.93*	2.51
	(1.81)	(1.81)	(1.97)	(1.49)	(1.51)	(1.62)	(1.64)
A.S. x Text	0.39	−0.49	−0.59	1.82	0.89	1.08	0.90
	(1.83)	(1.82)	(1.97)	(1.51)	(1.53)	(1.63)	(1.63)
A.S. x Skit	0.25	2.42	0.07	0.93	3.73**	0.77	4.53***
	(1.82)	(1.81)	(1.97)	(1.50)	(1.52)	(1.62)	(1.63)
Constant	80.89***	74.99***	81.06***	74.14***	74.52***	73.35***	72.52***
	(3.80)	(3.85)	(4.18)	(6.55)	(5.85)	(7.00)	(6.30)
Weighted?		✓			✓		✓
Low-Quality Dropped?			✓			✓	✓
Controls?				✓	✓	✓	✓
N	4,599	4,599	4,069	4,599	4,599	4,069	4,069
R^2	0.08	0.06	0.10	0.38	0.35	0.39	0.36
Adjusted R^2	0.08	0.06	0.09	0.37	0.34	0.39	0.36

*p < .1; **p < .05; ***p < .01

Notes: Reference category for medium is Video.

Table F20: **Predictors of Second-Stage Detection Accuracy**

	Detection Accuracy (% Correctly Classified)					
	(1)	(2)	(3)	(4)	(5)	(6)
Digital Literacy		0.25***	0.22***	0.22***	0.20***	0.21***
		(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Accuracy Prompt	−0.002		−0.01	−0.004	−0.005	−0.002
	(0.01)		(0.01)	(0.01)	(0.01)	(0.01)
Stage 1 Debrief			0.01*	0.01*	0.01*	0.01*
			(0.01)	(0.01)	(0.01)	(0.01)
Stage 1 Info Provided			−0.01	−0.001	−0.01	−0.004
			(0.01)	(0.01)	(0.01)	(0.01)
Political Knowledge			0.18***	0.19***	0.18***	0.20***
			(0.01)	(0.01)	(0.01)	(0.01)
Internet Usage			−0.002	−0.01	0.0001	−0.004
			(0.005)	(0.005)	(0.01)	(0.01)
Low-fake Env.			0.03***	0.04***	0.03***	0.05***
			(0.01)	(0.01)	(0.01)	(0.01)
No-fake Env.			0.04***	0.04***	0.04***	0.05***
			(0.01)	(0.01)	(0.01)	(0.01)
Age 65+			0.01	0.01	0.01	0.003
			(0.01)	(0.01)	(0.01)	(0.01)
High School			0.01	0.01	0.03	0.02
			(0.03)	(0.02)	(0.03)	(0.02)
College			0.02	0.02	0.04	0.03
			(0.03)	(0.02)	(0.03)	(0.02)
Postgrad			−0.01	−0.02	0.01	−0.01
			(0.03)	(0.02)	(0.03)	(0.02)
Republican			0.06***	0.07***	0.07***	0.08***
			(0.02)	(0.02)	(0.02)	(0.02)
CRT			−0.06**	−0.06**	−0.06**	−0.06**
			(0.03)	(0.03)	(0.03)	(0.03)
Republican x CRT			0.09***	0.07***	0.09***	0.08***
			(0.01)	(0.01)	(0.01)	(0.01)
Ambivalent Sexism			0.001	−0.002	0.001	−0.0001
			(0.004)	(0.004)	(0.004)	(0.004)
Constant	0.57***	0.36***	0.16***	0.20***	0.14***	0.16***
	(0.005)	(0.02)	(0.05)	(0.04)	(0.05)	(0.05)
Weighted?				✓		✓
Low-Quality Dropped?					✓	✓
N	5,497	5,497	5,496	5,496	4,870	4,870
R^2	0.0000	0.02	0.09	0.09	0.09	0.10
Adjusted R^2	−0.0002	0.02	0.09	0.09	0.09	0.10

*p < .1; **p < .05; ***p < .01

Notes: Reference category for environment is High-fake. PID pooled for brevity.

Table F21: **Predictors of Second-Stage False Positive Rate (FPR)**

	Detection FPR (% Real Videos Classified as Deepfakes)					
	(1)	(2)	(3)	(4)	(5)	(6)
Digital Literacy		-0.11^{***}	-0.08^{***}	-0.11^{***}	-0.06^{**}	-0.08^{***}
		(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Accuracy Prompt	-0.01		0.004	-0.01	0.003	-0.01
	(0.01)		(0.01)	(0.01)	(0.01)	(0.01)
Stage 1 Debrief			-0.03^{***}	-0.03^{***}	-0.03^{***}	-0.03^{***}
			(0.01)	(0.01)	(0.01)	(0.01)
Stage 1 Info Provided			0.01	0.001	0.01	0.004
			(0.01)	(0.01)	(0.01)	(0.01)
Political Knowledge			-0.13^{***}	-0.13^{***}	-0.13^{***}	-0.14^{***}
			(0.02)	(0.02)	(0.02)	(0.02)
Internet Usage			-0.002	0.002	-0.002	0.001
			(0.01)	(0.01)	(0.01)	(0.01)
Low-fake Env.			0.03^{***}	0.02^{***}	0.02^{***}	0.01
			(0.01)	(0.01)	(0.01)	(0.01)
No-fake Env.			0.23^{***}	0.23^{***}	0.22^{***}	0.21^{***}
			(0.01)	(0.01)	(0.01)	(0.01)
Age 65+			0.003	-0.003	0.01	0.0000
			(0.01)	(0.01)	(0.01)	(0.01)
High School			0.001	-0.003	-0.02	-0.02
			(0.03)	(0.02)	(0.04)	(0.02)
College			0.01	0.01	-0.02	-0.01
			(0.03)	(0.02)	(0.04)	(0.02)
Postgrad			0.03	0.02	0.001	0.002
			(0.04)	(0.02)	(0.04)	(0.02)
Republican			-0.06^{***}	-0.08^{***}	-0.07^{***}	-0.08^{***}
			(0.02)	(0.02)	(0.02)	(0.02)
CRT			0.03	0.03	0.04	0.03
			(0.03)	(0.03)	(0.03)	(0.03)
Republican x CRT			-0.07^{***}	-0.07^{***}	-0.08^{***}	-0.07^{***}
			(0.01)	(0.01)	(0.01)	(0.01)
Ambivalent Sexism			0.0005	0.002	-0.002	0.001
			(0.005)	(0.005)	(0.005)	(0.005)
Constant	0.28^{***}	0.37^{***}	0.42^{***}	0.44^{***}	0.42^{***}	0.44^{***}
	(0.01)	(0.02)	(0.06)	(0.05)	(0.06)	(0.05)
Weighted?				✓		✓
Low-Quality Dropped?					✓	✓
N	5,495	5,495	5,494	5,494	4,869	4,869
R^2	0.0002	0.003	0.16	0.16	0.16	0.16
Adjusted R^2	-0.0000	0.003	0.15	0.16	0.15	0.16

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

Notes: Reference category for environment is High-fake. PID pooled for brevity.

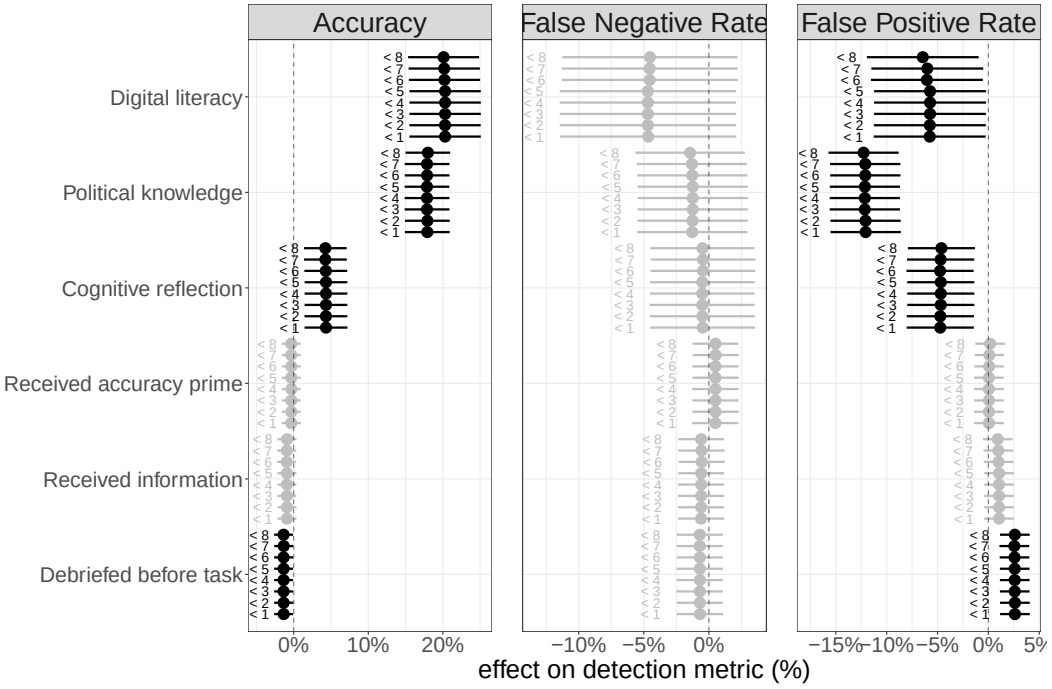
Table F22: **Predictors of Second-Stage False Negative Rate (FNR)**

	Detection FNR (% Deepfakes Classified as Real Videos)					
	(1)	(2)	(3)	(4)	(5)	(6)
Digital Literacy		−0.03	−0.04	−0.01	−0.05	−0.02
		(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
Accuracy Prompt	−0.01		−0.005	−0.01	0.003	−0.003
	(0.01)		(0.01)	(0.01)	(0.01)	(0.01)
Stage 1 Debrief			0.01	0.01	0.01	0.01
			(0.01)	(0.01)	(0.01)	(0.01)
Stage 1 Info Provided			−0.002	0.003	−0.01	−0.003
			(0.01)	(0.01)	(0.01)	(0.01)
Political Knowledge			−0.0004	−0.03	−0.003	−0.03
			(0.02)	(0.02)	(0.02)	(0.02)
Internet Usage			0.02***	0.01	0.02***	0.02**
			(0.01)	(0.01)	(0.01)	(0.01)
Low-fake Env.			0.01	0.001	0.01	−0.004
			(0.01)	(0.01)	(0.01)	(0.01)
No-fake Env.			0.01	0.02	0.02*	0.02*
			(0.01)	(0.01)	(0.01)	(0.01)
Age 65+			0.01	0.02	0.01	0.01
			(0.04)	(0.03)	(0.04)	(0.03)
High School			0.03	0.04	0.03	0.03
			(0.04)	(0.03)	(0.04)	(0.03)
College			0.08*	0.13***	0.08*	0.11***
			(0.04)	(0.03)	(0.05)	(0.03)
Postgrad			−0.05**	−0.04	−0.06**	−0.04
			(0.02)	(0.03)	(0.03)	(0.03)
Republican			0.09**	0.08*	0.11***	0.10**
			(0.04)	(0.04)	(0.04)	(0.04)
CRT			0.02***	0.02***	0.01**	0.01
			(0.01)	(0.01)	(0.01)	(0.01)
Republican x CRT			−0.04***	−0.03**	−0.05***	−0.04**
			(0.01)	(0.02)	(0.02)	(0.02)
Ambivalent Sexism	0.34***	0.36***	0.16**	0.21***	0.18**	0.23***
	(0.01)	(0.03)	(0.07)	(0.06)	(0.07)	(0.06)
Weighted?				✓		✓
Low-Quality Dropped?					✓	✓
N	3,690	3,690	3,690	3,690	3,266	3,266
R^2	0.0002	0.0003	0.02	0.03	0.02	0.02
Adjusted R^2	−0.0000	0.0000	0.02	0.02	0.01	0.01

*p < .1; **p < .05; ***p < .01

Notes: Reference category for environment is High-fake. PID pooled for brevity.

Figure F23: **Sensitivity of Predictors of Detection Experiment Performance to Non-Response Thresholding**

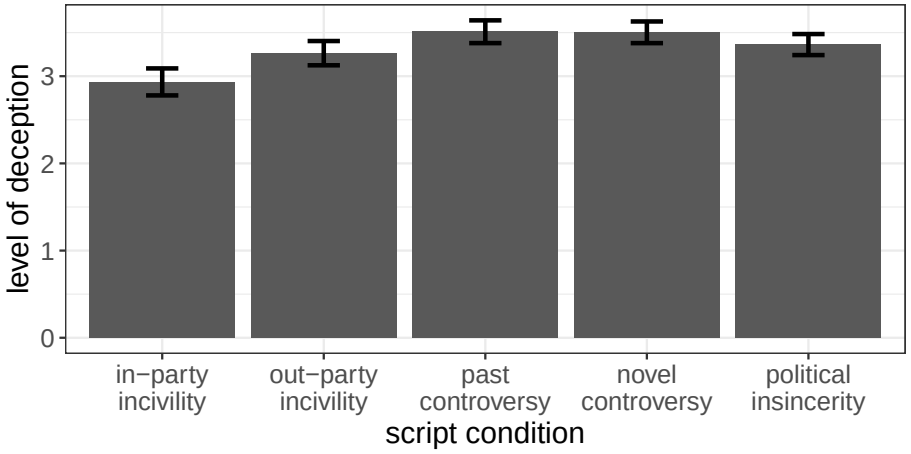


Notes: Each estimate is the effect of the corresponding predictor estimated from a model with full controls (see tables for detection-stage results) excluding respondents with $< x$ number of videos completed in the detection task.

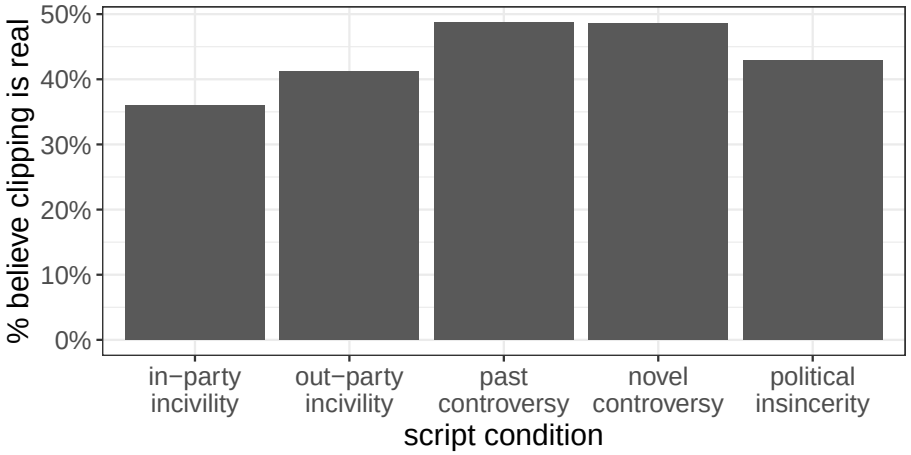
G Exploratory Analyses

Figure G24: **Exposure-Stage Heterogeneity in Deception by Scandal Script**

(a) **Categorical Response**



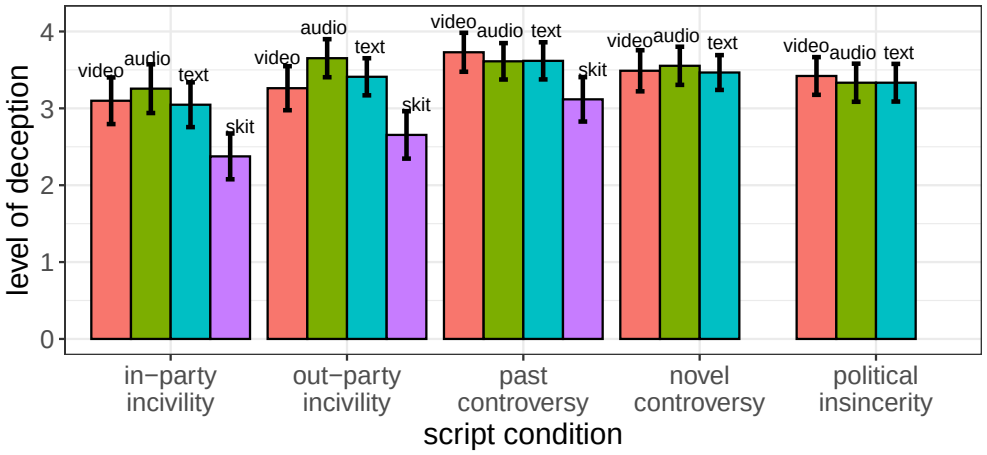
(b) **Binarized Response**



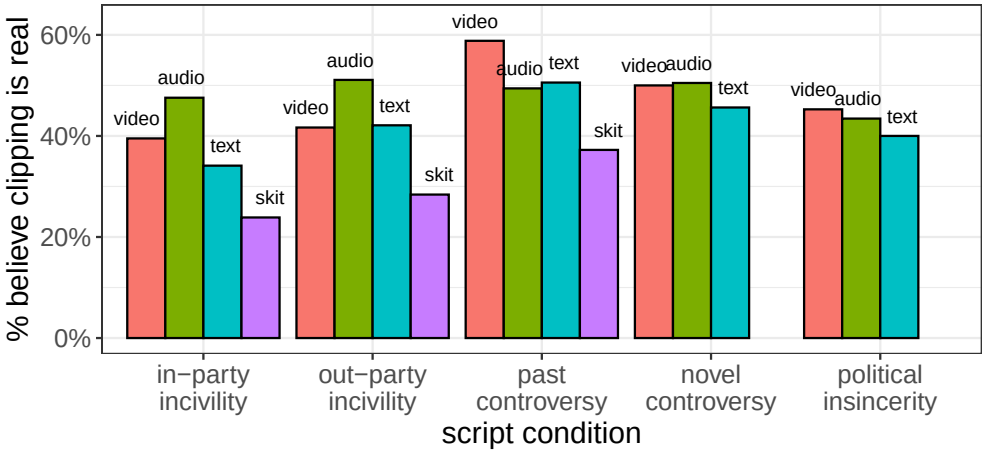
Notes: Results from the subset of respondents exposed to a scandal stimuli, not assigned an information treatment, and who provided a response to our deception question (n=1848).

Figure G25: **Exposure-Stage Heterogeneity in Deception by Scandal Script and Medium**

(a) **Categorical Response**

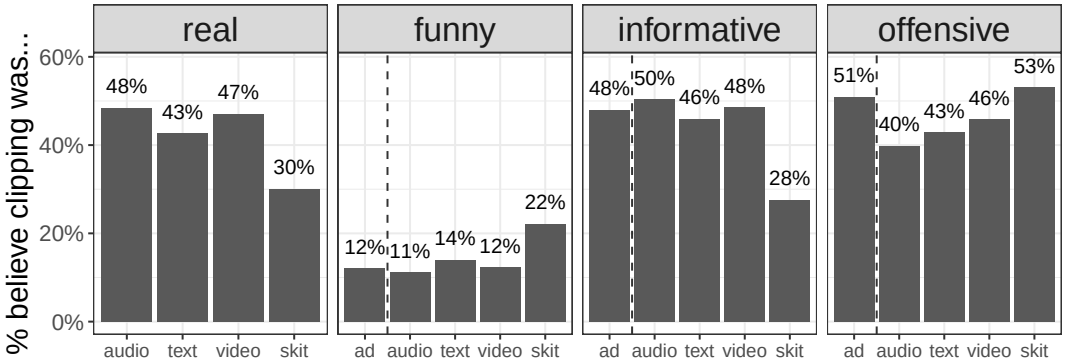


(b) **Binarized Response**



Notes: Results from the subset of respondents exposed to a scandal stimuli, not assigned an information treatment, and who provided a response to our deception question (n=1,848). To reduce the number of experimental cells, only three of five scripts were used as skit stimuli.

Figure G26: **Other Affective Responses to First-Stage Clip**

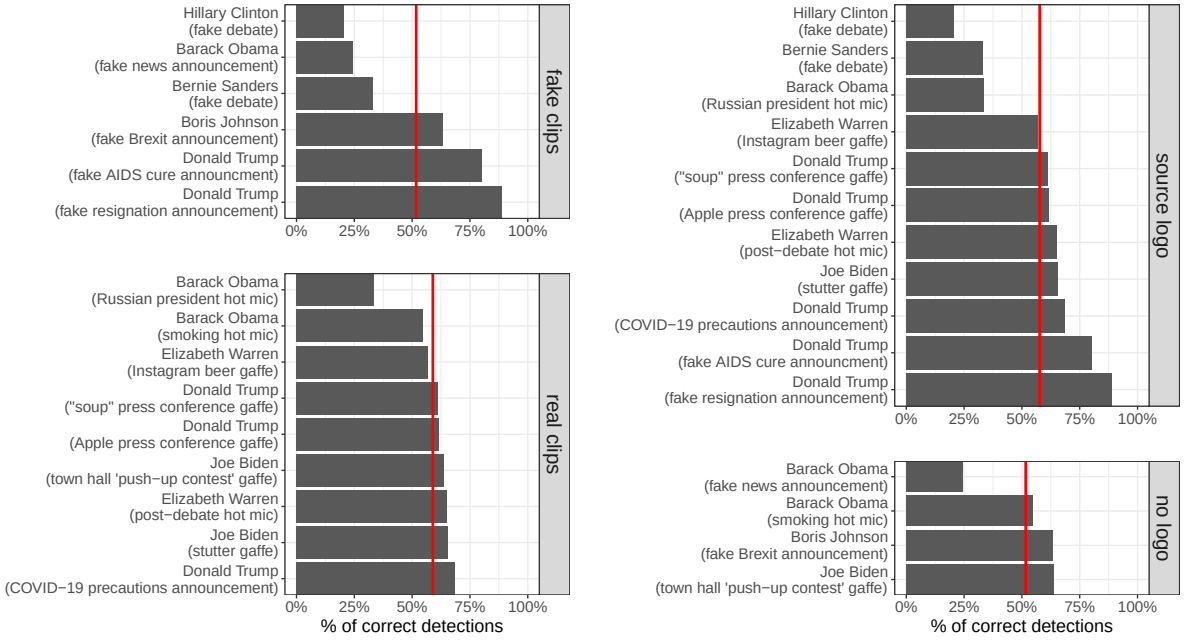


Notes: Responses evaluating whether clip was "funny," "informative," or "offensive" were solicited alongside belief that clip was not fake or doctored. The attack ad condition excluded since it is not a directly comparable clip of the scandal.

Figure G27: **Detection-Stage Performance for Specific Clips**

(a) **By Authenticity**

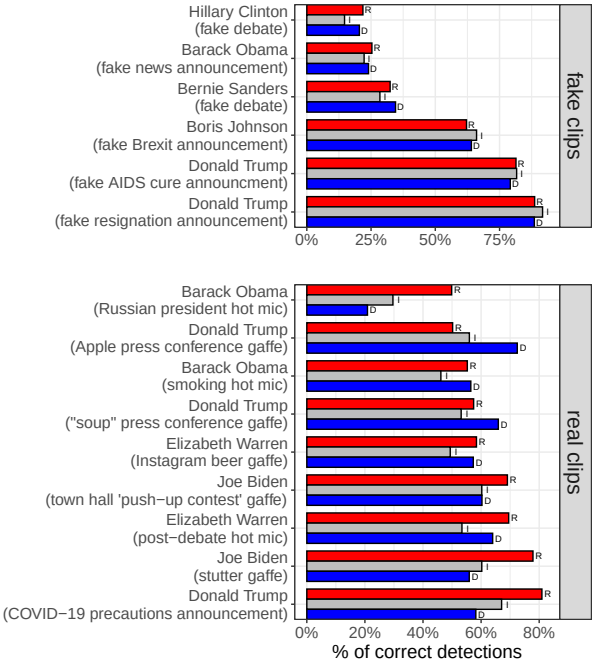
(b) **By Source Cue**



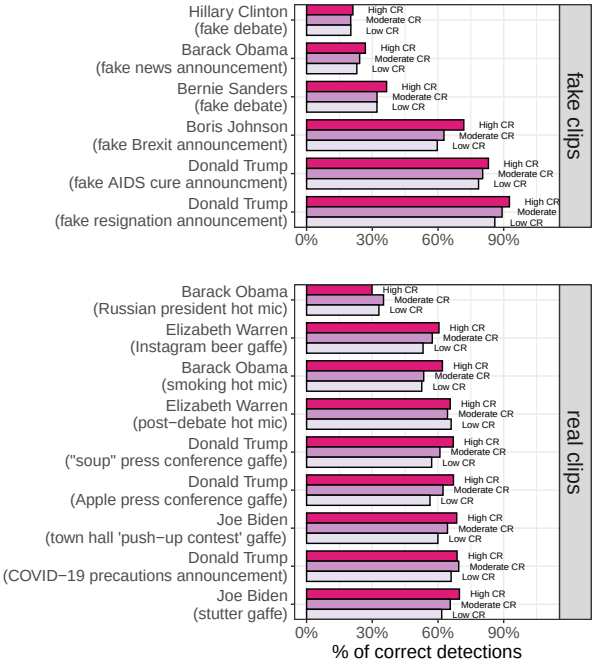
Notes: Results are for n =5,497 (99%) of respondents who provide a response to at least one video in the detection experiment. Fake clips are detected less well than real clips, but this difference (Δ) is not significant according to a t-test ($\Delta = -7.20\%, t = 0.57, p = 0.58$). Clips without source outlet logos are detected less well than clips with source logos, but this difference is also not significant ($\Delta = -6.03\%, t = 0.53, p = 0.61$).

Figure G28: **Detection-Stage Performance for Specific Clips by Subgroup**

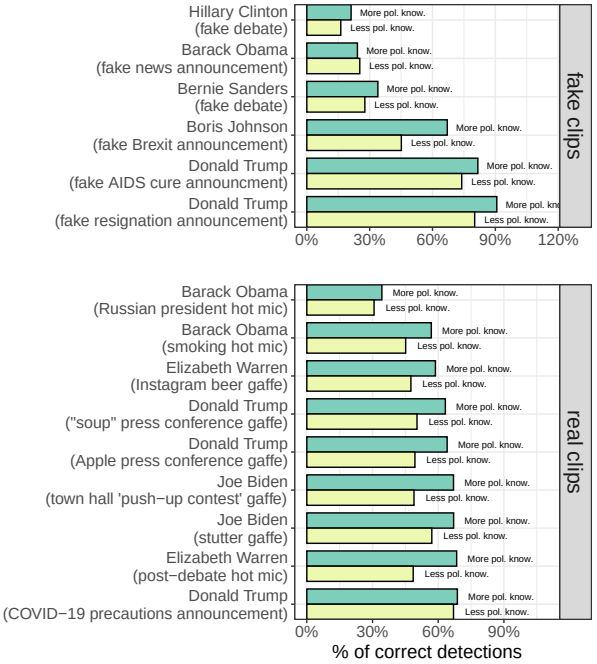
(a) **By Partisanship**



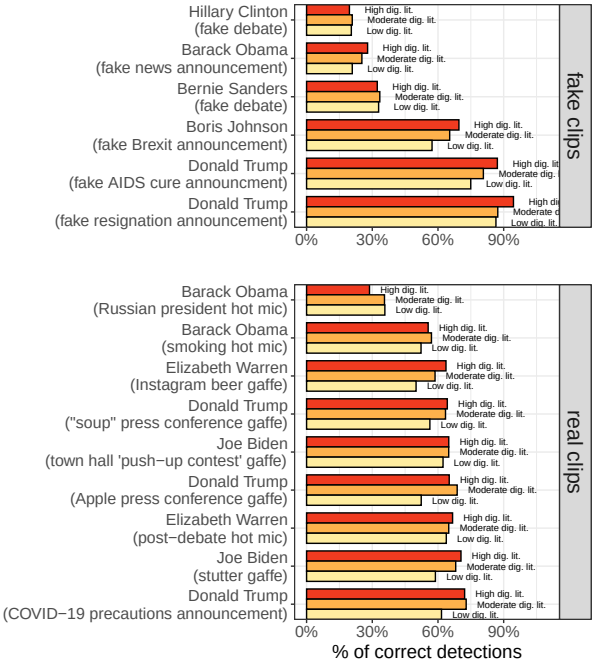
(b) **By Cognitive Reflection**



(c) **By Political Knowledge**



(d) **By Digital Literacy**



Notes: Results are for n =5,497 (99%) of respondents who provide a response to at least one video in the detection experiment. Cognitive reflection and digital literacy categories constructed as equal-sized quartiles.

Figure G29: **Relationship Between Belief in Authenticity of Video Clip and Confidence in Event**

(a) **Barack Obama (Russian president hot mic)**

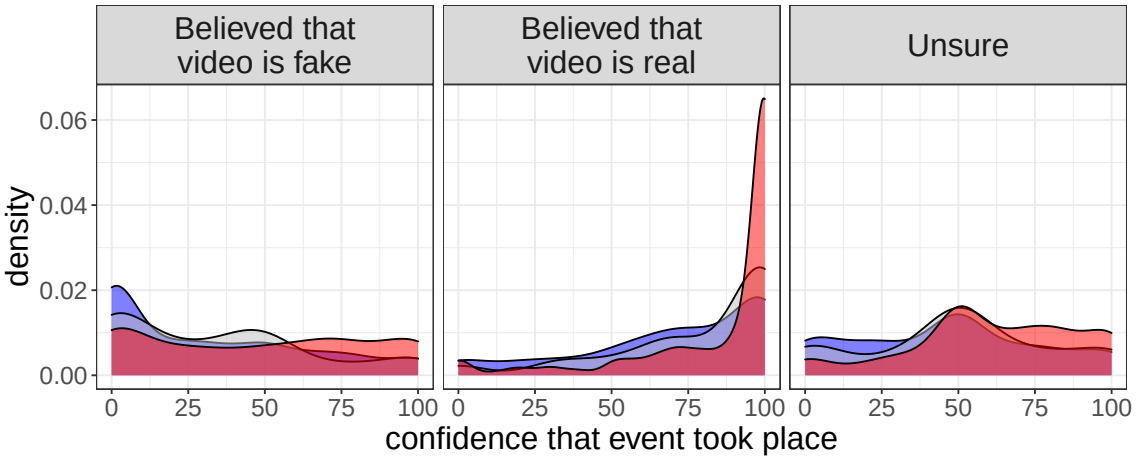
(b) **Doanld Trump (COVID-19 precautions announcement)**

(c) **Donald Trump (fake resignation announcement)**

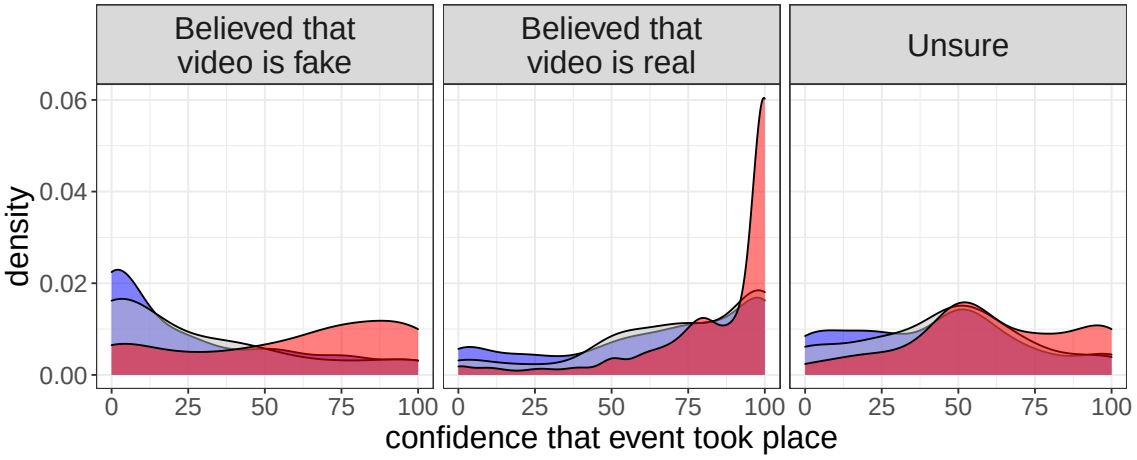
Notes: Clips (a) and (b) are real videos, clip (c) is a deepfake. Results are for n =5,497 (99%) of respondents who provide a response to at least one video in the detection experiment. A variety of regression specifications estimate large, robust and statistically significant positive relationship between a respondent's belief in the video's authenticity and confidence in the depited event's occurence.

Figure G30: **Relationship Between Belief in Authenticity of Video Clip and Confidence in Event by Partisanship**

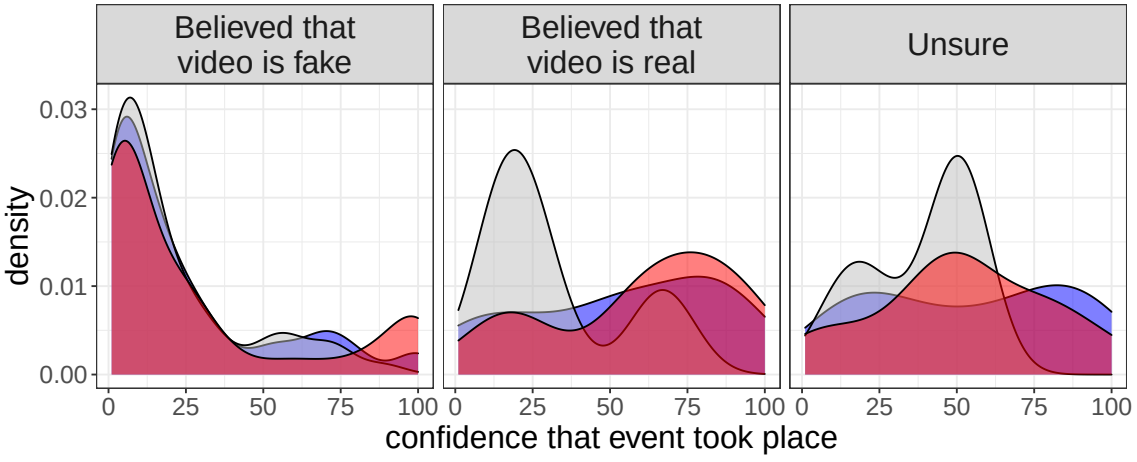
(a) **Barack Obama (Russian president hot mic)**



(b) **Doanld Trump (COVID-19 precautions announcement)**



(c) **Donald Trump (fake resignation announcement)**



Notes: Clips (a) and (b) are real videos, clip (c) is a deepfake. Results are for n =5,497 (99%) of respondents who provide a response to at least one video in the detection experiment.