Article

Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News

Cristian Vaccari and **Andrew Chadwick**

Abstract

Artificial Intelligence (AI) now enables the mass creation of what have become known as "deepfakes": synthetic videos that closely resemble real videos. Integrating theories about the power of visual communication and the role played by uncertainty in undermining trust in public discourse, we explain the likely contribution of deepfakes to online disinformation. Administering novel experimental treatments to a large representative sample of the United Kingdom population allowed us to compare people's evaluations of deepfakes. We find that people are more likely to feel uncertain than to be misled by deepfakes, but this resulting uncertainty, in turn, reduces trust in news on social media. We conclude that deepfakes may contribute toward generalized indeterminacy and cynicism, further intensifying recent challenges to online civic culture in democratic societies.

Keywords

misinformation, disinformation, uncertainty, political deepfakes, online civic culture

India, April 2018: A video goes viral on WhatsApp, the world's most popular mobile instant messaging platform. The footage, seemingly from a CCTV camera, shows a group of children playing cricket in the street. Suddenly, two men on a motorbike ride up and grab one of the smallest kids then speed away.[1] This "kidnapping" video creates widespread confusion and panic, spurring an 8-week period of mob violence that kills at least nine innocent people (BBC News, 2018).

The footage that sparked these vendettas was a clever fake—an edit of a video from a public education campaign in Pakistan, designed to raise awareness of child abductions. The educational video opens with the kidnapping but, soon after, one of the hired actors gets off the motorbike and shows a sign cautioning viewers to look after their children. In the fake video that went viral across India, this "big reveal" device was cut: all that remained was a shockingly realistic video of a child being snatched.

In the same month, BuzzFeed publishes a video showing former US President Barack Obama speaking directly to a camera, in what looks like the Oval Office. The first 35 seconds show only Obama's face. Following a few mildly out-of-character statements, Obama drops a bombshell: "President Trump is a total and complete dipshit." After a brief pause, he continues, "Now . . . you see, I would never

say these things, at least not in a public address, but someone else would . . . someone . . . like Jordan Peele." At this point, the true intention of BuzzFeed's video is revealed. This is not actually Obama speaking. A split screen appears showing Obama on the left while on the right is the renowned US actor, comedian, and director, Jordan Peele. Obama's and Peele's facial expressions and lip movements match perfectly. Using artificial intelligence (AI), Peele's production team has digitally reconstructed Obama's face to mirror his. As AI synthesizes Peele's face while Peele impersonates Obama with his voice it becomes clear that this is an ingenious public service announcement about how online video can be manipulated. The BuzzFeed video immediately went viral. Accompanied by the suitably clickbait tagline, "You Won't Believe What Obama Says In this Video! ☺," it notched up 5 million views and 83,000+ shares on Facebook, 5 million+ views on YouTube, and 4.75 million views and

Loughborough University, UK

Corresponding Author:
Cristian Vaccari, Loughborough University, Brockington Building, U3.19, Epinal Way, Loughborough, Leicestershire LE11 3TU, UK.
Email: c.vaccari@lboro.ac.uk

almost 52,000 retweets on Twitter (Facebook, 2018; Twitter, 2018; YouTube, 2018).

The Rise of Deepfakes

While "photoshopping" still images has long been a mainstay of digital culture, manipulated videos of people now increasingly find their way online. BuzzFeed created its video using increasingly common techniques known as "synthetic media" (Witness, 2018) or "deepfakes." Relying on machine learning algorithms, software applications create highly convincing "face-graft" videos where the expressions of one person are carefully superimposed onto the head of another (GitHub, 2019a, 2019b). Alternatively, existing recordings of a person's mouth movements and voice can be used to reverse engineer their speech to have them say any sentence. The results can be alarmingly convincing, especially with the low-resolution video that is common online.

Political deepfakes are an important product of the Internet's visual turn. They are at the leading edge of online, video-based disinformation and, if left unchallenged, could have profound implications for journalism, citizen competence, and the quality of democracy (Bennett & Livingston, 2018; Chadwick et al., 2018; Flynn et al., 2017; Rojecki & Meraz, 2016; Waisbord, 2018).[2] This study provides the first evidence on the deceptiveness of deepfakes. Anecdotal evidence suggests that the prospect of mass production and diffusion of deepfakes by malicious actors could present the most serious challenge yet to the authenticity of online political discourse. Images have stronger persuasive power than text and citizens have comparatively weak defenses against visual deception of this kind (Newman et al., 2015; Stenberg, 2006).

We ran an online experiment among a representative sample ($N = 2,005$) to identify the extent to which editing out the all-important educational "big reveal" in the BuzzFeed Obama/Peele deepfake results in individuals being misled or becoming uncertain about whether the video was true or false. In other words, our experiment reproduces the problem generated by the malicious fake kidnapping video that went viral in India.

While we do not find evidence that deceptive political deepfakes misled our participants, they left many of them uncertain about the truthfulness of their content. And, in turn, we show that uncertainty of this kind results in lower levels of trust in news on social media. Based on these results, we argue that, if unchecked, the rise of political deepfakes will likely damage online civic culture by contributing to a climate of indeterminacy about truth and falsity that, in turn, diminishes trust in online news.

The Renewed Power of Visual Communication

The power of visual communication has been a classic object of inquiry in political communication research. In a landmark experiment, Graber (1990) found that television viewers were more likely to accurately recall visual messages than verbal messages. Grabe and Bucy (2009) showed that "image bites" (i.e., clips where candidates are shown but not heard) are more powerful in shaping voters' opinions than "sound bites" (where candidates are heard talking, with or without images of them speaking). Prior (2013) found that survey respondents demonstrated higher levels of knowledge when questions probing factual recall featured both visual and verbal information.

Visuals enhance the transmission of information by helping citizens establish and retrieve memories. Stenberg (2006) shows that individuals process visual information more directly and with less effort than verbal information. Witten and Knudsen (2005) argue that, due to its perceived "precision," visual information is integrated more effectively than other types of sensory data. Misleading visuals are more likely than misleading verbal content to generate false perceptions because, based on the "realism heuristic" (Frenda et al., 2013; Sundar, 2008), individuals treat audio and images as more likely than text to resemble "the real world" of everyday experience.

When images and audiovisual content are easier to understand and process than written text this brings into play "metacognitive experience": the experientially derived feelings-about-our-thinking that shape our responses to tasks such as processing new information (Schwarz et al., 2007). One such experience, "fluency," is particularly important for understanding why people believe false information. People are more likely to accept messages as true if they perceive them as familiar (Berinsky, 2017). Familiarity elicits a "truthiness effect"—a sense of fluency that makes material easier to assimilate and therefore more credible (Newman et al., 2015). Due to their technical realism, and particularly if they depict already well-known public figures, deepfake political videos potentially intensify the already serious problem that fluency can be generated through familiarity, irrespective of the veracity of the video's content.

Social media users' sharing behavior also matters. Video and still images are more likely than news and online petitions to spread on Twitter (Goel et al., 2015, p. 186). During the 2016 US Presidential campaign, tweets from Donald Trump and Hillary Clinton that contained images or videos received significantly more likes and retweets (Pancer & Poole 2016).

Political Deepfakes as a Distinctive Form of Visual Disinformation

Deepfakes can be synthesized thanks to an AI technology called Generative Adversarial Networks (GANs; Goodfellow et al., 2014). The average person has a predictable range of jaw, lip, and head movements that correspond with the sounds they make when forming words. GANs use authentic video footage as a training set and create a competition between two

software neural networks, such that each improves based on the output of the other. Using this technique, Suwajanakorn et al. (2017) realistically synthesized both audio and video content of humans speaking. Thies et al. (2016) developed software that enables anyone with a webcam to generate replicas of other people's facial expressions. The most powerful technique produces "self-reenactment" video that reconstructs a speaker's facial expressions in real time (Rössler et al., 2018). Huge amounts of footage of political actors are currently available for free online. When used as training data for GANs (run by software that is also freely available), these materials enable users to create fabricated but realistic videos of public figures that may then be shared online without any obvious markers distinguishing them from genuine footage. AI is also being used to synthesize high quality audio mimicking human voices (Baidu Research, 2017; Gault, 2016).

Most people may be poorly equipped to discern when they are being deceived by deepfakes. Rössler et al. (2018) found that people correctly identify fakes in only about 50% of cases—statistically as good as random guessing. Detection is especially poor when evaluating videos with the smearing and blockiness artifacts caused by the compression commonly used on social media. AI-based methods are marginally better than humans, but their effectiveness also declines when video compression is used.

Theorizing Deepfakes' Impact: Deception, Uncertainty, and Trust

Deepfakes are a new and unique form of video-based visual disinformation. At the time of this writing, there is no academic research on their effects. In this study, we assess whether deepfakes affect individuals' perceptions of truth and falsity but, just as importantly, whether they create uncertainty about the information they convey. Finally, we consider whether the uncertainty elicited by deepfakes may reduce people's trust in news on social media.

Our initial focus is on *cognitive* outcomes. The obvious core of the problem is that deepfakes may deceive people. However, even if viewers are not deceived by a deepfake, they may become uncertain about whether their content is true or false. Uncertainty is conceptually distinct from ambivalence. Ambivalence arises when individuals are faced with a choice on which they have conflicting opinions, so that "additional information only heightens the internalized conflict" (Alvarez & Brehm, 1997, p. 346). By contrast, uncertainty is experienced when not enough information is available to make a choice, and thus it can be overcome by the introduction of new information (Alvarez & Brehm, 1997). As Downs (1957) argued, uncertainty arises among citizens because the costs of acquiring accurate information are too high. Deepfakes may increase the costs of getting accurate information, increasing uncertainty as a result. Thus, we focus on whether deceptive deepfakes generate uncertainty about the information they contain.

If deepfakes, among other methods of disinformation, succeed in increasing uncertainty, one of the main implications may be a reduction of trust in news on social media, where deepfakes are likely to circulate most widely. Hence, our second focus is on a potential *attitudinal* outcome of deepfakes: *trust in political news on social media*. Trust in news is declining across the world (Hanitzsch et al., 2018) and trust in news on social media is now lower than in news accessed through other channels (Newman et al., 2018).

Scholars have examined the relationship between trust and uncertainty from different perspectives. On the one hand, trust has often been conceptualized as providing "a solution to the problems caused by social uncertainty" (Yamagishi & Yamagishi, 1994, p. 131). Similarly, Tsfati and Cappella (2003, p. 505) argue that "for trust to be relevant there has to be some uncertainty on the side of the trustor." According to this approach, uncertainty precedes trust and, under certain conditions, elicits it. On the other hand, trusting others may become more difficult when uncertainty increases. Cook and Gerbasi (2011, p. 219) emphasize "situational factors (such as level of uncertainty and risk)" among the reasons why people do not trust each other. Increased uncertainty has been found to reduce trust in business decisions (Adobor, 2006), negotiated and reciprocal exchanges (Molm et al., 2009), use of e-commerce websites (Angriawan & Thakur, 2008), and reliance on market research (Moorman et al. 1993). With respect to problematic information online, increased uncertainty may explain why Van Duyn and Collier (2018) found that exposing people to elite tweets about the problem of fake news reduces the public's trust in news.

Hence, driving our study is the concern that, over time, in common with other sources of false information (e.g., Vosoughi et al., 2018), deepfakes may cultivate the assumption among citizens that a basic ground of truth cannot be established. Research shows that a "need for chaos"—a desire to "watch the world burn" without caring about the consequences—is one driver of false political rumors online (Petersen et al., 2018). Sowing uncertainty about what is true and what is not has become a key strategic goal of state-sponsored propaganda. Writing about Russian operations, Pomerantsev (2015) notes, "The aim is . . . to trash the information space so the audience gives up looking for any truth amid the chaos." The cumulative effect of multiple contradictory, nonsensical, and disorienting messages that malicious actors introduce into digital discourse (Chadwick et al., 2018; Phillips & Milner, 2017) may generate a systemic state of uncertainty. In this context, it becomes especially important to focus on whether deepfakes generate uncertainty and reduce trust.

Hypotheses

Using an experiment, we test three hypotheses, contrasting the responses of participants exposed to two deceptive versions of the BuzzFeed Obama/Peele deepfake and one educational, unedited version.

We are first interested in the extent to which deepfakes deceive people and create uncertainty. As we discussed earlier, there are good reasons to suggest that many people are not proficient at detecting deepfake video. We reason that (H1) *individuals who watch a deepfake political video that contains a false statement that is not revealed as false are more likely to be deceived* and (H2) are *more likely to experience uncertainty about its content, when compared with users who watch a deepfake political video where the false statement is revealed as false.*

Next, we are interested in the relationship between exposure to deceptive deepfakes and trust in news on social media, as mediated by the experience of uncertainty about the content of the deepfake. Based on the arguments outlined in the previous section, we reason that, if exposure to a deepfake through social media results in uncertainty about its content, this heightened uncertainty may then reduce levels of trust in news on social media. In other words, *uncertainty about the content of a deepfake mediates the relationship between exposure to a deceitful deepfake and trust in news on social media* (H3).[3]

Research Design, Data, and Method

Design

We assessed how a representative sample (*N*=2,005) responded to three variants of the BuzzFeed Obama/Peele video, two of which were deceptive, one of which contained the educational reveal. This between-subjects design enabled us to assess whether exposure to a deceptive versus an educational deepfake affects how participants evaluated the video and what levels of trust in social media they reported. Our study does not include a control group, an issue on which we reflect in the "Limitations" section below.

Treatments

We chose to use an existing political deepfake for two reasons. First, the fact that the video is a known viral success enhances the external validity of our experiment. Second, the BuzzFeed video can easily be split into different segments which, when watched in isolation, expose viewers to very different information. Appendix 1 in our Supplementary Information provides the full transcripts and video download links.

We edited the original video to create three separate videos. Two videos were *deceptive* because they left out the second half of the original video, which revealed that Obama's face is synthetically reconstructed and his voice impersonated. The third video was *educational*: it was the full-length video that included the split-screen revelation of the deepfake and that it is Peel, not Obama, speaking.

The first deceptive treatment shows the synthetic Obama saying, "President Trump is a total and complete dipshit." This video is four seconds long and does not provide any

cues to contextualize Obama's statement or suggest it may be false. The length of this message is comparable to the short videos that are often shared on social media. Hereafter, we label this treatment the "deceptive 4-second clip."

The second deceptive treatment contains the first 26 seconds of the BuzzFeed Obama/Peele video, hence we label it the "deceptive 26-second clip." As with the first treatment, viewers were not provided with any explicit information that may have led them to question this clip's authenticity. However, the video starts with the fake Obama saying, "We live an era in which our enemies can make us look like anyone is saying anything at any point in time, even if we would never say those things," so the video provides some subtle verbal cues that may alert viewers to its falsity. This cut of the video ends with Obama calling Trump a dipshit in the same way as the shortest video. We introduced this longer deceptive video because we wanted to explore if it might establish fluency and therefore acceptance, as viewers were exposed to the fake footage for a longer period of time than with the deceptive 4-second clip.

Finally, the third treatment is educational and comprises the full original video, which lasts for 1 minute 10 seconds and features two parts—one where the synthetic Obama speaks alone on camera and calls Trump a dipshit, the other when this is revealed as an artificial creation and Jordan Peele is shown impersonating Obama. The video ends with a warning about deepfakes, spoken by Peele impersonating Obama's voice but visually represented using Obama's synthetic face and Peele's real face. We label this treatment "Full video with educational reveal."

Measurement of Dependent Variables

Our dependent variables are subjects' evaluations of the truthfulness of the deepfake and trust in news on social media. We measured these after exposing subjects to the treatment.

To see if participants believed that the deepfake was truthful or not, we focused on the most outrageous and unlikely sentence uttered by the synthetic Obama. We asked, "Did Barack Obama ever call Donald Trump a 'dipshit'?" Respondents could answer "Yes," "No," or "I don't know." Asking such a direct and specific factual question enabled us to establish whether participants believed the least believable part of the deepfake—a more stringent test of our hypotheses than if we had focused on a more plausible statement from the video. Also, by using the word "ever" in this question we sought to avoid priming respondents to only factually recall the video they had watched, without reflecting on their beliefs about its veracity.

We used "Yes" answers as indicators that the deepfake deceived participants (H1). We chose to use "Don't know" (DK) answers as indicators that the deepfake had caused uncertainty (H2). While researchers often treat DK as missing data, Berinsky (2004) argues that DKs hold substantive meaning and can be explained by factors inherent

in everyday social interaction but mostly absent from the artificial context of the interview, such as individuals' need to hedge their bets, save face, or express uncertainty or ambivalence. We suggest that researchers of problematic information online could fruitfully take Berinsky's argument further still. Most work on uncertainty in political psychology has focused on issue positions and perceptions of candidate traits, where DK responses can be seen as indicating ambivalence due to conflicting opinions (Alvarez, 1997). In contrast, researchers interested in misperceptions typically ask questions that tap a respondent's assessment of the veracity of information—as we did in asking whether Obama ever called Trump a "dipshit." Here, DK responses can have important substantive meanings as parsimonious indicators of uncertainty. A DK response also has the advantage of being relatively untainted by the social desirability bias or embarrassment factor that might contaminate responses to questions that directly ask about uncertainty. We also deliberately identified DK, and not "no opinion" or a simple refusal to answer, as the expression of uncertainty because refusals differ in important and systematic ways from DKs. Refusals are more likely when a question asks for personally sensitive information; DKs are associated with cognitive effort and uncertainty (Shoemaker et al., 2002).

Finally, to measure trust in news on social media (H3), we asked, "How much do you trust the news and information about politics and public affairs that you see on social media?." Response modes were: "A great deal," "Somewhat," "A little," "Not at all," and "I don't know." Since this question taps into an attitude rather than a statement of fact, DK answers are likely to indicate ambivalence rather than uncertainty. Hence, we excluded from the analysis 6.5% of participants who answered DK to this question.

Participants

We administered our treatments to three randomly selected subsamples of British respondents to an online survey we conducted on a panel recruited by Opinium Research, a leading polling company.[4] We obtained a participation rate of 32.8% and 2,005 respondents completed the questionnaire. Information on the characteristics of our sample is reported in the Supplementary Information file. Compared with lab-based experiments, experiments embedded in online surveys offer greater representativeness and enable rich and realistic treatments, such as those employed in this study (c.f. Iyengar & Vavreck, 2012). Also, because we used a self-administered online questionnaire, responses were less affected by social desirability biases (Kreuter et al., 2008). This is particularly relevant for studies of disinformation, where social desirability may lead to under-reporting.

Procedure

The questionnaire included 8 questions measuring standard socio-demographic characteristics and 21 questions gauging political attitudes, social media usage, and access to and sharing of news on social media. After answering these questions, participants were randomly assigned to watching one of the three treatments, which they could replay one more time after the first viewing. They then answered questions measuring our dependent variables, as well as some response quality checks. The experiment then ended with a debriefing note.[5]

Confounding Factors

Random assignment to the three conditions was effective. Of the 2,005 respondents, 653 (32.5%) saw the deceptive 4-second clip, 683 (34.1%) saw the deceptive 26-second clip, and 669 (33.4%) saw the full video with educational reveal. Randomization checks confirm the three subsamples were evenly balanced in terms of demographic characteristics, political attitudes, digital media use, political talk on social media, and trust in news on social media, all of which we measured before the experiment.[6] Hence, we do not control for these factors in our subsequent analyses: random assignment neutralized their influence on our relationships of interest.

However, when testing our hypothesis on the effects on trust in news on social media, we control for a pretreatment measure of trust in news on social media, to ensure that our estimates are not biased by the fact that individuals with higher levels of trust were less likely to express uncertainty in the first place.[7]

Response Quality Checks

After watching the video, respondents were asked, "Please confirm that you were able to watch the above video successfully," to which all participants answered affirmatively. Our interface measured how much time participants spent on the page where the treatments were hosted. No participant stayed on the page for less than the duration of the video they had been assigned. After showing the video, we asked respondents, "Have you ever seen this video before?" to which 83 participants (4.1%) answered "Yes." We also asked, "Immediately after watching the video, did you do any research (e.g., a Google search) to find out more information around the video?" 35 respondents (1.7%) answered that they did. We did not exclude participants who claimed they had seen the video before or who admitted searching for information about the video because these questions were asked *after* exposure to the treatment. As shown by Montgomery et al. (2018), subsetting data based on post-treatment variables can statistically bias causal estimates and nullify the advantages of random assignment.

Analysis

We first test to what extent subjects exposed to each of the three videos incorrectly answered "Yes" to our question
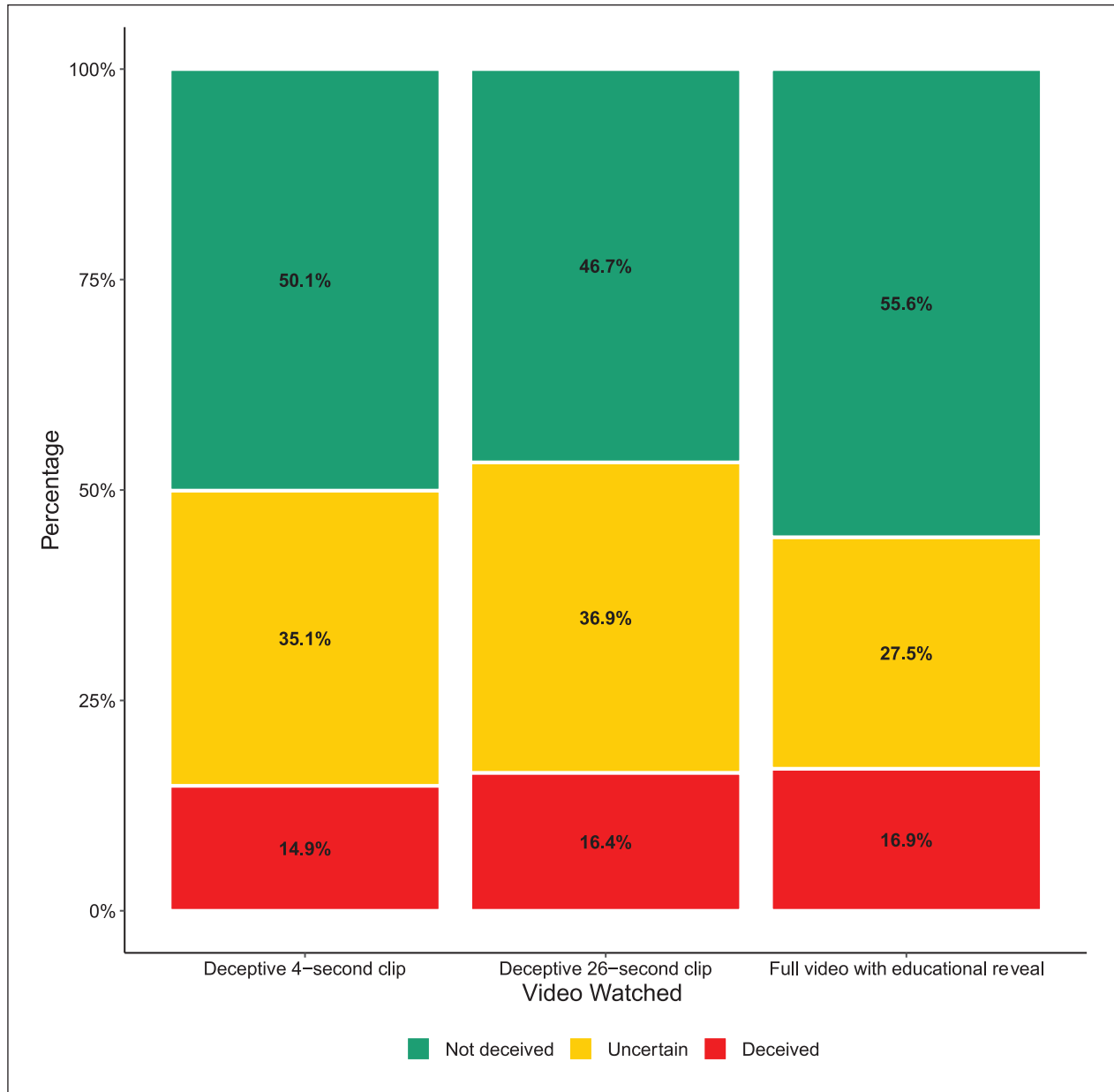
Figure 1. Assessment of the truthfulness of the video, by treatment.

asking if Obama called Trump a dipshit and thus were deceived (H1) and to what extent subjects answered "I don't know" and thus were uncertain about its content (H2). We compare responses across participants exposed to the two deceptive videos and to the full video with educational reveal, using Chi-square tests of independence and logistic regressions.

Overall, only 50.8% of subjects were not deceived by the deepfake. This finding is surprising given the statement was highly improbable. A smaller, though by no means negligible, group (16%) was deceived, while 33.2% were uncertain. However, responses differed based on the treatment participants watched. The Chi-Square coefficient (16.1, *df*=4,

$p=.003$) suggests that differences in the responses provided by subjects exposed to different treatments are statistically significant. Pairwise comparisons confirm that the answers of those who watched the full video with the educational reveal differed significantly from those of participants who watched the two deceptive deepfakes. By contrast, the responses elicited by the two deceptive videos did not differ significantly from each other.[8]

As Figure 1 shows, subjects exposed to either the 4-second or the 26-second deceptive deepfakes were *not* more likely to be deceived than those exposed to the full video with the educational reveal. The 4-second deceptive video was actually the least likely (14.9%) to deceive participants,

Table 1. Ordinary Least Squares Regression Mediation Model Predicting Trust in News on Social Media (Y) as a Function of Exposure to Deceptive Deepfake (X) and Uncertainty on the Truthfulness of the Video (M), Controlling for Baseline Levels of Trust in News on Social Media.

Antecedent		Consequent						
		Uncertainty (M)				Trust in news on social media (Y)		
		Coeff.	SE	p		Coeff.	SE	p
Exposure to deceptive deepfake (X)	a	0.085***	0.022	.000	c'	0.005	0.034	.887
Uncertainty (M)		–	–	–	b	−0.175***	0.034	.000
Baseline trust in news on social media	z_1	−0.003	0.037	.925	z_2	0.661***	0.057	.000
N		1,763				1,763		
R^2		0.061				0.075		
F		57.9 (2, 1,760)				54.1 (3, 2,001)		
p		.000				.000		

*$p \leq .05$; **$p \leq .01$; ***$p \leq .001$.

followed by the 26-second deceptive video (16.4%) and the full video with educational reveal (16.9%). These differences, however, are very small and they are not significant when we model participants' "Yes" responses in a logistic regression with the treatments as independent variables.[9] With the full video with educational reveal serving as the reference category, the coefficient for the deceptive 4-second video was −0.152 ($SE=0.151$, $p=.311$); the coefficient for the deceptive 26-second video was −0.035 ($SE=0.146$, $p=.807$). Overall, H1—that individuals who watch a deepfake political video that contains a false statement that is not revealed as false are more likely to believe the false statement—is rejected.

Importantly, however, the results support H2—watching a deepfake that contains a false statement that is not revealed as false is more likely to cause uncertainty. Exposure to either of the deceptive videos resulted in higher levels of uncertainty (35.1% among those who watched the 4-second version and 36.9% among those who watched the 26-second version) than exposure to the full-length video with the educational reveal (27.5%). To assess whether the deceptive videos elicited significantly higher levels of uncertainty, we ran a logistic regression predicting DK responses as a function of the treatment watched.[10] We obtained positive and significant coefficients for both deceptive videos, when compared to the full video with the educational reveal. The coefficient for the deceptive 4-second video was 0.353 ($SE=0.119$, $p=.003$, Holm-adjusted $p=.003$, Bonferroni-adjusted $p=.006$). The coefficient for the deceptive 26-second video was 0.432 ($SE=0.117$, $p=.000$, Holm-adjusted $p=.000$, Bonferroni-adjusted $p=.000$). Hence, both deceptive videos were significantly more likely to elicit uncertainty compared with the full educational video.

Finally, we test whether exposure to a deceptive deepfake reduces levels of trust in news on social media through a mediated path via increased levels of uncertainty (H3). We ran a simple mediation analysis based on an ordinary least

squares regression (Hayes, 2013).[11] The independent variable is exposure to either of the deceptive deepfakes (compared with exposure to the full video with the educational reveal). Uncertainty (i.e., answering DK to the question asking if Obama called Trump a dipshit) is the mediating variable, and trust in news on social media is the dependent variable.[12] The mediation model also controls for baseline levels of trust in news on social media measured before exposure. Table 1 and Figure 2 present the results.

Participants who were exposed to one of the deceptive deepfakes (as opposed to the full video with educational reveal) were significantly more likely to express uncertainty about the content of the video ($a=0.085$), further corroborating H1. In turn, participants who expressed uncertainty on the video manifested significantly lower levels of trust in news on social media ($b=−0.175$), even after controlling for pretreatment levels of trust. The indirect effect on trust in news on social media of exposure to a deceptive deepfake through increased uncertainty is the product of these two coefficients ($ab=−0.015$, $SE=0.005$). Hence, trust in news on social media decreases as a result of watching deceptive deepfakes, and the effect is mediated by the increased uncertainty arising from the treatment. 95% confidence intervals for this indirect effect did not include zero (−0.026 to −0.007, with 5,000 bootstrapped samples). By contrast, exposure to a deceptive deepfake did not affect trust in news on social media directly and independent of its effect on uncertainty ($c'=0.005$, not significant). Importantly, we found no significant correlation between baseline levels of trust in news on social media and uncertainty ($z_1=−0.003$, not significant). Hence, the relationship we found between uncertainty and trust is not confounded by the fact that participants with lower pretreatment levels of trust were more likely to express uncertainty. Instead, baseline levels of trust in news on social media strongly and significantly predicted posttreatment levels of trust ($z_2=0.661$), as expected. H3 is thus confirmed. Compared with the educational video, exposure to one of the
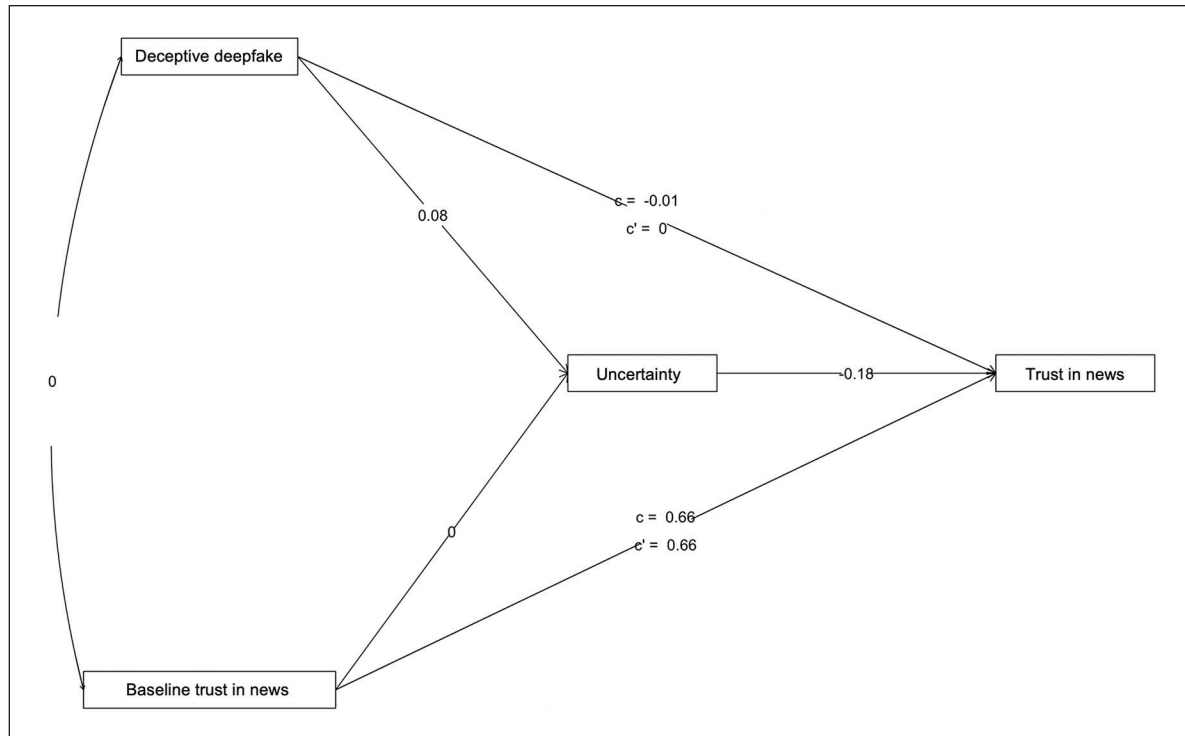
Figure 2. Graphical representation of the mediation model in Table 1.

deceptive deepfakes decreased trust in news on social media indirectly by eliciting higher levels of uncertainty, even after controlling for baseline levels of trust. The corollary is that, compared with the deceptive deepfakes, the full video with the educational reveal increased trust in news on social media through reduced uncertainty.[13]

Limitations

Before reflecting on the implications of our findings, we acknowledge some limitations of this study.

First, as our experiment was administered within an online questionnaire rather than in the field, we must be cautious about its external validity. In common with all survey experiments, this study could only establish the likely effects of exposure to a single deepfake video at a single moment in time. It did not account for interpersonal networks, algorithmic filtering, and competition with other messages—all factors that are likely to play a role in promoting or debunking deepfakes outside the contained environment of an experiment. The validity of our findings is, however, enhanced by the widely-recognized strengths of experimental methods. We randomly assigned carefully designed treatments to a large, representative sample and in a tightly controlled environment in which we assessed people's attitudes at the pre-exposure and post-exposure stages. Our findings therefore suggest that when individuals are exposed to deceptive deepfakes this may have broader social impact in spreading

uncertainty and trust, though we recognize that much further research is needed on the diversity of contextual conditions that will inevitably play a role in these processes.

Second, instead of producing our own treatments, we used basic editing technology to create different versions of an existing deepfake and thus we could not alter its content, apart from selectively cutting it. The key statement in the Obama/Peel deepfake involved a former President generally known for his composure insulting a sitting President with a slang curse word. A subtler message delivered by a video generated using the same AI tools may have sounded more credible. By the same token, the fact that the deceptive 26-second video starts with Obama warning about the ability of "our enemies" to "make us look like anyone is saying anything at any point in time" may have alerted participants and decreased the treatment's deceptive potential. Both these constraints may have made a Type II (false negative) error more likely, at least with respect to H1. As the technology to generate deepfakes develops, scholars may be able to develop custom-built treatments, but the ethical implications would need to be very carefully weighed against the risks.

Third, our experiment did not feature a control group. While this design prevents us from comparing subjects exposed to a deepfake video to those exposed to a "placebo" video, it does allow us to compare responses to the different versions (deceptive or educational) of the same deepfake—which was the focus of our hypotheses.

Fourth, our question measuring whether participants believed the key statement made by the synthetic Obama in the deepfake—"Did Barack Obama ever call Donald Trump a 'dipshit'?"—may have led us to overestimate the levels of uncertainty elicited by *all* treatments. Some participants may have interpreted the question as encompassing both public and private utterances by Obama, and thus they may have been led to answer that they did not know because they felt they *could* not know what Obama may have ever said in private. However, crucially, any such over-estimation of uncertainty would be evenly distributed across participants exposed to all treatments, and hence it should not affect our results when comparing responses among participants who saw different deepfakes.

Fifth, our measure of trust in news on social media is generic, but trust in news on social media is arguably platform specific. We chose to ask a parsimonious catch-all question on social media to avoid an overlong questionnaire and to avoid priming participants that we were particularly interested in this outcome. Future research could address whether the kinds of effects we documented vary across platforms.

Sixth, while we employed various measures of response quality, we did not perform manipulation checks to verify that participants perceived the deepfakes as deceptive or educational. While asking such questions is good practice in experimental research (Thorson et al., 2012), we did not employ them for two reasons. First, Rössler et al. (2018) show that most users have limited ability to discern between a deepfake and an authentic video, so manipulation checks would have been biased by a high degree of guessing. Second, we reasoned that many participants would feel compelled not to admit they had been deceived by the deepfake they watched if we asked them a direct question about its authenticity. While manipulation checks would have strengthened the validity of our findings, the particular object of our study made their use problematic.

Finally, we drew on an online panel-based sample to recruit our participants, and even though this makes our findings more generalizable than if we had drawn a convenience sample, results from nonprobability samples do not automatically generalize to the population (Pasek, 2015). That being said, our sample resembles the adult British population quite closely in terms of gender, age, and education.[14]

Conclusion

We have shown that political deepfakes may not necessarily deceive individuals, but they may sow uncertainty which may, in turn, reduce trust in news on social media.

In the long term, these effects may ripple out to online civic culture, potentially eliciting problematic norms and behaviors. Individuals are less likely to cooperate in contexts where trust is low, and this is particularly the case in high-conflict situations—such as the polarized politics of our times (Balliet & Van Lange, 2013). If social media users become even less trusting in the news they find online, they may become less likely to behave collaboratively and responsibly towards other users when they share news themselves. In the long term, the general expectation that little of what is available online can be trusted may further contribute to an attitudinal spiral that "anything goes" online. This may then diminish individuals' sense of responsibility for the information they share (Chadwick & Vaccari, 2019). It may also lead citizens to escape the news altogether, in order to avoid the stress resulting from uncertainty (Wenzel, 2019).

In this scenario, meaningful public debate would become more difficult, as citizens struggle to reconcile the human tendency to believe visual content with the need to maintain vigilance against manipulative deepfakes. Just as worryingly, at the elite level this online context may create new opportunities to campaign on promises to restore "order" and "certainty" through illiberal policies curtailing free speech and other civil rights (Arendt, 1951). As Hannah Arendt (1978) put it,

> A people that no longer can believe anything cannot make up its own mind. It is deprived not only of its capacity to act but also of its capacity to think and to judge. And with such a people you can then do what you please.

Widespread uncertainty may also enable deceitful politicians to deflect accusations of lying by claiming that nothing can be proved and believed.

Traditional responses to online disinformation may have limited efficacy in this context. Media literacy campaigners have focused on encouraging the public to seek out alternative sources of information and to juxtapose these with any utterance or source that claims to be authoritative (e.g., Aufderheide, 1992). But this aim relies on the premise that a political utterance clearly and observably took place, and what is required is simply contextualization. This model is at the heart of many (though not all) fact-checking organizations. Deepfakes present a distinctive problem for this model, for two reasons. First, because many fact checkers work on the basis that whatever is said in public, a real person has said it, even though the statement may be false. With a deepfake, this would not be the case. Second, and more fundamentally, fact-checking videos in a context where deepfakes abound would need to establish that a video is real, which is comparatively difficult due to deepfakes' technical competence and the fact that deepfakes are generated, in part, from videos that are already publicly available.

The kinds of juxtaposition and contextualization that will enable individuals to identify deepfakes might also prove difficult to institutionalize. Politicians will be quick to issue statements denying that they said what a deepfake video portrayed them as saying. Professional journalists may surface the truth eventually. Small communities of the technologically skilled may be able to discern the glitches introduced by GAN software and report the fakery online, though in the long term there is also the problem that AI-based methods of

detection will become victims of their own success because the training datasets will be used by malicious actors to further refine production of deepfakes. The question is, will all these efforts to counter disinformation through deepfakes be as timely and wide-ranging as necessary? Will they reduce all or most of the negative implications of deceptive deepfakes—spreading uncertainty and reducing trust—that we documented?

On a more optimistic note, we have shown that an educational video about political deepfakes can succeed in reducing uncertainty, and in so doing can increase trust in news on social media, compared with deceptive deepfakes. However, the educational video did not reduce outright deception—a finding that chimes with an important strand of research showing the limited effects of fact-checking (e.g., Garrett et al., 2013).

It is also possible that the reduction of trust in news on social media resulting from the uncertainty induced by deceptive deepfakes may not generate cynicism and alienation, but skepticism (Cappella & Jamieson, 1996). As Lewandowsky et al. (2012, p. 120) argue, "*skepticism* can reduce susceptibility to misinformation effects if it prompts people to question the origins of information that may later turn out to be false," while at the same time ensuring that accurate information is recognized and valued. While skepticism is no panacea (Green & Donahue, 2011), it is much less problematic for democracy than cynicism and may be a sign, or even a component, of a healthily critical but engaged online civic culture. Future research should carefully disentangle whether and under what conditions low trust in news on social media entails cynicism or skepticism.

The role political deepfakes play in public discourse in future will ultimately depend on how a range of different actors approach them. Technology companies are likely to further develop AI tools that generate synthetic representations of humans, but we hope that the same companies also endeavor to use their AI for maintaining the democratic good of authenticity by assisting in the detection of political deepfakes. Social media platforms will determine if automated and human forms of certification and control are likely to facilitate or hinder the publication and sharing of deepfakes. Domestic and international policy actors will employ deepfakes in different ways, from the relatively innocuous, such as public service chatbots, to the pernicious, such as creating and spreading false videos of opponents. Journalists and fact checkers will need to constantly assess the veracity of political deepfakes, identify malicious uses, and make reasoned choices about whether or how to alert the public to the danger. Citizens will try to navigate synthetic media as producers, viewers, commenters, and sharers, and the norms they abide by in adopting these behaviors will be crucial. Finally, political deepfakes will continue to generate significant empirical challenges and troubling normative puzzles for social scientists. It would be unwise to treat deepfakes as mere technological curiosities. The stakes are too high, and political

communication scholars are uniquely placed to understand the implications of political deepfakes for the quality of public debate and the formation of public opinion.

Declaration of Conflicting Interest

Funding

ORCID iD

Cristian Vaccari 🆔 https://orcid.org/0000-0003-0380-8921

Supplemental Material

Supplemental material for this article is available online.

Notes

1. When we wrote the first draft of this article, the video was available on the website of the Australian Broadcasting Corporation, but it has since been removed.
2. We follow others in defining online disinformation as intentional behavior that purposively misleads and online misinformation as unintentional behavior that inadvertently misleads. See Chadwick et al. (2018), who follow Jack (2017). Deepfakes are disinformation because they originate with intentional acts (the creation of the deepfake video). But they become misinformation, too, if circulated online by people who mistakenly believe them to be truthful representations. For the purposes of this study, this distinction is not germane because we do not seek to explain the factors shaping a decision to share a deepfake.
3. We do not expect that the deepfake should have any direct effect on trust in news on social media. As discussed by Hayes (2013, p. 88), testing a mediation model does not require hypothesizing or demonstrating a direct effect of the independent variable on the dependent variable.
4. We are grateful to Opinium Research for conducting the survey pro bono in support of the activities of the Online Civic Culture Centre at Loughborough University. Polling shows that 99% of the British public knows both Obama and Trump (YouGov, 2019a, 2019b).
5. See Supplementary Information, Appendix 3.
6. See Supplementary Information, Appendix 4.
7. For information on this measure see Supplementary Information, Appendix 5.
8. When comparing the full video with the educational reveal and the deceptive 4-second clip, Chi-Square=8.8, df=2, p=.012, adjusted p (Holm)=.024, adjusted p (Bonferroni)=.036; when comparing the full video with the educational reveal and the deceptive 26-second clip, Chi-Square=15, df=2, p=.000, adjusted p (Holm)=.002, adjusted p (Bonferroni)=.002; when comparing the deceptive 4-second and 26-second clips, Chi-Square=1.6, df=2, p=.572, adjusted p (Holm)=.448, adjusted p (Bonferroni)=1.000.

9. For full results of this regression see Supplementary Information, Appendixs 6.

10. For full results of this regression see Supplementary Information, Appendixs 7.

11. We ran the model using the "psych" package in *R* (Revelle, 2018).

12. The mediation model we test includes a mediator measured after the treatment—uncertainty about the deepfake. Montgomery et al. (2018) show that this may compromise random assignment and bias causal inferences. However, they also note that "The lesson here is not that studying mechanisms is impossible or that researchers should give up on trying to understand causal paths." As possible solutions, they cite designs that include "a treatment that affects the mediator but not the outcome" (Montgomery et al., 2018, p. 772). Relatedly, Pearl (2014, p. 4) argues that "there is no need to require that covariates [including mediators] be pretreatment, as long as they are causally unaffected by the treatment." Our model meets these criteria because our treatments affected the mediator (uncertainty), as shown in our discussion of H2 below, but not the outcome of our mediation model—trust in news on social media. Average levels of posttreatment trust in news on social media were 0.673 among participants who watched the deceptive 4-second clip, 0.711 among those who watched the deceptive 26-second clip, and 0.707 among those exposed to the full video with educational reveal. The ANOVA F coefficient was 0.467 ($p = .627$), indicating there was no significant association between treatment watched and trust in news on social media. This is also confirmed by the mediation regression we ran to test H3 (Table 1), which shows no significant direct effect of the treatment on trust in news on social media (Coeff. $= 0.005$, $SE = 0.034$, $p = .887$).

13. For the full results of this regression mediation model see Supplementary Information, Appendix 8. The *ab* coefficient for the indirect effect is the same as in the model in Table 1, but with a positive sign: $-0.085 \times -0.175 = 0.015$ (95% CI $= [0.007, 0.026]$).

14. See Supplementary Information, Appendix 2.

References

Adobor, H. (2006). Optimal trust? Uncertainty as a determinant and limit to trust in inter-firm alliances. *Leadership & Organization Development Journal*, *27*(7), 537–553.

Alvarez, R. M. (1997). *Information and elections*. University of Michigan Press.

Alvarez, R. M., & Brehm, J. (1997). Are Americans ambivalent towards racial policies? *American Journal of Political Science*, *41*(2), 345–374.

Angriawan, A., & Thakur, R. (2008). A parsimonious model of the antecedents and consequence of online trust: An uncertainty perspective. *Journal of Internet Commerce*, *7*(1), 74–94.

Arendt, H. (1951). *The origins of totalitarianism*. Harcourt Brace.

Arendt, H. (1978, October 26). Hannah Arendt: From an Interview. *The* New York Review of Books https://www.nybooks.com/articles/1978/10/26/hannah-arendt-from-an-interview/

Aufderheide, P. (Ed.). (1992). *Media literacy: A report of the national leadership conference on media literacy*. Aspen Institute.

Baidu Research. (2017). *Deep voice 3: 2000-speaker neural text-to-speech*. http://research.baidu.com/Blog/index-view?id=91

Balliet, D., & Van Lange, P. A. M. (2013). Trust, conflict, and cooperation: A meta-analysis. *Psychological Bulletin*, *139*(5), 1090–1112.

BBC News. (2018, June 11). *India WhatsApp "child kidnap" rumours claim two more victims*. https://www.bbc.co.uk/news/world-asia-india-44435127

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122–139.

Berinsky, A. J. (2004). *Silent voices: Public opinion and political participation in America*. Princeton University Press.

Berinsky, A. J. (2017). Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science*, *47*(2), 241–262.

Cappella, J. N., & Jamieson, K. H. (1996). News frames, political cynicism, and media cynicism. *The Annals of the American Academy of Political and Social Science*, *546*(1), 71–84.

Chadwick, A., & Vaccari, C. (2019). *News sharing on UK social media: Misinformation, disinformation, and correction*. https://www.lboro.ac.uk/media/media/research/o3c/Chadwick%20Vaccari%20O3C-1%20News%20Sharing%20on%20UK%20Social%20Media.pdf

Chadwick, A., Vaccari, C., & O'Loughlin, B. (2018). Do tabloids poison the well of social media? Explaining democratically dysfunctional news sharing. *New Media & Society*, *20*(11), 4255–4274.

Cook, K. S., & Gerbasi, A. (2011). Trust. In P. Hedström, P. Bearman, & P. S. Bearman (Eds.), *The Oxford handbook of analytical sociology*. Oxford University Press.

Downs, A. (1957). *An economic theory of democracy*. Harper.

Facebook. (2018, April 17). *You won't believe what Obama says in this video!* https://www.facebook.com/watch/?v=10157675129905329

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, *38*, 127–150.

Frenda, S. J., Knowles, E. D., Saletan, W., & Loftus, E. F. (2013). False memories of fabricated political events. *Journal of Experimental Social Psychology*, *49*(2), 280–286.

Garrett, R. K., Nisbet, E. C., & Lynch, E. K. (2013). Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory. *Journal of Communication*, *63*(4), 617–637.

Gault, M. (2016, November 6). After 20 minutes of listening, new Adobe tool can make you say anything. *Motherboard*. https://www.vice.com/en_us/article/jpgkxp/after-20-minutes-of-listening-new-adobe-tool-can-make-you-say-anything

GitHub. (2019a). *Faceswap*. https://github.com/deepfakes/faceswap

GitHub. (2019b). *DeepFaceLab*. https://github.com/iperov/DeepFaceLab#Where_can_I_get_the_FakeApp

Goel, S., Anderson, A., Hofman, J., & Watts, D. J. (2015). The structural virality of online diffusion. *Management Science*, *62*(1), 180–196.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . .Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *3*, 2672–2680.

Grabe, M. E., & Bucy, E. P. (2009). *Image bite politics: News and the visual framing of elections*. Oxford University Press.

Graber, D. A. (1990). Seeing is remembering: How visuals contribute to learning from television news. *Journal of Communication*, *40*(3), 134–156.

Green, M. C., & Donahue, J. K. (2011). Persistence of belief change in the face of deception: The effect of factual stories revealed to be false. *Media Psychology*, *14*(3), 312–331.

Hanitzsch, T., Van Dalen, A., & Steindl, N. (2018). Caught in the nexus: A comparative and longitudinal analysis of public trust in the press. *The International Journal of Press/Politics*, *23*(1), 3–23.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* Guilford Publications.

Jack, C. (2017) *Lexicon of Lies: Terms for Problematic Information.* Data & Society Research Institute.

Iyengar, S., & Vavreck, L. (2012). Online panels and the future of political communication research. In H. Semetko & M. Scammell (Eds.), *The SAGE handbook of political communication* (pp. 225–240). SAGE.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and Web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131.

Molm, L. D., Schaefer, D. R., & Collett, J. L. (2009). Fragile and resilient trust: Risk and uncertainty in negotiated and reciprocal exchange. *Sociological Theory*, *27*(1), 1–32.

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, *62*(3), 760–775.

Moorman, C., Deshpande, R., & Zaltman, G. (1993). Factors affecting trust in market research relationships. *Journal of Marketing*, *57*(1), 81–101.

Newman, E. J., Garry, M., Unkelbach, C., Bernstein, D. M., Lindsay, D., & Nash, R. A. (2015). Truthiness and falsiness of trivia claims depend on judgmental contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1337–1348.

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., & Nielsen, R. K. (2018). *Reuters Institute digital news report 2018*. Reuters Institute for the Study of Journalism. http://media.digitalnewsreport.org/wp-content/uploads/2018/06/digital-news-report-2018.pdf

Pancer, E., & Poole, M. (2016). The popularity and virality of political social media: Hashtags, mentions, and links predict likes and retweets of 2016 US presidential nominees' tweets. *Social Influence*, *11*(4), 259–270.

Pasek, J. (2015). When will nonprobability surveys mirror probability surveys? Considering types of inference and weighting strategies as criteria for correspondence. *International Journal of Public Opinion Research*, *28*(2), 269–291.

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychological Methods*, *19*(4), 459–481.

Petersen, M. B., Osmundsen, M., & Arceneaux, K. (2018, September 1). A "need for chaos" and the sharing of hostile political rumours in advanced democracies. *PsyArXiv Preprints*. https://psyarxiv.com/6m4ts/

Phillips, W., & Milner, R. M. (2017). *The ambivalent internet: Mischief, oddity, and antagonism online.* Polity.

Pomerantsev, P. (2015, January 4). Inside Putin's information war. *Politico*. https://www.politico.com/magazine/story/2015/01/putin-russia-tv-113960

Prior, M. (2013). Visual political knowledge: A different road to competence? *Journal of Politics*, *76*(1), 41–57.

Revelle, W. (2018). *psych: Procedures for personality and psychological research*. Northwestern University. https://www.scholars.northwestern.edu/en/publications/psych-procedures-for-personality-and-psychological-research

Rojecki, A., & Meraz, S. (2016). Rumors and factitious informational blends: The role of the web in speculative politics. *New Media & Society*, *18*(1), 25–43.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2018). FaceForensics: A large-scale video dataset for forgery detection in human faces. https://arxiv.org/pdf/1803.09179.pdf

Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in Experimental Social Psychology*, *39*, 127–161.

Shoemaker, P. J., Eichholz, M., & Skewes, E. A. (2002). Item nonresponse: Distinguishing between don't know and refuse. *International Journal of Public Opinion Research*, *14*(2), 193–201.

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

Sundar, S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. Metzger & A. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). MIT Press.

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics*, *36*(4), Article 95.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2387–2395).

Thorson, E., Wicks, R., & Leshner, G. (2012). Experimental methodology in journalism and mass communication research. *Journalism & Mass Communication Quarterly*, *89*(1), 112–124.

Tsfati, Y., & Cappella, J. N. (2003). Do people watch what they do not trust? Exploring the association between news media skepticism and exposure. *Communication Research*, *30*(5), 504–529.

Twitter. (2018, April 17). *You won't believe what Obama says in this video!* https://twitter.com/BuzzFeed/status/986257991799222272

Van Duyn, E., & Collier, J. (2018). Priming and fake news: The effects of elite discourse on evaluations of news media. *Mass Communication & Society*, *22*(1), 29–48. https://doi.org/10.1080/15205436.2018.1511807

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Waisbord, S. (2018). Truth is what happens to news: On journalism, fake news, and post-truth. *Journalism Studies*, *19*(13), 1866–1878.

Wenzel, A. (2019). To verify or to disengage: Coping with "fake news" and ambiguity. *International Journal of Communication*, *13*, Article 19.

Witness. (2018, June 11). *Mal-uses of AI-generated synthetic media and deepfakes: Pragmatic solutions discovery convening.* http://witness.mediafire.com/file/q5juw7dc3a2w8p7/Deepfakes_Final.pdf/file

Witten, I. B., & Knudsen, E. I. (2005). Why seeing is believing: Merging auditory and visual worlds. *Neuron*, *48*(3), 489–496.

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and Emotion*, *18*(2), 129–166.

YouGov. (2019a, November 1). *Public figure—Barack Obama.* https://yougov.co.uk/topics/politics/explore/public_figure/Barack_Obama

YouGov. (2019b, November 1). *Public figure—Donald Trump.* https://yougov.co.uk/topics/politics/explore/public_figure/Donald_Trump

YouTube. (2018, April 17). *You won't believe what Obama says in this video!* https://www.youtube.com/watch?v=cQ54GDm1eL0

Author Biographies

Cristian Vaccari (PhD, IULM University in Milan) is professor of Political Communication and co-director of the Centre for Research in Communication and Culture at Loughborough University.

Andrew Chadwick (PhD, London School of Economics) is professor of Political Communication in the Department of Communication and Media at Loughborough University, where he also directs the Online Civic Culture Centre (O3C).