The Negative Consequences of Informing Voters about Deepfakes: Evidence from Two Survey Experiments

John Ternovski, Joshua Kalla, P. M. Aronow

Abstract. Advances in machine learning have made possible "deepfakes," or realistic, computer-generated videos of public figures saying something they have not actually said. Policymakers have expressed concern that deepfakes could mislead voters, but prior research has found that such videos have minimal effects. There has nevertheless been extensive media coverage of the dangers of deepfakes, urging voters to be critical consumers of political videos. We explore whether these well-intentioned messages have an unintended consequence: if voters are warned about deepfakes, they may begin to distrust *all* political videos. We conducted two online survey experiments, and found that informing participants about deepfakes did not enhance participants' ability to successfully spot manipulated videos but consistently induced them to believe the videos they watched were fake, even when they were real. Our findings suggest that even if deepfakes are not themselves persuasive, information about deepfakes can nevertheless be weaponized to dismiss real political videos.

1 Introduction

Breakthroughs in machine learning have led to the development of software that can seamlessly fabricate videos of any individual. Computer-generated videos, so-called "deepfakes," can be made in which a politician appears to say something they never actually said.[1] Computer and social scientists have raised concerns that deepfakes may mislead voters and sway election outcomes (e.g., Dack 2019). Policymakers have echoed these concerns. For example, during a hearing of the U.S. House of Representatives Intelligence Committee, Adam Schiff, the committee's chair, noted that deepfakes allow "malicious actors to foment chaos, division or crisis," and that such videos "have the capacity to disrupt entire campaigns, including that for the presidency" (O'Sullivan 2019, p. 1). Since this hearing, Congress has passed two laws (S.2904 2020; H.R.6395 2021) that explicitly directed "the Department of Homeland Security (DHS), the Department of Defense (DOD), and the National Science Foundation (NSF) to issue reports on and bolster research into deepfakes... These bills ask for recommendations that could lay

1. For a sociological overview of the development of deepfakes, see Paris and Donovan (2019).

10.54501/jots.v1i2.28

the predicate for federal regulations of such media" (Ferraro 2020, p. 1). However, recent randomized experiments on the impacts of deepfakes in American politics have found no evidence that people believe the content of the manipulated videos (Wittenberg et al. 2021; Vaccari and Chadwick 2020). Barari, Lucas, and Munger (2021) found that deepfakes were persuasive, but that their effects were comparable to those of textual misinformation.[2]

Despite this evidence from social scientists, news coverage of deepfakes continues to be extensive and predominantly emphasizes their potential threat (e.g., Gosse and Burkell 2020; Yadlin-Segal and Oppenheim 2021). A cursory search of the five most popular news websites in the U.S. YouGov (according to 2021) for the term "deepfake"[3] on the news aggregator Google News returned 11,700 news articles discussing deepfakes, 62.5% of which use cautionary language ("threat," "worried," "danger," "warn," "risk").[4] Attempts to inform the public of the dangers of deepfakes even led to the creation of a widely viewed deepfake of former President Barack Obama, in which the comedian Jordan Peele partnered with BuzzFeed Video to create a deepfake of himself impersonating Obama as a warning to Americans (Castillo 2018). This public service announcement has accrued over 8.4 million views on YouTube. But do these well-intentioned attempts to inform and educate the public have an unintended consequence? Does information about the existence and dangers of deepfakes cause voters to distrust all political video footage—whether real or fake?

While this question has not been answered in the context of deepfakes, a robust literature on textual misinformation (for a review, see Lazer et al. (2018)) has found that elite discourse about "fake news" may lower trust in the media and prime participants to disbelieve the veracity of real news (Van Duyn and Collier 2019). Official warnings about fake news similarly induce participants to disbelieve true headlines (Clayton et al. 2020; Pennycook, Bear, et al. 2020). There are already high-profile cases of American voters alleging that real political videos are deepfakes. For example, in January 2021, supporters of Donald Trump suggested that a video Trump shared via Twitter, in which he conceded the 2020 election, was a deepfake (Villarreal 2021).

It isn't only highly motivated partisans who disbelieve real video footage. In a recent high-profile case, U.S. law enforcement officials erroneously alleged that real video footage was deepfaked. A Pennsylvania woman was accused of making a deepfake of high school cheerleaders vaping "to try to get them kicked off the squad" (AP 2021). But upon closer examination, video forensic experts found no evidence that the video was manipulated (Harwell 2021). "When pressed on how police made their determination that the footage had been manipulated... [the police officer who made the arrest said] that he had relied on his 'naked eye'" (Thalen 2021, p. 1).

This paper explores whether information warning about the existence of deepfakes makes American voters more likely to disbelieve real political videos. Or do these efforts to inform voters about the dangers of deepfakes work as intended—leading to the more critical consumption of political videos? If the former is true, as media coverage of deepfakes increases, Americans' trust in political videos may continue to erode. Another

2. A recent deepfake study in the Netherlands found modest persuasive effects among a subset of participants (Dobber et al. 2021).

3. The search was conducted on March 30, 2021 with the input: deepfake AND (site:news.yahoo.com OR site:nbc.com OR site:cbs.com OR site:nbcnews.com OR site:cnn.com).

4. The search was conducted on March 30, 2021 with the input: deepfake AND ("threat*" OR "worried" OR "danger*" OR "warn*" OR "risk*") AND (site:news.yahoo.com OR site:nbc.com OR site:cbs.com OR site:nbcnews.com OR site:cnn.com). We qualitatively assessed the results of these searches. While there were, unsurprisingly, many false positives for both search queries, we found that, in concordance with Gosse and Burkell (2020) and Yadlin-Segal and Oppenheim (2021), the overwhelming majority of articles that specifically addressed deepfakes focused on their potential danger or, at best, described them as "creepy."

danger is that politicians could use factually true statements (e.g., "deepfakes exist") to subtly disavow and dismiss video recordings of their past statements and behavior. Such outcomes could undermine political accountability; thus it is imperative to understand the social impacts of information about deepfakes before political deepfakes become commonplace.

Across two online survey experiments, we demonstrate that informing voters about deepfakes increases disbelief in both real and manipulated videos without improving participants' ability to successfully identify deepfakes. Study 1 used an actor posing as a politician sharing an extreme policy position. Using a factorial design, participants were randomized in a first factor to see either a real video of the politician or a deepfake version of the video and, in a second factor, to receive information about deepfakes or not. Study 2 leveraged Americans' low levels of policy knowledge (e.g., Barabas et al. 2014) to show a real video of a real politician making a policy statement that is atypical of his party and is not widely known (and thus might reasonably be thought to be a deepfake). Both studies measured belief in the content of the videos and trust in video as a source of political information. We found that participants were unable to discriminate between real and deepfaked videos even when they were informed about the existence of deepfakes. Information about deepfakes instead induced participants to disbelieve any political video they were shown as part of the experiment—real or fabricated. In other words, a general statement about the dangers of deepfakes, as one might see in a headline from a trusted news source, encouraged participants to disbelieve real video clips of politicians making policy statements. The effects were large and consistent across both studies. Information about deepfakes even affected what policy stances participants associated with a real US politician. Thus providing information about the dangers of deepfakes not only made participants suspect that real videos of politicians speaking were fake, but even affected how they internalized the content of the videos.

This paper is organized as follows. First, we briefly overview the data and designs of both survey experiments. Next, we report our main findings and discuss the implications of the results. We conclude with broader policy implications and suggest avenues for further research.

2 Experimental Design

We ran two pre-registered[5] online survey experiments on Lucid Theorem[6] using convenience samples[7] in the spring and fall of 2020. The Appendix contains the full survey questionnaires.

2.1 Study 1: Fictional Politician

We created a video of an actor playing a fictional politician advocating an extreme policy position: support for a law requiring doctors to use essential oils to treat cancer before

5. See https://osf.io/rqfz5/?view_only=e2807b367a534262bb6c7aeb5727b999 for our pre-analysis plans.

6. Lucid is an increasingly popular alternative to Amazon Mechanical Turk for social science survey research. Many well-known findings have been replicated on Lucid, suggesting the platform is capable of providing high-quality data (Coppock and McClellan 2019; Peyton, Huber, and Coppock 2021). During the COVID-19 pandemic, Ternovski and Orr (2022) found that Lucid data can provide reliable data when researchers screen on attentiveness, which we do here.

7. Treatment effects from online convenience samples have been shown to generalize to nationally representative samples (e.g., Mullinix et al. 2015).

attempting treatments with conventional medicine.[8] (Figure 1 contains a screenshot of the video.) The actor also recorded another video that we used as training footage for our deepfake, in which he recited a generic political speech.[9] This video of a generic speech was then broken down into still images, which were used as training data in a SAEHD (High Definition Styled AutoEncoder) model (trained over 65,000 iterations). The model encodes the face in the training data and the face in the "destination" video (the video in which our hired actor talks about essential oils in medicine) into a latent space to create a common latent representation. This process makes a face swap in a video clip possible. The resulting model is essentially a moldable mask that is superimposed on the actor in the destination video. The full technical details of the deepfake model are found in the Appendix (Section 1). (For a more detailed overview of commonly used methods of creating deepfakes, please see Barari, Lucas, and Munger (2021).) This process emulates a malicious agent using a look-alike actor as a training set for the deepfake. The real video on essential oils and the deepfake thereof were used as treatments in Study 1.

Figure 1: Schreenshot of the video used in Study 1.

Study 1 began with a series of screening and pre-treatment measures (for details, please see the survey script in the Appendix). Participants were then shown a separate page with the following text: "On the next page, you will watch a brief political speech. You will then be asked about how you feel about the person making the speech and whether or not you would vote for them in the upcoming election." We randomized participants to receive either the real video or the deepfake version and either information about deepfakes or no additional information in a 2x2 fully factorialized experiment. The four conditions are summarized in Table 1 on the next page.

After viewing the video, we measured participants' perceived favorability of (and intention to vote for) the politician, their view on essential oils in medicine, whether they believed the video was real, and their overall confidence in other video footage of politicians speaking. Due to the risk of priming latent beliefs by asking participants if the video they

8. We confirmed that the fictional politician appeared sufficiently non-partisan by asking participants to guess his political party at the end of the survey. The plurality of respondents (30%) said they "didn't know" (or it was "unclear" as to) what party he was affiliated with; 27% thought the politician was a Democrat, 16% thought he was a Republican, and 27% reponded that he was Independent. Respondents' partisanship was weakly associated with their guesses (Cramer's V = 0.08; see Appendix Table OA2 for more details).

9. The speech was composed using excerpts from Swaim and Nussbaum (2016), a FiveThirtyEight article in which Republican and Democratic speechwriters collaborated "to write a totally pandering bipartisan stump speech for an imaginary presidential candidate — one who espouses only positions that a majority of voters agree with" (p. 1).

	No Information (N=696)	**Information about Deepfakes (N=700)**
Real Video (N=696)	**(1)** Participants were shown a real video of the actor stating that, if elected, he would require doctors to try using essential oils to treat cancer before they are allowed to use conventional medicine. *(N=347)*	**(2)** Participants were shown the same video as (1) but with the following message above the clip "WARNING: Computer scientists are increasingly concerned about "DeepFake" videos. With DeepFakes, it's now possible to use a computer to convincingly manipulate videos of people to make them appear to say things they've never said." *(N=349)*
Deepfake Video (N=700)	**(3)** This treatment arm was identical to (1) except participants were shown the deepfake version of the video. *(N=349)*	**(4)** This treatment arm was identical to (2) except participants were shown the deepfake version of the video. *(N=351)*

Table 1: Description of Study 1 treatment arms.

just saw was real,[10] we attempted to measure this outcome unobtrusively by asking whether they were convinced that the politician "believes what is said." We only explicitly ask participants whether they believed the video was deepfaked at the very end of the survey.

We found that survey attrition varied significantly across treatment conditions (Pearson $\chi^2(3) = 11.3$, $p = 0.01$): 4.9% of participants in the No Information + Real Video condition did not finish the survey, compared to 1.8% in the three remaining conditions combined. Because our pre-analysis plan did not explicitly stipulate how we would address differential attrition, we used Manski-type worst-case bounds (Manski 2003), which require only that the support of the outcome is bounded when constructing our confidence intervals.[11] These bounds are considered the gold-standard approach when analyzing experiments with attrition (Gerber and Green 2012). Additional details are provided in Section 2 of the Appendix.

10. Research on textual fake news has found that asking participants about the accuracy of a specific piece of misinformation affected how likely they were to share that misinformation online (Pennycook et al. 2021; Pennycook, McPhetres, et al. 2020). If priming accuracy can affect behavior, it also has the potential to affect latent beliefs.

11. The "best" and "worst" cases used to estimate the confidence interval are also covariate adjusted and use robust standard errors.

2.2 Study 2: Real Politician

One weakness of Study 1 is that the extreme policy stance in the video could plausibly be genuine. Participants were not familiar with the politician (as he is fictional) and may not have had any expectation that the video they were watching could be deepfaked. McDonald (2019) empirically illustrated how people make use of prior knowledge of real politicians in online survey experiments and how studies using solely hypothetical politicians can produce misleading estimates of real-world political behavior. We therefore replicated the impact of providing information about deepfakes using a real politician in Study 2.

We identified a well-known politician who had previously expressed a policy stance that was atypical given his party affiliation. We found 2002 video footage of Republican Mitt Romney asserting, in a Massachusetts gubernatorial debate, that he would protect a woman's right to choose-—an unusual stance for a Republican.[12]

While Study 2's use of a real politician may allay some concerns about generalizability from a fictional politician in Study 1, the findings in Study 2 should not be construed as broadly generalizable to all American politicians. Study 2 tests the information treatment in the context of one American politician, and there may be unexpected heterogeneity across different politicians. In particular, given his criticisms of Donald Trump and support for the Black Lives Matter movement, Romney was not representative of the typical Republican politician. Participants might therefore have responded differently to him than they would have to other Republican politicians. Future studies should replicate these treatment effects with a more diverse pool of prominent politicians.

The design of this study was nearly identical to that of Study 1 with the following two modifications. First, Study 2 had only two conditions, Information about Deepfakes and No Information (see Table 2 on the next page), as we did not create a deepfake of Mitt Romney given ethical considerations. Second, our analysis was restricted to participants who knew Romney was a Republican pre-treatment, as Romney's (formerly) pro-choice position would not appear surprising to participants with no knowledge of his party affiliation. In Study 2, we found no evidence of differential attrition: survey completion rates did not differ significantly across treatment arms.

3 Results and Discussion

News articles about the dangers of deepfakes and public service announcements like Jordan Peele's Obama deepfake are intended to make viewers more critical when consuming political videos. Ideally, a viewer will believe real videos and disregard fake ones. It is possible that the intent of such messages is to make Americans more skeptical of *all* information, but such a goal could have deleterious impacts on democratic functioning. Belief in fake videos may lead to misinformed voters, but disbelief in real videos of politicians discussing their policy positions may lead to uninformed voters (for further analysis of uninformed vs. misinformed voters, see Kuklinski et al. (2000)). High levels of uninformed voters have been linked to serious electoral consequences; for instance, Fowler and Margolis (2014) found that "[a] lack of knowledge on the policy positions of the parties significantly hinders the ability of low-socioeconomic-status citizens to translate their preferences into partisan opinions and vote choices" (p. 100).

As such, our first analysis assesses whether informing participants about the danger of deepfakes affects the rate at which voters disbelieve a deepfaked political video. The primary outcome measure used asked participants the extent to which they agreed with

12. As of 2020, there were only two pro-choice Republicans in the Senate (Sussman 2020).

	No Information (N=966)	**Information about Deepfakes (N=959)**
Real Video (N=1,925)	**(1)** Participants watched a real video of Mitt Romney saying, "And I've been very clear on that, I will preserve and protect a woman's right to choose and I'm devoted and dedicated to honoring my word..."	**(2)** Participants were shown the same video as (1) but with the following message above the clip "WARNING: Computer scientists are increasingly concerned about "DeepFake" videos. With DeepFakes, it's now possible to use a computer to convincingly manipulate videos of people to make them appear to say things they've never said."

Table 2: Description of Study 2 treatment arms.

the statement "This video was doctored, manipulated and/or faked by a computer (i.e. it is a 'Deep Fake')" on a 7-point agree/disagree scale ranging from strongly disagree (-3) to strongly agree (3).[13]

The leftmost panel of Figure 2 on the next page illustrates that when the information about deepfakes was randomly added to a deepfake video, participants were 0.5 points more likely to believe the video they were watching was fabricated ($p<0.001$).[14] This treatment effect was not driven by people affirmatively identifying the deepfake. When participants watched a deepfaked video without information about deepfakes, they were fairly confident that what they were watching was real (-0.4 points on our 7-point scale or approximately halfway between "somewhat disagree that the video is a deepfake" and "neither agree nor disagree that the video is a deepfake"); when information about deepfakes was added, they became more uncertain about whether the video was a deepfake (0.1 points).[15] This result is consistent with Vaccari and Chadwick's 2020 conclusion that deepfakes increase uncertainty. It is important to note that the treatment effect does not simply increase the likelihood that a subject will select what may be construed as the "I don't know" option. Namely, in the absence of the information treatment, 43.7% of participants disagreed that the video they watched was a deepfake, 32.1% neither agreed nor disagreed, and 24.2% agreed. However, in the information treatment group, 25.7% of participants disagreed, 38.9% neither agreed nor disagreed, and 35.4% agreed that the video was a deepfake. (See Table OA4a for a full breakdown of responses.)

The next step is to investigate whether the effect of providing information about deep-

13. Alternative survey instruments found similar (but smaller) treatment effects for all analyses. The results of these alternative outcome measures can be viewed in Appendix Figures OA2a and OA2b.

14. Although there is no objective scale for the size of effects in all contexts, Cohen's d statistic has been used to provide a descriptive interpretation of the magnitude of an effect based on the standard deviation of the outcome variable, particularly for psychological interventions. In Study 1, the Cohen's d is 0.3, which suggests that this effect size is small to medium (Cohen 2013).

15. Participants were not able to discern between a real video and a deepfake version of the same video without the information treatment (see Section 3 of the Appendix for more details).

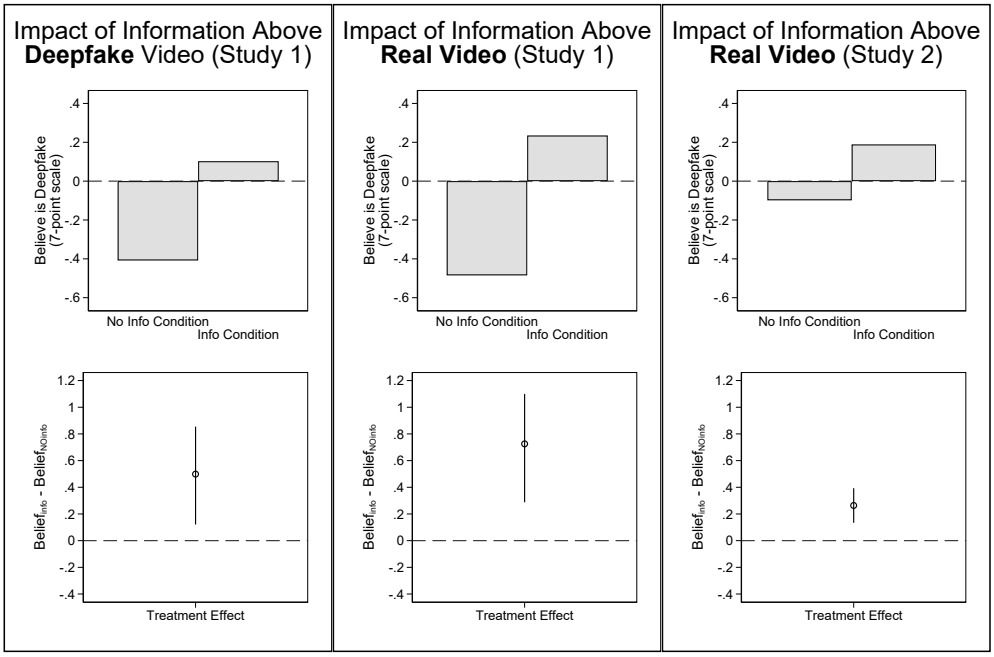

Figure 2: Information about deepfakes induced disbelief in the accompanying video, regardless of whether it was real or fake.

fakes has the unintended consequence of reducing belief in real videos. We see a comparable treatment effect (0.73 points [p<0.001]) when information about deepfakes was randomly added to a real video with the same content (middle panel in Figure 2).[16] For a full breakdown of the responses selected in each condition, see Table OA4b. The interaction of the deepfake treatment and the information treatment is not statistically significant.[17] Rather than helping participants detect deepfake videos, information about deepfakes instead caused them to disbelieve the video they were watching—real or fake. This interaction effect is directionally opposite from the normative ideal: adding information about deepfakes to videos made participants more likely to disbelieve the real video rather than successfully identify the deepfake.

We find similar effects with the real video of Romney in Study 2. The results are summarized in the rightmost panel of Figure 2. As before, information about deepfakes induced participants to disbelieve real videos—an effect size of 0.26 points on a 7-point scale (p<0.001).[18] As in Study 1, the difference is not simply driven by an increase in doubt; in fact, in Study 2, uncertainty appeared to decrease slightly. In the absence of the information treatment, 29.9% of participants disagreed that the video they watched was a deepfake, 45.9% neither agreed nor disagreed, and 24.2% agreed. However, in the information treatment condition, 24.8% of participants disagreed, 40.6% neither agreed nor disagreed, and 34.6% agreed that the video was a deepfake. (Table OA4c shows a detailed breakdown of the responses selected by participants in each condition.)

Pooling the data from both studies and including a study fixed effect, we find that information about deepfakes increased participants' belief that the video was fabricated by 0.40

16. The Cohen's d of this effect is 0.5, which indicates a medium effect size (Cohen 2013).
17. Even without bounds, the p-value is 0.19 and the effect is in the opposite direction from what would be speculated.
18. Cohen's d of this effect is 0.2, indicating a small (but nontrivial) effect size (Cohen 2013).

points on a 7-point scale ($p<0.001$).[19] Information about deepfakes also increased how unconvinced participants were that the politician actually believed what he was saying by 0.13 points on a 3-point scale ($p<0.001$). Finally, information about deepfakes increased the rate at which participants said they "didn't know whether the politician believed what was said in the video" (a binary variable) by 6.0 percentage points ($p=0.001$).

We also evaluated whether information about deepfakes can affect how participants internalize the *content* of the video. Namely, does information about deepfakes change what facts participants associate with the politician in the video? Towards the end of Study 2, we asked participants to name three facts about Romney in an open-ended question. We found that information about deepfakes did not significantly change how likely participants were to mention abortion ($p=0.95$), but there was a major shift in what participants perceived Romney's position on abortion to be. The information about deepfakes caused a 3.1-percentage-point drop in the percentage of participants who associated Romney with a pro-choice position ($p<0.001$). (For more details on the open-ended question, please see Table OA3.)[20] An alternative measure of this outcome asked participants in a close-ended question if Romney had ever "supported women's access to abortion"; we found that adding information about deepfakes decreased the likelihood that participants marked "true" by 14.1 percentage points ($p<0.001$).

3.1 Heterogeneous Treatment Effects

We also examined heterogeneous treatment effects, but as we noted in both pre-analysis plans, we were underpowered for most of these analyses. We briefly overview the most noteworthy results here and report the full results in Sections 4–5 of the Appendix. We found some evidence that participants surveyed before the election displayed higher levels of distrust in political videos than those surveyed afterwards ($p<0.05$). While we originally planned to investigate motivated reasoning in Study 2, by the time the study was launched, Romney had already become a polarizing figure among Republicans, so our data does not allow us to adequately address this question. Section 6 in the Appendix discusses this point in more detail.

4 Discussion and Conclusions

Our findings add to the growing body of literature on textual misinformation, which has consistently found that warnings about fake news may help readers reject misinformation but may have the undesirable effect of increasing Americans' disbelief in true news stories (e.g., Pennycook, Bear, et al. 2020; Clayton et al. 2020). We find strong evidence that cautionary information about deepfakes increases disbelief in accompanying video clips—regardless of whether the video is fake or real. This is particularly problematic as the information treatment did not state that the accompanying video was fake, only that deepfakes exist and are challenging to spot, which is something that an American might hear on a news program. Such a nudge nevertheless induced people to disbelieve a video that revealed a real politician's little-known but true policy stance. And participants not only suspected the video was fake; the information treatment changed their beliefs

19. We exclude Study 1 participants randomized to the deepfake condition in the reported specification, but the results do not change meaningfully when we include them.

20. We report the most conservative coding so that ambiguous responses like "support women" and "fight for women's causes" are coded as "other." As a robustness check, we conduct the same analysis but with a liberal interpretation of the same responses (i.e., we code these types of ambiguous responses as "pro-choice"). Using this liberal interpretation, we recover a slightly larger treatment effect of 3.2 percentage points ($p<0.001$). The full list of these ambiguous responses can be found in the Appendix, Section 8, Study 2 Free Response Handcoding Cleanup Code.

about the politician's policy stances.

We must caution against generalizing our finding to mean that *any* kind of information about deepfakes is likely to induce disbelief in real political information. Rather, our study is the first to experimentally measure the impact of alarmist messaging about deepfakes that often appears in the popular press.[21] This is not to say that this is the only way of informing voters about deepfakes. Indeed, other approaches may inform without increasing disbelief in real political videos. For instance, is it possible to train humans to detect deepfakes? This type of preemptive training is sometimes referred to as inoculation against misinformation.

Recent research has found that in some contexts, certain inoculation methods may improve respondents' ability to discern between deepfake and real videos somewhat (Barari, Lucas, and Munger 2021; Groh et al. 2021), but there is evidence that the inoculation effects vary depending on individuals' levels of digital literacy, political knowledge, partisanship, and the content of the deepfake (Barari, Lucas, and Munger 2021). Future research should therefore expand on this nascent research and investigate other ways of informing and training individuals to detect deepfakes (e.g., by drawing their attention to context cues, such as the source of the video (Cf. Swire et al. 2017)).

We also highlight that our study is rooted in the current media environment, where political deepfakes are fairly uncommon. Thus it is possible that repeated exposure to deepfakes may reduce some of the backlash effects we present in this paper. That said, Groh et al. (2021) find that discernment did *not* increase with practice (over the course of ten trials). Since we were concerned that mentioning deepfakes prior to our treatment might prime participants' suspicion, we did not ask whether they had encountered a deepfake before participating in our study. Previous studies have found relatively low levels of prior exposure. For instance, Vaccari and Chadwick (2020) found that only 4.1% of their sample had seen the deepfake of Jordan Peele posing as Obama, so unless that rate increased markedly, our study would have been underpowered to measure differences in treatment effects based on prior exposure. Future studies should investigate how increases in the number of political deepfakes in the information environment affect the backlash effects we report here.

There are at least three notable limitations to these findings. First, the data is sourced from online survey experiments with convenience samples of participants, which means that there are legitimate concerns about the external validity of our results (see Coppock 2019, for a thorough analysis of these general concerns). Second, the way the deepfake information treatment is presented in our studies is not how the average person is likely to learn about deepfakes. In the real world, such information usually comes from a news source. One advantage of our treatment is that we remove many of the contextual cues of real-world exposure (e.g., news source), which helps isolate the effect of being informed about deepfakes from other related effects (e.g., having an emotional reaction to the source of the information). Still, future studies should investigate the impact of exposure to real-world news articles about deepfakes. Third, the time between exposure to information about deepfakes and videos of politicians making policy statements may be much longer in practice than in our survey experiments. Future studies should investigate if (and how quickly) these treatment effects decay. One countervailing possibility is that in the real world, people may receive higher dosages of deepfake information (e.g., through

21. Indeed, our treatment may understate the intensity of alarm found in deepfake coverage in the news media. For instance, Gosse and Burkell (2020) conducted a qualitative analysis of deepfake news coverage and found that news articles often presented extreme hypothetical scenarios that have not yet occurred (e.g., "a young Bernie Sanders can be shown in KKK rallies…or Vladimir Putin declaring war on Britain" (p. 8)). Yadlin-Segal and Oppenheim (2021) come to similar conclusions, noting that the overwhelming majority of news articles on deepfakes focus on a hypothetical future in which real audiovisual media is completely indistinguishable from deepfakes.

more extensive news coverage), which may cause larger increases in skepticism. As such, future research should assess the impact of dosage.

Our results illustrate that, while well intentioned, attempts to warn the public about deepfakes may inadvertently cause the delegitimization of true information. Our findings therefore suggest that the news media, elites, and social media platforms may need to take great care in their attempts to educate the public. We show that providing information about the existence and potential dangers of deepfakes erodes trust; thus future research should explore other approaches.

References

Associated Press News. 2021. "Cheerleader's mom accused of making 'deepfakes' of rivals" (March 15, 2021). https://apnews.com/article/pennsylvania-doylestown-cheerleading-0953a60ab3e3452b87753e81e0e77d7f.

Barabas, Jason, Jennifer Jerit, William Pollock, and Carlisle Rainey. 2014. "The question (s) of political knowledge." *American Political Science Review* 108 (4): 840–55. https://doi.org/doi:10.1017/S0003055414000392.

Barari, Soubhik, Christopher Lucas, and Kevin Munger. 2021. "Political Deepfake Videos Misinform the Public, But No More than Other Fake Media." *OSF Preprints. January* 13. https://osf.io/preprints/cdfh3/.

Castillo, Michelle. 2018. "Fake video news is coming, and this clip of Obama 'insulting' Trump shows how dangerous it could be" (April 17, 2018). https://www.cnbc.com/2018/04/17/jordan-peele-buzzfeed-psa-edits-obama-saying-things-he-never-said.html.

Clayton, Katherine, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 2020. "Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media." *Political Behavior* 42 (4): 1073–95. https://doi.org/10.1007/s11109-019-09533-0.

Cohen, Jacob. 2013. *Statistical power analysis for the behavioral sciences.* Academic press.

Coppock, Alexander. 2019. "Generalizing from survey experiments conducted on Mechanical Turk: A replication approach." *Political Science Research and Methods* 7 (3): 613–28. https://doi.org/doi:10.1017/psrm.2018.10.

Coppock, Alexander, and Oliver A McClellan. 2019. "Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents." *Research & Politics* (London, England) 6 (1). https://doi.org/10.1177/2053168018822174.

Dack, Sean. 2019. "Deep Fakes, Fake News, and What Comes Next" (March 20, 2019). https://jsis.washington.edu/news/deep-fakes-fake-news-and-what-comes-next/.

Dobber, Tom, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes de Vreese. 2021. "Do (microtargeted) deepfakes have real effects on political attitudes?" *The International Journal of Press/Politics* (Los Angeles, CA) 26 (1): 69–91. https://doi.org/10.1177/1940161220944364.

Ferraro, Matthew. 2020. "Congress's deepening interest in deepfakes" (December 29, 2020). https://thehill.com/opinion/cybersecurity/531911-congresss-deepening-interest-in-deepfakes.

Fowler, Anthony, and Michele Margolis. 2014. "The political consequences of uninformed voters." *Electoral Studies* 34:100–110. https://doi.org/10.1016/j.electstud.2013.09.009.

Gerber, Alan, and Donald Green. 2012. *Field Experiments: Design, Analysis, and Interpretation.* W.W. Norton.

Gosse, Chandell, and Jacquelyn Burkell. 2020. "Politics and porn: how news media characterizes problems presented by deepfakes." *Critical Studies in Media Communication* 37 (5): 497–511. https://doi.org/10.1080/15295036.2020.1832697.

Groh, Matthew, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2021. "Comparing Human and Machine Deepfake Detection with Affective and Holistic Processing." *arXiv preprint,* arXiv: arXiv:2105.06496. https://arxiv.org/pdf/2105.06496.

H.R.6395 — 116th Congress (2019-2020). 2021. *William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021.* 116th Congress (2019-2020). https://www.congress.gov/bill/116th-congress/house-bill/6395.

Harwell, Drew. 2021. "Remember the 'deepfake cheerleader mom'? Prosecutors now admit they can't prove fake-video claims" (March 14, 2021). https://www.washingtonpost.com/technology/2021/05/14/deepfake-cheer-mom-claims-dropped/.

Kuklinski, James H, Paul J Quirk, Jennifer Jerit, David Schwieder, and Robert F Rich. 2000. "Misinformation and the currency of democratic citizenship." *The Journal of Politics* 62 (3): 790–816. https://www.jstor.org/stable/2647960.

Lazer, David MJ, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. "The science of fake news." *Science* 359 (6380): 1094–96. https://doi.org/10.1126/science.aao2998.

Manski, Charles F. 2003. *Partial identification of probability distributions.* Springer Science & Business Media.

McDonald, Jared. 2019. "Avoiding the Hypothetical: Why "Mirror Experiments" are an Essential Part of Survey Research." *International Journal of Public Opinion Research* 32 (2): 266–83. https://doi.org/10.1093/ijpor/edz027.

Mullinix, Kevin J, Thomas J Leeper, James N Druckman, and Jeremy Freese. 2015. "The generalizability of survey experiments." *Journal of Experimental Political Science* 2 (2): 109–38. https://doi.org/doi:10.1017/XPS.2015.19.

O'Sullivan, Donie. 2019. "House Intel chair sounds alarm in Congress' first hearing on deepfake videos" (June 13, 2019). https://www.cnn.com/2019/06/13/tech/deepfake-congress-hearing/index.html.

Paris, Britt, and Joan Donovan. 2019. "Deepfakes and Cheapfakes: The Manipulation of Audio and Visual Evidence" (September 18, 2019). https://datasociety.net/library/deepfakes-and-cheap-fakes/.

Pennycook, Gordon, Adam Bear, Evan T Collins, and David G Rand. 2020. "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings." *Management Science* 66 (11): 4944–57. https://doi.org/10.1287/mnsc.2019.3478.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. "Shifting attention to accuracy can reduce misinformation online." *Nature* 592 (7855): 590–95. https://doi.org/10.1038/s41586-021-03344-2.

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* (Los Angeles, CA) 31 (7): 770–80. https://doi.org/10.1177/0956797620939054.

Peyton, Kyle, Gregory A. Huber, and Alexander Coppock. 2021. "The Generalizability of Online Experiments Conducted During the COVID-19 Pandemic." *Journal of Experimental Political Science,* 1–16. https://doi.org/10.1017/XPS.2021.17.

S.2904 — 116th Congress (2019–2020). 2020. *Identifying Outputs of Generative Adversarial Networks Act.* 116th Congress (2019–2020). https://www.congress.gov/bill/116th-congress/senate-bill/2904.

Sussman, Anna. 2020. "The Loneliness of the Pro-Choice Republican Woman" (October 30, 2020). https://www.newyorker.com/news/us-journal/the-loneliness-of-the-pro-choice-republican-woman.

Swaim, Barton, and Jeff Nussbaum. 2016. "The Perfect Presidential Stump Speech" (October 3, 2016). https://projects.fivethirtyeight.com/perfect-stump-speech/.

Swire, Briony, Adam J Berinsky, Stephan Lewandowsky, and Ullrich KH Ecker. 2017. "Processing political misinformation: comprehending the Trump phenomenon." *Royal Society open science* 4 (3): 160802. https://doi.org/10.1098/rsos.160802.

Ternovski, John, and Lilla Orr. 2022. "Evidence of rising rates of inattentiveness on Lucid in 2020," https://osf.io/preprints/socarxiv/8sbe4/.

Thalen, Mikael. 2021. "Case against 'deepfake mom' falls apart after prosecutors backtrack on manipulated video claims" (May 17, 2021). https://www.dailydot.com/debug/deepfake-mom-cheerleaders-prosecutors-backtrack/.

Thomas, Daniel R, Sergio Pastrana, Alice Hutchings, Richard Clayton, and Alastair R Beresford. 2017. "Ethical issues in research using datasets of illicit origin." In *Proceedings of the 2017 Internet Measurement Conference,* 445–62. https://doi.org/10.1145/3131365.3131389.

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* (London, England) 6 (1): 2056305120903408. https://doi.org/10.1177/2056305120903408.

Van Duyn, Emily, and Jessica Collier. 2019. "Priming and fake news: The effects of elite discourse on evaluations of news media." *Mass Communication and Society* 22 (1): 29–48. https://doi.org/10.1080/15205436.2018.1511807.

Villarreal, Daniel. 2021. "Parler Users Call Trump's Concession Video 'Deep Fake,' and Worry He'll 'Sell Us Out'" (January 7, 2021). https://www.newsweek.com/parler-users-call-trumps-concession-video-deep-fake-worry-hell-sell-us-out-1559895.

Wittenberg, Chloe, Ben M Tappin, Adam J Berinsky, and David G Rand. 2021. "The (minimal) persuasive advantage of political video over text." *Proceedings of the National Academy of Sciences* 118 (47). https://doi.org/10.1073/pnas.2114388118.

Yadlin-Segal, Aya, and Yael Oppenheim. 2021. "Whose dystopia is it anyway? Deepfakes and social media regulation." *Convergence* (London, England) 27 (1): 36–51. https://doi.org/10.1177/1354856520923963.

YouGov. 2021. "The most popular news websites in America" (March 30, 2021). https://today.yougov.com/ratings/media/popularity/news-websites/all.

Authors

John Ternovski is a Postdoctoral Fellow at the McCourt School of Public Policy at Georgetown University. **Joshua Kalla** is an Assistant Professor of Political Science and Statistics & Data Science at Yale University. **P. M. Aronow** is an Associate Professor, with tenure, of Political Science, Biostatistics, and Statistics & Data Science at Yale University.

Acknowledgements

We thank Gregory A. Huber, David G. Rand, and Todd Rogers for useful discussions. This paper was presented at the Harvard Experimental Political Science Conference, and we thank all involved for the opportunity to present and for their helpful feedback. We are grateful to Grace Kang and Rosa Kleinman for their research assistance.

Data Availability Statement

Replication files are available at:
https://osf.io/rqfz5/?view_only=e2807b367a534262bb6c7aeb5727b999

The study preanalysis plan is registered with:
https://osf.io/rqfz5/?view_only=e2807b367a534262bb6c7aeb5727b999

Funding Statement

This research was funded by P. M. Aronow's Research Account at Yale University.

Ethical Standards

The authors declare that the human subjects research in this article was reviewed and approved by the Yale University Human Subjects Committee (Protocol #2000027228). The authors affirm that this article adheres to the APSA's Principles and Guidance on Human Subjects Research. Participants were compensated for their participation by the panel provider.

These studies were designed such that there was no deception; they avoided ethical gray areas of similar deepfake studies. If we use existing ethical frameworks (e.g., Thomas et al. 2017), the typical risks of an experimental study on deepfakes are 1) the possibility of behavioral/sentiment change (i.e., the potential to introduce new information into the broader informational environment), 2) the potential for abuse (i.e., malicious actors could use the results of the research) and 3) necessary use (i.e., is it possible to answer the same research question using an alternative design)? We avoid those risks completely, as these studies do not introduce any new misinformation. And because we do not create a deepfake of a real politician, there is no potential for abuse by malicious actors (i.e., there is no material detailing how to create a deepfake of a real politician). Finally, our studies illustrate two alternative designs to study the impact of political deepfakes without creating a new deepfake of a real politician: 1) using a fictional politician or 2) using a real video of a real politician in which only the content appears spurious.

Keywords

deepfakes; experiments; fake news; misinformation

5 Appendix

For the Appendix, please see
https://osf.io/rqfz5/?view_only=e2807b367a534262bb6c7aeb5727b999