Do (microtargeted) deepfakes have real effects on political attitudes?

Dobber, T.; Metoui, N.; Trilling, D.; Helberger, N.; de Vreese, C.

DOI
10.1177/1940161220944364

Publication date
2021

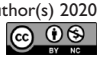
Document Version
Final published version

Published in
International Journal of Press/Politics

Article

Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?

The International Journal of Press/Politics
2021, Vol. 26(1) 69–91



DOI: 10.1177/1940161220944364
journals.sagepub.com/home/hij


Tom Dobber[1], Nadia Metoui[1], Damian Trilling[1], Natali Helberger[1], and Claes de Vreese[1]

Abstract

Deepfakes are perceived as a powerful form of disinformation. Although many studies have focused on detecting deepfakes, few have measured their effects on political attitudes, and none have studied microtargeting techniques as an amplifier. We argue that microtargeting techniques can amplify the effects of deepfakes, by enabling malicious political actors to tailor deepfakes to susceptibilities of the receiver. In this study, we have constructed a political deepfake (video and audio), and study its effects on political attitudes in an online experiment ($N = 278$). We find that attitudes toward the depicted politician are significantly lower after seeing the deepfake, but the attitudes toward the politician's party remain similar to the control condition. When we zoom in on the microtargeted group, we see that both the attitudes toward the politician and the attitudes toward his party score significantly lower than the control condition, suggesting that microtargeting techniques can indeed amplify the effects of a deepfake, but for a much smaller subgroup than expected.

Keywords

deepfake, political microtargeting, disinformation, political attitudes

So-called "deepfakes," many argue, may be a new form of disinformation that is especially challenging to society. These manipulated videos are the result of machine learning, and can make it seem as if a person says or does something, while in reality, they have never said or done anything of the sorts. Using a lot of real examples of

[1]University of Amsterdam, Amsterdam, the Netherlands

Corresponding Author:
Tom Dobber, Amsterdam School of Communication Research, University of Amsterdam, Nieuwe Achtergracht 166, 1018 Amsterdam WV, the Netherlands.
Email: t.dobber@uva.nl

speech and moving images, a so-called neural network is trained that can be used to create a deepfake and deceive citizens. Barack Obama, for example, was once heard and seen calling Donald Trump "a total and complete dipshit" in an online video. In reality, this never occurred. A deepfake made by Jordan Peele made it seem that way.[1] Disinformation conveyed via deepfakes could pose a challenge during elections, since, to the untrained eye, a deepfake may be difficult to distinguish from a real video. Any political actor could try to discredit an opponent or try to incite some political scandal with the goal of furthering their own agenda. After being exposed to a deepfake, citizens may, for instance, change their attitudes toward the politician depicted in the deepfake, or toward the politician's party. As a result, citizens then cast their votes on the basis of false information, and potentially in line with the goals of the political actor behind the deepfake. This can raise questions about the legitimacy of democratic institutions (Bennett and Livingston 2018), the quality of public debate (Xia et al. 2019), the power of citizens (Flynn et al. 2017), and the power of malicious political actors (Bradshaw and Howard 2018).

Whether people indeed "fall for" deepfakes is unclear and understudied, but not unimaginable. Deepfakes consist of largely real images, and producers only manipulate relatively small elements of the video (e.g., facial expressions, voice), which contributes to the realism of the deepfake. In this sense, a deepfake is qualitatively different from a photoshopped image: a deepfake deceives not just the eyes, but the ears as well.

There are several reasons to believe deepfake disinformation can have a detrimental societal impact, which is why studying effects of deepfake disinformation is worth the scientific scrutiny. For one, deepfakes can be realistic disinformation. Automatically generated images and sounds can be as convincing as real sounds and images. An ordinary citizen may struggle to distinguish fact from fiction. Second, deepfakes can be used to amplify existing mis-, dis- or malinformation. A producer could create a deepfake where the pope is seen and heard to endorse Donald Trump, or a public health official ostensibly seen and heard confirming that vaccinations indeed cause autism. Third, deepfakes can also be a form of efficient disinformation. If a political actor has enough training data, the actor can make many different, realistic deepfakes of the same person in a short period of time. In combination with political microtargeting (PMT) techniques, deepfakes can be especially impactful. We are not there yet. Deepfakes do not yet flood the public sphere, let alone *microtargeted* deepfakes. But (microtargeted) deepfakes have the characteristics that make them potentially very powerful modes of disinformation in the near future.

In this paper, we argue that it is not only the technical possibility of creating deepfakes that is troubling, but also the potential consequences of deploying deepfakes *in combination with* PMT. In particular, we expect that the use of PMT techniques is an important amplifier of the disinformation effects of deepfakes.

PMT is a relatively new technique used by political campaigns worldwide (Baldwin-Philippi 2019; Dobber et al. 2017; Dommett 2019; Dommett et al. 2020; Kreiss 2016; Matsumoto 2018; Moura and Michelson 2017). PMT is "a type of personalized communication that involves collecting information about people, and using that information to show them targeted political advertisements" (Borgesius et al. 2018: 82). While

tailored messages are often seen as textual messages or traditional campaigning material, we can easily imagine how a deepfake can be used to try and influence particular subgroups of the electorate.

While there is substantial literature on the technical side of deepfakes, such as detection methods (e.g., Afchar et al. 2018; Agarwal and Varshney 2019; Li and Lyu 2018; Yang et al. 2019), as of yet, deepfakes are only marginally studied in the political communication field. Vaccari and Chadwick (2020) used an existing deepfake to show that deepfakes poison the public debate by confusing people about what is real and what is not. To the best of our knowledge, effects of self-produced (microtargeted) deepfakes on people's political attitudes have never been studied. More knowledge about these effects is needed to understand and counter the threats that deepfake disinformation poses to our democratic societies, for instance, to better inform strategies to combat deepfakes. For this study, we have produced a political deepfake ourselves (video and audio). Using an online experiment, we aim to study the effects of (microtargeted) deepfakes by answering the following key question: To what extent does a (microtargeted) deepfake meant to discredit a politician affect citizens' attitudes toward that politician and his party?

Theoretical Background

Deepfakes as a Form of Disinformation

False information generally can be placed in one of three categories: disinformation, misinformation, or malinformation (Wardle and Derakhshan 2017). Deepfakes fit best in the disinformation category, which encompasses "manipulated content," "imposter content," and "fabricated content" (Wardle and Derakhshan 2017: 5). Disinformation can be seen as "intentional behavior that purposively misleads" (Chadwick et al. 2018: 4257). In contrast, misinformation differs from disinformation in the sense that the former does not imply the intention to deceive (Jack 2018), and malinformation is different from disinformation in that malinformation requires a (slim) factual basis.

Often, disinformation is meant to achieve some political goal. Actors behind disinformation can be domestic actors as well as foreign political actors. Legitimate domestic political actors can use illegitimate means such as disinformation to further their goals.

Foreign actors may try to intervene in domestic debates by injecting lies and conflict in the public sphere (Asmolov 2018; Bradshaw and Howard 2018; Lukito 2019; Xia et al. 2019). Foreign actors may even try to confuse citizens to a point where they become cynical and suspicious of legitimate information and legitimate institutions (Arendt 1951; see also Vaccari and Chadwick 2020). Therefore, disinformation is increasingly regarded a matter of (inter)national security (see, for example, Atlantic Council 2019; European Commission 2019; Metodieva 2018).

Literature about disinformation is growing rapidly, but there is still a "substantive research gap" about its effects (Tucker et al. 2018: 57). Existing literature of paints a nuanced picture. Guess et al. (2018) found that the effects of false news articles on

citizens' political attitudes are likely dampened because only a specific small group of citizens (the people with the most conservative online media diets) is exposed to misinformation. Similarly, Bail et al. (2020: 1) found no evidence for the idea that Russian trolls affected Americans' political attitudes: those engaging with the Russian trolls "were already highly polarized." These studies occurred in a very specific context (the U.S. context), focused on specific modes of disinformation (false news stories and Russian trolling on Twitter), and in specific time periods (October 7 to November 14, 2016; and October and November, 2017).

These studies offer insights into the limits of online disinformation campaigns. We argue that deepfake disinformation could be more impactful than Twitter trolling and false news stories. Vaccari and Chadwick (2020) found in a U.K.-based study that deepfakes confuse people about what is real, and consequently reduce trust in news on social media. Zimmermann and Kohring (2020) found that the less people trust news media, the more likely they are to fall for disinformation, which in turn can affect their voting behavior. However, the latter study did not focus on deepfake disinformation, and Vaccari and Chadwick (2020) did not measure how deepfakes affect political attitudes.

Moreover, the potential impact of deepfakes may be amplified by microtargeting. After all, the reason for Bail et al.'s (2020) conclusion of limited effects was not that the efforts in itself were ineffective, but that they were essentially targeted at the wrong, already polarized, audience.

The Amplifying Role of PMT

The Trump-electorate example mentioned in the introduction illustrates how sending several *different* deepfakes to *different* voters could be a way to amplify the effect of a deepfake. One could argue that by using microtargeting techniques, the actor spreading the deepfake could reach only those people who are perceived as susceptible to the specific disinformation served by the deepfake and, as a result, are most likely to alter their attitudes because of the disinformation. The people who are unsusceptible to one specific deepfake message, however, are potentially susceptible to *other* disinformation messages tailored to them personally. PMT is the instrument that allows deepfake producers to send the "right" deepfake to the "right" person.

Our expectations about the potential amplifying role of PMT in deepfake disinformation campaigns are informed by two contrasting theoretical perspectives. First, one could argue that tailored deepfakes are perceived to be more relevant, and, thus, are more likely to be scrutinized by the receiver (which may amplify the effects of the deepfake). Second, one could also argue that a tailored deepfake would cause motivated reasoning: Confronted with incongruent information, the receiver reasons this incongruence away to "maintain their extant values, identities and attitudes" (Slothuus and De Vreese 2010: 652). Motivated reasoning likely decreases the deepfake's effects.

Amplification. PMT allows political actors to expose people to tailored messages, which should amplify the effects of those tailored messages. The idea is that people

would only receive messages that are personally relevant. People who perceive a message as relevant and engage in greater message scrutiny than those who perceive a message as generic (Chang 2006; Wheeler et al. 2005). Scrutinized messages are more likely to influence citizens (Petty and Cacioppo 1986; Wheeler et al. 2008). A tailored deepfake is more likely to be perceived as relevant, which increases the chances of message scrutiny, which, in turn, increases the chances of influencing the citizen. Evidence of how PMT amplifies effects on political behavior is scarce. Endres (2019: 1) found that targeting Democratic voters on issues on which they hold similar positions with the Republican candidate "is associated with decreased support" for the Democratic candidate, and "increased abstention, and increased support for" the Republican candidate. Haenschen and Jennings (2019) found that microtargeted online ads could increase turnout conditionally under Millennial voters: only in competitive districts. Both studies were conducted in a U.S. context. Decreasing support for the citizen's "own" candidate and increasing support for the opponent, as demonstrated by Endres (2019), are arguably impressive in a polarized two-party context such as the United States (Abramowitz 2013; Webster and Abramowitz 2017). In a multiparty context, citizens are more likely to switch to parties within their "consideration set" rather than to a party outside of the consideration set or rather than abstaining altogether (Rekker and Rosema 2019).

Highly competitive districts such as those studied by Haenschen and Jennings (2019) are difficult to find in a multiparty context due to their (often) system of proportional representation. As such, it is difficult to see how the findings of Haenschen and Jennings (2019) can be generalized to multiparty contexts.

Present research on PMT focuses on legitimate forms of communication, but not on disinformation. To further explore what happens when people are exposed to a (microtargeted) deepfake, we turn to the literature on *gaffes* and scandals.

Gaffes and scandals. A gaffe is an "unintentional and/or inappropriate statement or behavior bringing into question his or her knowledge, wisdom, and/or politically acceptable attitudes that lead others to question a person's judgment, ability or character" (Frantzich 2012: 4). A prominent example of a gaffe is a recording of 2012 U.S. presidential candidate Mitt Romney, who was covertly filmed when discounting 47 percent of Americans as entitled, dependent victims who will vote for Obama no matter what (Sheinheit and Bogard 2016). According to Sheinheit and Bogard (2016), this gaffe correlated with a decrease in support for the 2012 U.S. Republican candidate. A different example is the "Dean scream," which correlated with the deterioration of the 2004 U.S. campaign of Howard Dean (Kreiss 2012).

There is little literature on gaffes. But the closely related field of political scandals is more mature. In the political scandal literature, scandals are causally associated with a decline of political attitudes toward politicians and political parties (Brody and Shapiro 1989; Chanley et al. 2000; Maier 2011).[2] Considering literature on gaffes and scandals, we expect that

Hypothesis 1a (H1a): A deepfake meant to discredit a political candidate negatively affects people's attitudes toward the depicted politician.
Hypothesis 1b (H1b): A deepfake meant to discredit a political candidate negatively affects people's attitudes toward the politician's party.

Considering the literature on the potential amplifying quality of PMT, we expect that

Hypothesis 1c (H1c): The effects of the deepfake are stronger for the microtargeted group than the untargeted group.

Inoculation. A deepfake is meant to cause an incongruence between expectation and perceived reality. Seeing a known political figure say or do something offensive or shocking can induce motivated reasoning if people identify with the same political party as the depicted politician. Partisanship plays an important role in activating motivated reasoning (Bolsen et al. 2014; Slothuus and De Vreese 2010). While partisan motivated reasoning can seem to decrease when presented with clear evidence (Parker-Stephen 2013), this type of reasoning has been shown to be highly adaptive in finding ways to still maintain one's predispositions, despite clear evidence (Bisgaard 2015). Indeed, corrections of misleading statements made by a political candidate have been shown to have no impact on evaluations of that candidate (Nyhan et al. 2019). Moreover, partisans that are confronted with information have been found to interpret this information along party lines (Lauderdale 2016), and in line with prior beliefs (Gaines et al. 2007). However tenacious, Redlawsk et al. (2010) demonstrated that there likely is an "affective tipping point" where citizens stop motivated reasoning. Potentially, microtargeted deepfakes can play a role in reaching that tipping point by confronting citizens with highly relevant discomforting information. Considering the literature on motivated reasoning, we formulate the following research question:

Research Question (RQ): How are the attitudes of supporters of the depicted politician's party affected by a deepfake meant to discredit the political candidate?

Method

Sample

The sample consisted of 278 participants. Participants were recruited by Kantar Lightspeed, a Dutch company specialized in recruitment for academic purposes, and paid a small amount for their participation. Data collection took place in October 2019. The mean year of birth in the sample was 1970 ($SD = 14.68$), and 54.7 percent was female. One percent completed only elementary school, 20 percent completed only high school, 35 percent completed only vocational school, and 43 percent held only a bachelor's degree or higher.[3] About 49.3 percent of the sample indicated to be Christian. We purposely oversampled Christians to get large enough groups, as the

Figure 1. Deepfake (L) and original video (R).

incidence rate of Christians in the Netherlands is only 31 percent (de Hart and van Houwelingen 2018).

Experimental Design

Christian religious identity serves as a blocking variable. After answering a filter question about religion, participants were first placed in either the Christian or non-Christian block and then randomly distributed into either the deepfake or the original video condition (see Table 1 for experimental design).

Independent Variables

Deepfake. The deepfake stimulus is a thirteen-second subtitled video showing a leading politician of Dutch Christian Democrats "CDA," which is one of the largest center-right parties in the Netherlands. The first eight seconds of the video calls for the attention of the participant and announces to the participants that they are going to see a short video of [name politician], politician of the CDA. The following five seconds are a manipulated video that makes it seem as if, in a television show, the politician jokes about Christ's crucifixion: "But, as Christ would say: don't crucify me for it."[4] It is a play on words, essentially making a joke out of Christ's crucifixion. It would move attitudes because the politician is a prominent Christian politician and the base of his party is to a large extent Christian. Figure 1 shows a screenshot from the deepfake and a screenshot from the original video.

Making the deepfake. First, we produced a fake speech with the politician's voice using a text-to-speech-based learning approach ("Tacotron2"). Then, we produced a fake "silent video" of the politician from a real video for which we modified the lip movements (frame by frame) to match the new fake speech using artificial intelligence (AI)-based lip synchronization techniques (Suwajanakorn et al. 2017).

To produce the silent video, we collected approximately twenty-five hours of publicly available videos of the politician. These videos where split into frames and used to train a deep learning model that predicts the mouth/lip shape of the politician from

a given input audio. From these videos, we also extracted approximately twelve hours of audio that we then transcribed and used to train another model that generates audio with the voice of the politician (the fake speech) from a given input text. We then used our first model to predict the lip movements corresponding the fake speech. Then, we reconstructed the lips and mouth texture for each frame and added the fake audio to produce the final video using the ffmpeg library.

Control condition. The stimulus in the control condition was the original, nonmanipulated, subtitled version of the video of the politician that was also used for the deepfake. The first eight seconds of the control video calls for the attention of the participant and announces to the participants that they are going to see a short video of [name politician], politician of the CDA. The following five seconds show the original version of the television interview that was manipulated in the experimental condition.

Microtargeted or untargeted appeal. The participants who indicated in a filter question that they were Christians were regarded as a group that received a microtargeted stimulus. Christianity served as a blocking variable, and was used to simulate microtargeting. The stimulus is about Christianity, and therefore catered to the personal interest of the Christian participants—but not the nonreligious participants. People who indicated in the filter question that they were not religious were regarded as a group that received an untargeted stimulus. Hence, we speak of a microtargeted and an untargeted appeal.

Being Christian. Participants who answered the question, "I consider myself a Christian" positively, were randomly placed in either the experimental or the control microtargeted group (Group 1 or 3). Participants who indicated to be religious, but not Christian were screened out. Participants who declared themselves to be nonreligious were randomly placed in either the experimental or control untargeted group (Group 2 or 4).

Degree of religiosity. The participants who considered themselves Christian were asked how often they pray at home, using a 7-point scale from the European Social Survey. We then dichotomized this variable into "heavy prayers": those who pray once a week or more often ($N = 81$; 11 percent prayed once a week, 13 percent more than once a week, 76 percent prayed every day) and those who prayed between at least once a month (but not once a week or more) or who prayed less ($N = 52$; 21 percent prayed at least once a month, 10 percent prayed only on religious holidays, 12 percent prayed once a year, 58 percent prayed never). Three participants declined to volunteer how often they prayed and were considered missing. We dichotomized this variable for two reasons. First, in reality, citizens who are being profiled are often classified as either belonging to one subgroup (1) or not (0) (e.g., Briggs Meyers and Myers 2010). Second, the conceptual difference of, for instance, praying once a week or every day is relatively small, but the difference between praying at least once a month and every week is much larger. Consequently, it is hard to imagine how someone who prays every week should receive a different deepfake than someone who prays several times

a week or every day. But it is easier to imagine that for someone who prays once a month, religion is not as central to their life as it is to someone who prays at least every week.

Voted CDA. This variable was used to register the potential occurrence of motivated reasoning. Participants were asked whether they, in the past five years, ever cast their votes for the CDA (in European, national, provincial, or local elections). Participants could answer either yes or no (217 no, sixty yes, one missing).

Dependent Variables

Attitude toward politician. This dependent variable was measured after the stimulus and entailed the attitude toward the politician. The nine-item measure is derived from Boomgaarden et al. (2016). On a 7-point scale, participants were asked to assess the politician's competence, experience, authenticity, corruptness, determination, fairness, responsibility, honesty, and friendliness (eigenvalue = 5.5; Cronbach's $\alpha = .93$). We added one item about "authenticity" to the original eight-item measure of Boomgaarden et al. (2016). Authenticity is a relevant part of the attitude when studying deepfake disinformation.

Attitude toward political party. This dependent variable is measured with one item, on a 11-point Likert scale. Participants were asked about their stance regarding eight political parties, including the CDA to which the politician belongs. The value 0 stood for negative and 10 stood for positive (see Seltzer and Zhang 2011).

Ethics. The experimental protocol has been approved by the ethical review board of our institution. We debriefed participants immediately after the experiment, and stressed among others that the video was manipulated and that the politician in reality never made the Christ remark and that he likely never would. We also informed the participants about the Christian roots of the CDA and linked to the values page of the CDA Web site. Moreover, our experiment took place in a controlled environment and not during an election.

Procedure. Participants were contacted by Kantar Lightspeed. Kantar Lightspeed sampled from a nationally representative sample. They oversampled Christians. They used noninterlocking quotas to get enough Christians but also keep the sample representative on gender, age, and education. Before participants started with the online survey experiment, they were informed and asked for their consent. In the first part of the survey, participants were asked about their religiosity and then sorted into a group or screened out. After completing the survey, the participants were debriefed about the real purpose of the study. Participants in the experimental condition were explained that what they saw was manipulated and that actually the politician never has and likely never will make such a remark about Christ. Participants saw information about the ideology and the Christian fundament of the CDA and were offered a link to CDA's policy positions if they wanted to read more.

Manipulation Check

Credibility. We measured the degree to which participants found the deepfake credible with two 7-point scale items ($r = .77$): a tweaked version of the scale used by Appelman and Sundar (2016: 76). *I find the video authentic* ($M_{\_deepfake} = 3.78$; $SD = 1.32$; $M_{\_control} = 4.14$; $SD = 1.27$), and *I find the video credible* ($M_{\_deepfake} = 3.62$; $SD = 1.52$; $M_{\_control} = 4.22$; $SD = 1.30$). When the participants scored lower than 4 on either the first or the second item (or both items), they were asked in an open question why they deemed the authenticity and credibility (somewhat) low. On the scale that consisted of both items combined and averaged, the deepfake ($M = 3.70$, $SD = 1.32$) was considered significantly less credible than the control video ($M = 4.18$, $SD = 1.23$): $t(274) = 3.08$, $p = .01$. However, upon inspecting the answers to the open question about why participants found the deepfake not so credible, we learned that many participants had given a low credibility score because they considered not the deepfake noncredible but rather had circumstantial issues, for example, "all politicians are noncredible and only care about their own interest and glory" or "because it is not in line with how I think and live my life." Respondents interpreted "credibility" in this experiment in a broad sense, not necessarily relating to the deepfake, but rather to the politician or politics in general. Of the eighty-four participants who found the deepfake (somewhat) noncredible, only twelve noted that the video was likely manipulated: For example, "The voice is not in line with the mouth movements and his movements look unnatural and manipulated" or "Because Christ would not have said 'don't crucify me for it.' But neither would [the politician]. I think this is what we would call 'Fake News.'"

Because the credibility of the deepfake was close to the credibility of the original video, and because only twelve of the participants actually recognized the deepfake as being a manipulated video, *and* because people can be influenced even if they are aware of the efforts to influence them (Evans and Park 2015), we decided to carry out our analyses with all participants, regardless of whether they perceived the deepfake as (somewhat non-)credible.[5]

Finally, at the end of the survey, we checked whether participants had ever heard of "deepfakes." Of the experimental group, almost 75 percent did never heard of deepfakes, 20 percent had heard of them, and 4 percent did not know. To ascertain whether these three groups differed on the degree in which they recognized the deepfake as being manipulated, we used an analysis of variance (ANOVA) and found no significant differences between the groups: $F(1, 128) = 3.03$, $p = .08$.

Scrutiny. Scrutiny was measured using four 7-point scale items from Wheeler et al. (2005). We dropped one item to improve scale reliability. The remaining three items were "to what extent did you watch the video attentively?" "To what extent did you think deeply about the content of the video?" and "How much effort did you put into understanding the content of the video?" The three items were combined and averaged (eigenvalue = 1.54; Cronbach's $\alpha = .79$).

The experimental ($M = 4.41$, $SD = 1.25$, $N = 143$) and the control group ($M = 4.53$, $SD = 1.27$, $N = 133$) did not differ significantly on the degree to which they

scrutinized the stimulus, $t(274) = 0.74$, $p = .23$. However, comparing "heavy prayers" in the experimental group ($M = 4.90$, $SD = 1.08$, $N = 41$) with "light prayers" in the same group ($M = 4.18$, $SD = 1.24$, $N = 26$) did yield a significant difference: $t(65) = -2.51$, $p = .001$. A comparison between "heavy prayers" and nonreligious participants ($M = 4.22$, $SD = 1.28$, $N = 74$) showed that "heavy prayers" also scrutinized the message significantly more elaborate than the nonreligious participants, $t(113) = -2.91$, $p = .002$. However, the heavy prayers in the control group did not score significantly different on scrutiny ($M = 4.85$, $SD = 1.12$, $N = 40$) from the heavy prayers in the experimental group, $t(79) = -0.21$, $p = .42$. This means that heavy prayers scrutinized the messages elaborately, suggesting that a message from the Christian politician is considered especially relevant by the group of heavy prayers.

Power

For H1a and H1b, we compare a group of 144 participants with a group of 133 participants. We conducted an a priori power analysis by using G*Power (Faul et al. 2007) with a significance level of $\alpha = .05$, a moderate effect size of $d = .3$, and a statistical power of $(1 - \beta) = 0.80$ (Cohen 1988). For the power analysis, we have followed all the steps as outlined by Perugini et al. (2018). This revealed an estimated sample size of 278. This means we fall short one observation. H1c is predicated on a small sample. Not finding an effect would come with a relatively high chance of making a type II error. Finding an effect, however, would not suggest a type I error.

Randomization Check

A randomization check showed no significant differences between the experimental condition and the control condition regarding year of birth, $t(263) = 1.01$, $p = .31$; gender, $t(276) = 0.07$, $p = .95$; and education, $t(275) = 1.01$, $p = .31$. Looking closer at the four conditions, the randomization check showed no significant differences between the four groups regarding year of birth, $F(3, 261) = 0.51$, $p = .68$; gender, $F(3, 274) = 0.92$, $p = .43$; and education, $F(3, 273) = 0.55$, $p = .65$.

Results

Main Analyses

Comparing the two groups that either saw the deepfake or the control video, we find that the experimental group held significantly worse attitudes toward the politician after seeing the deepfake ($M = 4.31$, $SD = 1.10$, $N = 144$) than the control group ($M = 4.62$, $SD = 0.96$, $N = 133$): $t(275) = 2.48$, $p = .01$. This means *H1a* is supported.

Focusing on the attitudes toward the political party of the depicted politician (CDA), the difference between the experimental group ($M = 4.46$, $SD = 2.26$, $N = 144$) and the control group ($M = 4.76$, $SD = 2.38$, $N = 133$) is nonsignificant: $t(275) = 1.08$, $p = .14$.[6] This means that *H1b* is not supported.

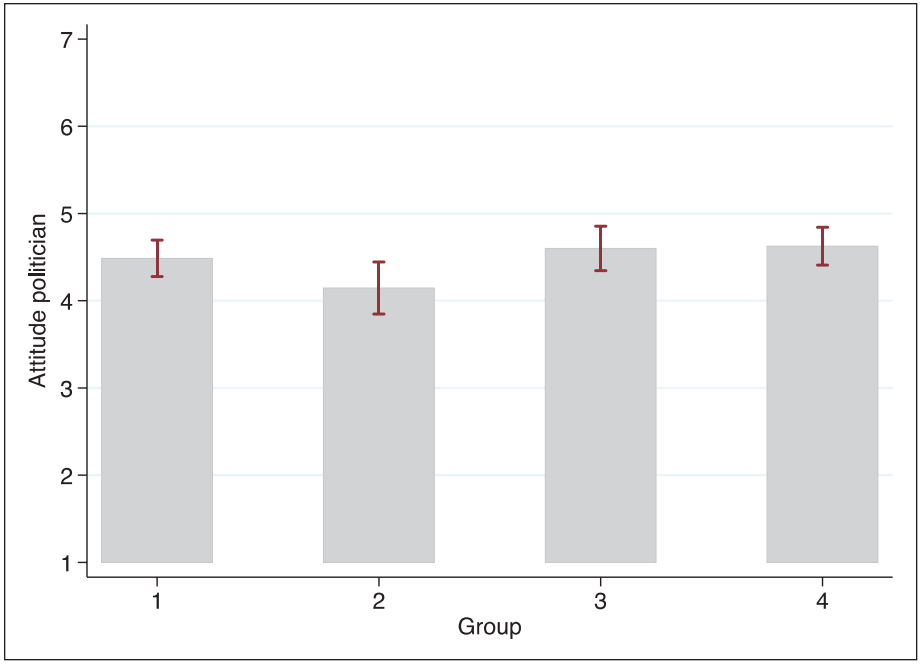


Figure 2. CI plot attitude toward politician after exposure, per group.
Note. Group 1 = Christian experimental (95% CI = [4.28, 4.70]), Group 2 = nonreligious experimental (95% CI = [3.85, 4.44]), Group 3 = Christian control (95% CI = [4.35, 4.85]), Group 4 = nonreligious control (95% CI = [4.41, 4.84]). CI = confidence interval.

Zooming in and comparing the four groups, using an ANOVA, we find a significant difference between the four groups regarding attitude toward the politician: $F(3, 273) = 3.21$, $p = .02$.[7] A Bonferroni post hoc comparison (see Figure 2) showed that nonreligious experimental group scored significantly ($p = .04$) lower ($M = 4.15$, $SD = 1.28$, $N = 73$) than the nonreligious control group ($M = 4.62$, $SD = 0.89$, $N = 67$). An ANOVA yielded no meaningful significant differences between the four groups regarding their attitudes toward the politician's party CDA: $F(3, 273) = 4.72$, $p = .001$. Upon closer inspection, using a post hoc Bonferroni comparison, the difference was between Experimental Group 1 and Experimental Group 2.

Further testing H1c (*The effects of the deepfake are stronger for the microtargeted group than the untargeted group*), we zoom in on the group of voters that would be the most vulnerable targets of this specific piece of disinformation: Christians who are very religious, and who have voted for CDA in the past. To see whether the deepfake's effects are stronger for this specific subsample in comparison with the Christians in the control group who are also very religious and have voted for CDA in the past, we conducted a *t* test to compare the scores of the experimental group of such participants ($N = 14$) with their counterparts in the control condition ($N = 10$). Figure 3 shows the mean attitude scores toward the depicted politician for this specific group.

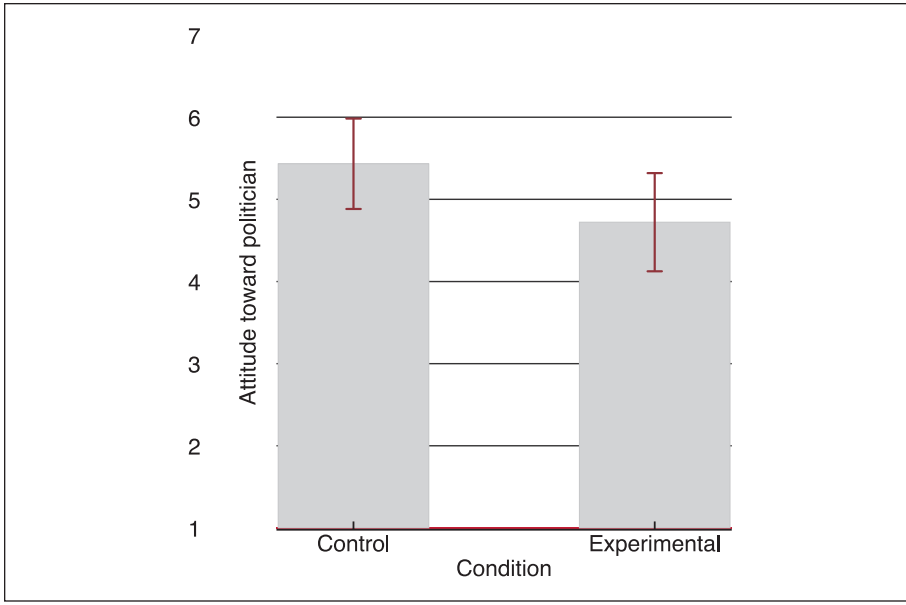


Figure 3. Hypothesis 1c.
Note. CI plot comparing scores on attitude toward the politician for very religious Christians who voted CDA in the past: $t(22) = 1.84$, $p = .04$. Control group: $M = 5.43$, $SD = 0.77$, 95% CI = [4.88, 5.98], $N = 10$. Experimental group: $M = 4.72$, $SD = 1.04$, 95% CI = [4.12, 5.32], $N = 14$. CI = confidence interval; CDA = Dutch Christian Democrats.

Figure 4 displays the same subsample's mean attitudes toward the political party and shows that the mean scores of the experimental condition are significantly lower than the mean scores of the control condition. Both these findings support *H1c*, but only on a granular level.

Small subsamples. For both attitude toward the politician and attitude toward the party, a Kolmogorov–Smirnov test and a Shapiro–Wilk test were conducted. Table 2 shows that both conditions and both variables were distributed normally, making the *t* test a robust test even for these very small subsamples.

Motivated reasoning. To answer the research question (*How are the attitudes of supporters of the depicted politician's party affected by a deepfake meant to discredit the political candidate?*), we looked at whether CDA voters in the experimental condition held different attitudes than CDA voters in the control condition. We did the same for the non-CDA voters in both conditions. CDA voters in both conditions did not score significantly different, while non-CDA voters did. This holds for attitude toward the politician (Table 3) as well as attitude toward his political party[8] (Table 4). This suggests the occurrence of motivated reasoning, which inoculated the CDA voters in the

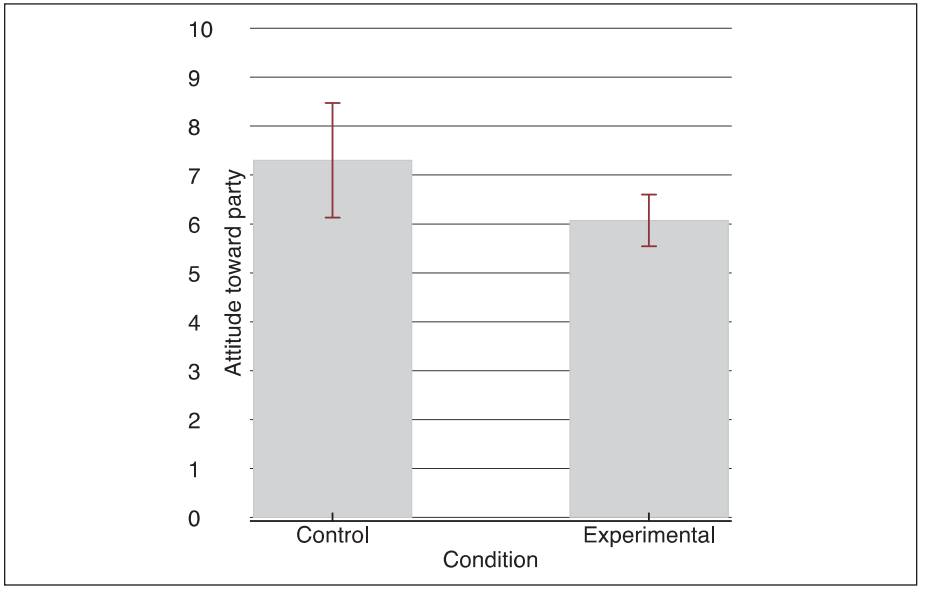


Figure 4. Hypothesis 1c.
Note. CI plot comparing scores on attitude toward the political party (CDA) for very religious Christians who voted CDA in the past: $t(22) = 2.35$, $p = .01$. Control group: $M = 7.30$, $SD = 1.64$, 95% CI = [6.13, 8.47], $N = 10$. Experimental group: $M = 6.07$, $SD = 0.92$, 95% CI = [5.54, 6.47], $N = 14$. CI = confidence interval; CDA = Dutch Christian Democrats.

experimental group from negative effects of the deepfake (answering the RQ). See the Supplementary Information file for a comparison of the control and experimental group per moderator for both dependent variables.

Discussion

In this study, we set out to investigate the extent to which a (microtargeted) deepfake meant to discredit a politician can affect citizens' attitudes toward that politician and his party.[9] This experiment indicates that indeed it is possible to stage a political scandal with a deepfake. The negative attitudinal consequences toward the politician and the party that are found in the scandal literature (Brody and Shapiro 1989; Chanley et al. 2000; Maier 2011) are found in this study as well. While especially the attitude toward the politician is directly affected by the deepfake, attitudes toward the politician's party are only conditionally affected. As such, our findings differ from Guess et al. (2018) and from Bail et al. (2020), who found no effects of disinformation on political behavior and attitudes. The current study provides a first careful support for the idea that indeed deepfakes are a more powerful mode of disinformation in comparison with the false news stories studied by Guess et al. (2018) and the Russian Twitter trolls studied by Bail et al. (2020).

Table 1. Experimental Design.

	Deepfake	Original
Christian	Group 1 ($N = 72$)	Group 3 ($N = 66$)
Non-Christian	Group 2 ($N = 73$)	Group 4 ($N = 67$)

Table 2. Tests for normal distribution attitude toward politician in experimental group ($n = 14$; Kolmogorov-Smirnov $p = .38$; Shapiro-Wilk $p = .12$) and control group ($N = 10$; Kolmogorov-Smirnov $p = .52$; Shapiro-Wilk $p = .17$), and for normal distribution attitude toward political party experimental group ($N = 14$; Kolmogorov-Smirnov $p = .38$; Shapiro-Wilk $p = 1.00$) and control group ($N = 10$; Kolmogorov Smirnov $p = .52$; Shapiro-Wilk = .65).

Test for Normal Distribution	Adj χ^2_ Experimental	Adj χ^2_ Control	W_Experimental	W_Control
Attitude politician				
Kolmogorov–Smirnov	1.95	1.31		
Shapiro–Wilk			.90	.89
Attitude party				
Kolmogorov-Smirnov	1.95	1.31		
Shapiro–Wilk			.99	.95

Table 3. Results of *t*-Test Comparing CDA Voters in Experimental Group ($N = 36$) with CDA Voters in Control Group ($N = 24$): $t(58) = 0.25$, $p = .40$, and Results of *t* test Comparing Non-CDA Voters in Experimental Group ($N = 107$) with CDA Voters in Control Group ($N = 109$): $t(214) = 3.05$, $p = .001$ on Attitudes toward the Politician.

Attitude Politician	*M_exp* (*SD*)	*M_control* (*SD*)	*t*	*p*
Voted CDA (yes)	4.89 (0.97)	4.95 (0.85)	0.25	.40
Voted CDA (no)	4.12 (1.09)	4.54 (0.97)	3.05	.001

Note. CDA = Dutch Christian Democrats.

Table 4. Results of *t*-Test Comparing CDA Voters in Experimental Group ($N = 36$) with CDA Voters in Control Group ($N = 24$): $t(58) = 0.88$, $p = 19$, and Results of *t*-Test Comparing Non-CDA Voters in Experimental Group ($N = 107$) with CDA Voters in Control Group ($N = 109$): $t(214) = 1.66$, $p = .05$ on Attitudes toward the Party (CDA).

Attitude Party	*M_Exp* (*SD*)	*M_Control* (*SD*)	*t*	*p*
Voted CDA (yes)	6.28 (1.54)	6.63 (1.44)	0.88	.19
Voted CDA (no)	3.84 (2.14)	4.35 (2.35)	1.66	.05

Note. CDA = Dutch Christian Democrats.

Amplification

We theorized that PMT could function as an amplifier that would make the deepfake more effective. Our findings suggest that PMT can indeed amplify the effects of the deepfake, but only for a much smaller portion of the sample than we expected. In particular, it turns out that the group that one needs to target to are *the very religious* Christian *CDA voters*, instead of *all* Christians. But why would other Christians who have not voted CDA be less susceptible, even though they should be equally discomforted by the deepfake? The explanation might lie in the multiparty system. There are two other, more orthodox Christian parties in the Netherlands, ChristenUnie (CU), and Staatkundig Gereformeerde Partij (SGP), and many heavily religious people may consider CDA as too distant from "pure" Christian views anyway (also see the classification of the Christian consideration set by Rekker and Rosema 2019). Next to that, the less religious participants would be less susceptible to the deepfake, because their Christianity is less central in their lives. Consequently, PMT should be based on more than simply "belonging to one group," but rather on the intersection of two or more characteristics. In this case, being both heavily religious *and* voted CDA.

In sum, concordant with Endres (2019), we found that partisans can be negatively affected by a microtargeted message regarding their own candidate. In contrast with Endres (2019) and Matthes and Marquart (2015), we find that a message meant to be incongruent with the opinions of the receiver can have a significant and substantial negative attitudinal effect.

Inoculation. For part of the "mistargeted group" (the CDA voters who were not heavy prayers), it appears that motivated reasoning inoculated them from any negative effects of the deepfake (see Kahan et al. 2017 for more information on measuring motivated reasoning). Motivated reasoning is sometimes considered negative in the face of truthful information—Richey (2012: 511) even reflected on whether motivated reasoning was the "death knell of deliberative democracy." But inoculation through motivated reasoning can be positive when facing disinformation meant to be incongruent with people's prior beliefs.

It may be a reassuring thought that the supporters of the politician who was negatively depicted in the deepfake are to some extent protected from deepfake manipulation by their tendency to engage in motivated reasoning. Even when facing clear evidence of the incongruency occurring, the partisans indeed did not hold worse attitudes than their counterparts in the control condition did, while the nonpartisan groups did (in line with Bisgaard 2015). Still, if, for instance due to microtargeting, the message is personally relevant *and* (therefore) discomforting enough, a potential affective tipping point may be reached instantly, and the motivated reasoning will cease (see Redlawsk et al. 2010).

Credibility. The open question about why participants did not find the deepfake too credible made it evident that only a small fraction of the sample recognized the deepfake as a manipulated video. Moreover, the open question showed that the credibility scale did not measure the credibility of the deepfake in a narrow sense, but rather in a

broader sense where participants interpreted the credibility items as how credible they find the depicted politician or politics in general. One participant, for instance, explained their low credibility score as follows: "Nothing in politics is credible." Someone else explained, "I have trouble taking politics in the Netherlands seriously."

Frankly, the deepfake can be improved upon. The mouth movement of the politician sometimes reminds of a dummy used by a ventriloquist, the voice is acceptable but not good, and the video is only five seconds. But even with these points of improvements, almost no participant raised concerns with the veracity of the video itself. This can be partly attributed to the novelty of the technique, but is also because seeing and hearing a person say something can be so realistic.

Unexpected effects. The potential threat of (microtargeted) deepfakes lies in their use by a malicious political actor with the desire to achieve some illegitimate political goal. Similar to Maarek (2003), who attributed the shocking loss of a French presidential candidate to a too professional campaign, or similar to Adams et al. (1986) who found that an antinuclear warfare television broadcast actually increased American viewers' support for then-president Reagan instead of vice versa, our experiment shows that pursuing a goal with microtargeted deepfakes may also come with some unforeseen outcomes. For instance, not the general group of Christians who saw the deepfake held significantly worse attitudes toward the politician in comparison with the control group, but rather the non-Christians who saw the deepfake did. Moreover, we found that, after exposure, this experimental group of non-Christians held significant and substantial better attitudes toward populist party Forum voor Democratie ($M = 4.27$, $SD = 3.36$, $N = 73$) in comparison with their counterparts in the control group ($M = 3.03$, $SD = 3.12$, $N = 67$): $t(138) = 2.27$, $p = .01$. These unforeseen outcomes suggest that impacting a dynamic and chaotic event that is an election in a controlled and predictable way is challenging, if not impossible.

Should we worry about (microtargeted) deepfakes? Yes, we should worry, but more about deepfakes in general than about *microtargeted* deepfakes. Making a deepfake requires a sizable amount of work, but technology progresses quickly. Having the right tools (advanced video card, adequate computing power, quality training data) makes it easier to produce a quality deepfake.

For now, the limited, but significant main effect of our imperfect deepfake on the attitudes of the general sample is more or less aligned with the idea of "minimal effects" of political communication (see Bennett and Iyengar 2008). The idea of minimal effects in political communication was later substantiated by Kalla and Broockman (2018), who in an important meta-analysis estimated that the persuasive effects of campaign contact and advertising on American voters were zero. They also found that identifying and persuading specific subgroups of persuadable voters appears to be a successful persuasive strategy. But "identifying cross- pressured persuadable voters requires much more effort than simply applying much-ballyhooed 'big data'" (p. 2). Similarly, while that meta-analysis is not easily generalizable to a non-U.S., multiparty, less-affectively polarized context, the current study also finds that making

several, or even hundreds or thousands of *tailored* deepfakes is for now a bridge too far. Not necessarily because of technical hurdles, but rather because microtargeting "correctly" is challenging. Even in this experiment, we correctly microtargeted only fourteen people in our sample.

Over time, it could get easier to get accurate perceptions of what characteristics make a voter group susceptible to a tailored deepfake. But for a malicious actor operating present day, taking a less subtle approach and spreading one discomforting deepfake would be the most realistic option. Worrisomely, a better quality and longer deepfake, repeated exposure and distribution in a dynamic real-life context could easily produce larger effects. Furthermore, the notion that the main barrier protecting the electorate from large persuasive effects is the difficulty to microtarget a deepfake correctly, is hardly comforting. But for now, as Karpf (2019) has argued, the largest threat of present-day disinformation does not lie in individual-level effects, but rather in the *belief* that individuals can be swayed so easily:

> If the public is made up of easily-duped partisans, then there is no need to take difficult votes. If the public simply doesn't pay attention to policymaking, then there is no reason to sacrifice short-term partisan gains for the public good.

Directions for Future Research

Future research should map the effects of deepfakes, potentially by comparing deepfakes with differing levels of quality, and different degrees of shock the deepfakes induce, and by comparing source effects. More importantly, future research should map ways to counter deepfakes' effects. A regulatory focus should be directed against this potential new frontier of disinformation warfare. The surprising low number of participants who recognized the deepfake as being manipulated is a clear sign that public awareness and knowledge of deepfakes should improve. But informing the public about deepfakes must not lead to cynicism in citizens and in politicians.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Tom Dobber https://orcid.org/0000-0002-6657-4037

Damian Trilling https://orcid.org/0000-0002-2586-0352

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Buzzfeed.com. 2018. https://www.buzzfeed.com/raigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed.
2. Much of the scandal literature focuses on corruption. This study does not focus on corruption, but rather on a politician's character. On a more abstract level, one could argue that voters' responses to a corrupt politician are the result of a (negative) judgment of the corrupt politician's character as well.
3. Does not add up to 100 due to rounding.
4. In Dutch: "Maar zoals Christus zou zeggen, pin mij er niet op vast."
5. We regard this as a conservative approach. As a robustness check, we have also analyzed the data with only the participants who had a credibility score of >3. The main findings do not change. Meaningful differences are discussed in the footnote in the results section.
6. In the credibility >3 sample, the *M_exp* (*SD*) was 4.90 (2.02), and the *M_control* (*SD*) was 5.30 (1.98), $t = 1.47$, $p = .07$.
7. In the credibility >3 sample, this difference was not significant.
8. For the credibility >3 sample, this difference was nonsignificant: *M_exp* (*SD*) = 6.17 (1.60), *M_control* (*SD*) = 6.73 (1.39), $t = 1.32$, $p = .10$.
9. Did we register the effect of a deepfake or the effect of a gaffe? There is one vital difference between a deepfake and a gaffe: a gaffe actually happened, while a deepfake is not real. This vital difference shapes a completely different context, of which we have virtually no empirical knowledge yet. For example, an issue following this difference is credibility. Potentially, although the gaffe literature is rather thin, credibility for gaffes is much higher simply because they actually happened—in other words, there is less "noise" that may be induced by the creator of a deepfake. Second, a deepfake differs from a gaffe in that a deepfake gives the malicious political actor control. The deepfake producer controls (1) who to depict, (2) what that person says, and (3) to what citizen groups the deepfake can be tailored. A gaffe gives the malicious actor much less control, limiting the strategic value of gaffes. Gaffes and deepfakes are comparable, but only up to a certain level. We believe we indeed did measure the effects of a deepfake because the stimulus was a deepfake, and not a gaffe.

References

Abramowitz, Alan I. 2013. "The Electoral Roots of America's Dysfunctional Government." *Presidential Studies Quarterly* 43 (4): 709–31.

Adams, William C., Dennis J. Smith, Allison Salzman, Ralph Crossen, Scott Hieber, Tom Naccarato, William Vantine, and Nine Weisbroth. 1986. "Before and after the Day after: The Unexpected Results of a Televised Drama." *Political Communication* 3 (3): 191–213.

Afchar, Darius, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. "MesoNet: A Compact Facial Video Forgery Detection Network." 10th IEEE International Workshop on Information Forensics and Security, WIFS December 2018, 1–7.

Agarwal, Sakshi, and Lav R. Varshney. 2019. "Limits of Deepfake Detection: A Robust Estimation Viewpoint." http://arxiv.org/abs/1905.03493.

Appelman, Alyssa, and Shyam Sundar. 2016. "Measuring Message Credibility: Construction and Validation of an Exclusive Scale." *Journalism & Mass Communication Quarterly* 93 (1): 59–79.

Arendt, Hannah. 1951. *The Origins of Totalitarianism*. Orlando, Florida: Harcourt.

Asmolov, Gregory. 2018. "The Disconnective Power of Disinformation Campaigns." *Journal of International Affairs* 71 (15): 69–77.

Atlantic Council. 2019. "Democratic Defense against Disinformation 2.0." https://www.brookings.edu/research/democratic-defense-against-disinformation-2-0/.

Bail, Christopher A., Brian Guay, Emily Maloney, Aidan Combs, D. Sunshine Hillygus, Friedolin Merhout, Deen Freelon, and Alexander Volfovsky. 2020. "Assessing the Russian Internet Research Agency's Impact on the Political Attitudes and Behaviors of American Twitter Users in Late 2017." *Proceedings of the National Academy of Sciences of the United States of America* 117:243–50.

Baldwin-Philippi, Jessica. 2019. "Data Campaigning: Between Empirics and Assumptions." *Internet Policy Review* 8 (4): 18. https://doi.org/10.14763/2019.4.1437.

Bennett, W. Lance, and Shanto Iyengar. 2008. "A New Era of Minimal Effects? The Changing Foundations of Political Communication." *Journal of Communication* 58 (4): 707–31.

Bennett, W. Lance, and Steven Livingston. 2018. "The Disinformation Order: Disruptive Communication and the Decline of Democratic Institutions." *European Journal of Communication* 33 (2): 122–39.

Bisgaard, Martin. 2015. "Bias Will Find a Way: Economic Perceptions, Attributions of Blame, and Partisan-Motivated Reasoning during Crisis." *Journal of Politics* 77 (3): 849–60.

Bolsen, Toby, James N. Druckman, and Fay Lomax Cook. 2014. "The Influence of Partisan Motivated Reasoning on Public Opinion." *Political Behavior* 36 (2): 235–62.

Boomgaarden, Hajo G., Mark Boukes, and Aurora Iorgoveanu. 2016. "Image versus Text: How Newspaper Reports Affect Evaluations of Political Candidates." *International Journal of Communication* 10:2529–55.

Borgesius, Frederik J. Zuiderveen, Judith Möller, Sanne Kruikemeier, Ronan Fathaigh, Kristina Irion, Tom Dobber, Balazs Bodo, and Claes de Vreese. 2018. "Online Political Microtargeting: Promises and Threats for Democracy." *Utrecht Law Review* 14 (1): 82–96.

Bradshaw, Samantha, and Philip N. Howard. 2018. "The Global Organization of Social Media Disinformation Campaigns." *Journal for Internal Affairs* 71:23–31. https://jia.sipa.columbia.edu/.

Briggs Meyers, Isabel, and Peter Myers. 2010. *Gifts Differing: Understanding Personality Type*. London: John Murray Press.

Brody, Richard A., and Catherine R. Shapiro. 1989. "Policy Failure and Public Support: The Iran-Contra Affair and Public Assessment of President Reagan." *Political Behavior* 11 (4): 353–69.

Chadwick, Andrew, Cristian Vaccari, and Ben O'Loughlin. 2018. "Do Tabloids Poison the Well of Social Media? Explaining Democratically Dysfunctional News Sharing." *New Media and Society* 20 (11): 4255–74.

Chang, Chingching. 2006. "Seeing the Small Picture: AD-Self versus Ad-Culture Congruency in International Advertising." *Journal of Business and Psychology* 20 (3): 445–65.

Chanley, Virginia A., Thomas J. Rudolph, and Wendy M. Rahn. 2000. "The Origins and Consequences of Public Trust in Government: A Time Series Analysis." *The Public Opinion Quarterly* 64 (3): 239–56.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale: Lawrence Erlbaum.

de Hart, Joep, and Pepijn van Houwelingen. 2018. "*Christenen in Nederland*." https://www.scp.nl/publicaties/publicaties/2018/12/19/christenen-in-nederland

Dobber, Tom, Damian Trilling, Natali Helberger, and Claes H. De Vreese. 2017. "Two Crates of Beer and 40 Pizzas: The Adoption of Innovative Political Behavioural Targeting Techniques." *Internet Policy Review* 6 (4): 1–25. https://policyreview.info/node/777/pdf

Dommett, Katharine. 2019. "Data-Driven Political Campaigns in Practice: Understanding and Regulating Diverse Data-Driven Campaigns." *Internet Policy Review* 8 (4). doi:10.14763/2019.4.1432

Dommett, Katharine, Luke Temple, and Patrick Seyd. 2020. "Dynamics of Intra-Party Organisation in the Digital Age: A Grassroots Analysis of Digital Adoption." *Parliamentary Affairs* 1–20. doi:10.1093/pa/gsaa007.

Endres, Kyle. 2019. "Targeted Issue Messages and Voting Behavior." *American Politics Research* 48:317–28

European Commission. 2019. *"Twentieth Progress Report towards an Effective and Genuine Security Union EN*." https://ec.europa.eu/home-affairs/sites/homeaffairs/files/what-we-do/policies/european-agenda-security/20191030_com-2019-552-security-union-update-20_en.pdf

Evans, Nathaniel J., and Dooyeon Park. 2015. "Rethinking the Persuasion Knowledge Model: Schematic Antecedents and Associative Outcomes of Persuasion Knowledge Activation for Covert Advertising." *Journal of Current Issues & Research in Advertising* 36 (2): 157–76.

Faul, Franz, E. Erdfelder, A. G. Lang, and A. Buchner. 2007. "G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences." *Behavioral Research Methods* 39 (2): 175–91. doi:10.3758/BF03193146.

Flynn, D. J., Brendan Nyhan, and Jason Reifler. 2017. "The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs about Politics." *Political Psychology* 38:127–50.

Frantzich, Stephen. 2012. *O.O.P.S.: Observing Our Politicians Stumble: The Worst Candidate Gaffes and Recoveries in Presidential Campaigns*. Santa Barbara: Praeger.

Gaines, Brian J., James H. Kuklinski, Paul J. Quirk, Buddy Peyton, and Jay Verkuilen. 2007. "Same Facts, Different Interpretations: Partisan Motivation and Opinion on Iraq." *Journal of Politics* 69 (4): 957–74.

Guess, Andrew, Brendan Nyhan, and Jason Reifler. 2018. "Selective Exposure to Misinformation: Evidence from the Consumption of Fake News during the 2016 U. S. Presidential Campaign." European Research Council. https://www.dartmouth.edu/~nyhan/fake-news-2016.pdf.

Haenschen, Katherine, and Jay Jennings. 2019. "Mobilizing Millennial Voters with Targeted Internet Advertisements: A Field Experiment." *Political Communication* 36:357–75. https://doi.org/10.1080/10584609.2018.1548530

Jack, Caroline. 2018. "Lexicon of Lies: Terms for Problematic Information*." Data & Society Research Institute*. https://datasociety.net/output/lexicon-of-lies/.

Kahan, Dan, Ellen Peters, Erica Cantrell Dawson, and Paul Slovic. 2017. "Motivated Numeracy and Enlightened Self-Government." *Behavioural Public Policy* 1 (1): 54–86. doi:10.1017/bpp.2016.2.

Kalla, Joshua L., and David E. Broockman. 2018. "The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments." *American Political Science Review* 112 (1): 148–66.

Karpf, David. 2019. "On Digital Disinformation and Democratic Myths*." MediaWell SSRC*. https://mediawell.ssrc.org/expert-reflections/on-digital-disinformation-and-democratic-myths/.

Kreiss, Daniel. 2012. *Taking Our Country Back: The Crafting of Networked Politics from Howard Dean to Barack Obama*. New York: Oxford University Press.

Kreiss, Daniel. 2016. *Prototype Politics: Technology-Intensive Campaigning and the Data of Democracy*. New York: Oxford University Press.

Lauderdale, Benjamin E. 2016. “Partisan Disagreements Arising from Rationalization of Common Information.” *Political Science Research and Methods* 4 (3): 477–92.

Li, Yuezun, and Siwei Lyu. 2018. “Exposing DeepFake Videos By Detecting Face Warping Artifacts.” ArXiv Preprint ArXiv:1811.00656v3. http://arxiv.org/abs/1811.00656.

Lukito, Josephine. 2019. “Coordinating a Multi-Platform Disinformation Campaign: Internet Research Agency Activity on Three U.S. Social Media Platforms, 2015 to 2017.” *Political Communication* 37:238–55. doi:10.1080/10584609.2019.1661889.

Maarek, Philippe J. 2003. “Political Communication and the Unexpected Outcome of the 2002 French Presidential Elections.” *Journal of Political Marketing* 2 (2): 13–24.

Maier, Jürgen. 2011. “The Impact of Political Scandals on Political Support: An Experimental Test of Two Theories.” *International Political Science Review* 32 (3): 283–302.

Matsumoto, Asuka. 2018. “Internet Campaigning in the US and Japan: Battles in Cyber Space.” *SAIS Review of International Affairs* 38 (1): 27–38. doi:10.1353/sais.2018.0003.

Matthes, Jörg, and Franziska Marquart. 2015. “A New Look at Campaign Advertising and Political Engagement: Exploring the Effects of Opinion-Congruent and -Incongruent Political Advertisements.” *Communication Research* 42 (1): 134–55.

Metodieva, Asya. 2018. “Disinformation as a Cyber Threat in the V4: Capabilities and Reactions to Russian Campaigns.*”* www.stratpol.sk.

Moura, Mauricio, and Melissa R. Michelson. 2017. “WhatsApp in Brazil : Mobilising Voters through Door-to-Door and Personal Messages.” *Internet Policy Review* 6 (4): 1–17.

Nyhan, Brendan, Ethan Porter, Jason Reifler, and Thomas J. Wood. 2019. “Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability.” *Political Behavior*. doi:10.1007/s11109-019-09528-x.

Parker-Stephen, Evan. 2013. “Tides of Disagreement: How Reality Facilitates (and Inhibits) Partisan Public Opinion.” *Journal of Politics* 75 (4): 1077–88.

Perugini, Marco, Marcello Gallucci, and Giulio Costantini. 2018. “A Practical Primer To Power Analysis for Simple Experimental Designs.” *International Review of Social Psychology* 31 (1): 1–23.

Petty, Richard E., and John T. Cacioppo. 1986. “The Elaboration Likelihood Model of Persuasion.” *Advances in Experimental Social Psychology* 19:123–205.

Redlawsk, David P., Andrew J. W. Civettini, and Karen M. Emmerson. 2010. “The Affective Tipping Point: Do Motivated Reasoners Ever ‘Get It’?” *Political Psychology* 31 (4): 563–93.

Rekker, Roderik, and Martin Rosema. 2019. “How (Often) Do Voters Change Their Consideration Sets?” *Electoral Studies* 57:284–93. doi:10.1016/j.electstud.2018.08.006.

Richey, Mason. 2012. “Motivated Reasoning in Political Information Processing: The Death Knell of Deliberative Democracy?” *Philosophy of the Social Sciences* 42 (4): 511–42.

Seltzer, Trent, and Weiwu Zhang. 2011. “Toward a Model of Political Organization-Public Relationships: Antecedent and Cultivation Strategy Influence on Citizens’ Relationships with Political Parties.” *Journal of Public Relations Research* 23 (1): 24–45.

Sheinheit, Ian, and Cynthia J. Bogard. 2016. “Authenticity and Carrier Agents: The Social Construction of Political Gaffes.” *Sociological Forum* 31 (4): 970–93.

Slothuus, Rune, and Claes H. De Vreese. 2010. “Political Parties, Motivated Reasoning, and Issue Framing Effects.” *Journal of Politics* 72 (3): 630–45.

Suwajanakorn, Supasorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. “Synthesizing Obama: Learning Lip Sync from Audio.” *ACM Transactions on Graphics* 36 (4): 1–12.

Tucker, Joshua A., Andrew Guess, Pablo Barberá, Christian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brenden Nyhan. 2018. “Social Media, Political Polarization,

and Political Disinformation: A Review of the Scientific Literature." *Hewlett Foundation*. doi:10.2139/ssrn.3144139.

Vaccari, Cristian, and Andrew Chadwick. 2020. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News." *Social Media & Society* 6:1–13. doi:10.1177/2056305120903408

Wardle, Claire, and Hossein Derakhshan. 2017. "Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*." Report to the Council of Europe*. https://www.rcmediafreedom.eu/Publications/Reports/Information-disorder-Toward-an-interdisciplinary-framework-for-research-and-policy-making.

Webster, Steven W., and Alan I. Abramowitz. 2017. "The Ideological Foundations of Affective Polarization in the U.S. Electorate." *American Politics Research* 45 (4): 621–47.

Wheeler, S. Christian, Kenneth G. DeMarree, and Richard E. Petty. 2008. "A Match Made in the Laboratory: Persuasion and Matches to Primed Traits and Stereotypes." *Journal of Experimental Social Psychology* 44 (4): 1035–47.

Wheeler, S. Christian, Richard E. Petty, and George Y. Bizer. 2005. "Self-Schema Matching and Attitude Change: Situational and Dispositional Determinants of Message Elaboration." *Journal of Consumer Research* 31 (4): 787–97.

Xia, Yiping, Josephine Lukito, Yini Zhang, Chris Wells, Sang Jung Kim, and Chau Tong. 2019. "Disinformation, Performed: Self-Presentation of a Russian IRA Account on Twitter." *Information Communication & Society* 22:1646–64.

Yang, Xin, Yuezun Li, and Siwei Lyu. 2019. "Exposing Deep Fakes Using Inconsistent Head Poses." ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing—Proceedings, May 2019, 8261–65.

Zimmermann, Fabian, and Matthias Kohring. 2020. "Mistrust, Disinforming News, and Vote Choice: A Panel Survey on the Origins and Consequences of Believing Disinformation in the 2017 German Parliamentary Election." *Political Communication* 37:215–37. doi:10.1080/10584609.2019.1686095

Author Biographies

Tom Dobber is a postdoctoral researcher at the Amsterdam School of Communication Research (ASCoR), University of Amsterdam. His research focuses on political microtargeting, disinformation and electoral pledges.

Nadia Metoui is a postdoctoral researcher at the Amsterdam School of Communication Research (ASCoR), University of Amsterdam. She is a computer scientist with a special interest in political microtargeting.

Damian Trilling is associate professor at the Amsterdam School of Communication Research (ASCoR), University of Amsterdam. He uses computational social science to study topics such as media use, social media, and big data.

Natali Helberger is university professor Law & Digital Technology and professor of Information Law at the Institute for Information Law, University of Amsterdam.

Claes de Vreese is faculty professor of Artificial Intelligence, Data & Democracy and professor of Political Communication, at the Amsterdam School of Communication Research (ASCoR), University of Amsterdam.