Data & Democracy

Political microtargeting: A threat to electoral integrity?

Dobber, T.

Publication date
2020
Document Version
Other version
License
Other

Chapter 5

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

Abstract

Deepfakes are perceived as a powerful form of disinformation. Though 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.

Introduction

So-called 'deepfakes', many argue, may be the next disinformation challenge 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 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 (Silverman, 2018). 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 & Howard, 2018).

Whether people indeed 'fall for' deepfakes is unclear, 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 deepfakes can have a detrimental societal impact, which is why deepfakes are worth the scientific scrutiny. For one, deepfakes can be realistic. 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 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* political microtargeting (PMT). In particular, we expect that the use of PMT techniques is an important amplifier of the effectiveness of deepfakes.

PMT is a relatively new technique used by political campaigns worldwide (Anstead, 2017; Kreiss, 2016; Dobber et al., 2017; Moura, 2017). It involves 1) the collection of personal data, 2) using those data to identify subgroups of people who share characteristics that make them susceptible to a specific message, 3) sending a tailored message. 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.

Imagine, for example, that a political actor wants to discredit Donald Trump. It would be wasteful to target all U.S. voters. Rather, the malicious political actor could focus on Trump's base, and ignore Democrats altogether. Because Trump's base consists of many different voter groups, it may be ineffective to send one specific deepfake to all voter groups. Rather, the malicious political actor should make several deepfakes and microtarget those deepfakes to the different voter subgroups that constitute Trump's base. NRA-members, for example, would receive a deepfake in which Donald Trump is seen and heard to unfold a plan about gun control. Coal mine workers would receive a deepfake in which Trump ostensibly offends coal miners. Evangelicals are targeted with a deepfake where Trump say something negative about their religion, and so on. While it takes a large effort to create the first deepfake (for instance, gathering enough training material can be challenging), it takes very little *additional* effort to create the second, third, or thousandth deepfake.

Considering that the European Court of Human Rights regards broadcasting as the most pervasive medium due to the power of visuals (ECHR Jersild, 1994), it is no wonder that lawmakers monitor the developments around deepfakes closely and sometimes even attempt to regulate. In an effort to curtail deepfakes' negative potential, lawmakers in California have banned the distribution of political deepfakes that are released within 60 days of an election (AB730, 2019). In the European Union, the European Commission states to be "aware of the technology (...) but cannot yet gauge its impact" (European Commission, 2018). The Commission has taken some active measures (a code of practice on disinformation for platforms and advertisers, investments in online verification tools, setup of a rapid alert system) and supports the creation of "an independent European network of fact-checkers" as part of an action plan to tackle disinformation in the EU (Action Plan Against Disinformation, 2018).

While there is substantial literature on the technical side of deepfakes, such as detection methods (e.g., Güera and Delp, 2019; Li and Lyu, 2018; Yang et al., 2019; Afchar et al.,

2018; Matern et al., 2019), as of yet, deepfakes are only marginally studied in the political communication field. Vaccari and Chadwick (2019) showed 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 (microtargeted) deepfakes on people's political attitudes have never been studied. Knowledge about how (microtargeted) deepfakes affect political behavior could help 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 disinformation

False information generally can be placed in one of three categories: disinformation, misinformation, or malinformation (Wardle & Derakhshan, 2017). Deepfakes fit best in the disinformation category, which encompasses 'manipulated content', 'imposter content', and 'fabricated content' (Wardle & Derakhshan, 2017; p. 5). Disinformation can be seen as "intentional behavior that purposively misleads" (Chadwick et al. 2018; p. 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 (Bradshaw & Howard, 2018; Asmolov, 2019; 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 & Chadwick, 2019). Therefore, disinformation is increasingly regarded a matter of (inter)national security (see, e.g., European Commission, 2019; Atlantic Council, 2019; Metodieva, 2018).

Research on the effects of mis- and disinformation 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. (2019; p. 1) found no evidence for the idea that Russian trolls impacted Americans' political attitudes: those engaging with the Russian trolls "were already highly polarized". These studies occurred in a very specific context (the US 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). But the studies offer a first glimpse into the limits of online disinformation campaigns. We argue that deepfake disinformation could be more impactful than Twitter trolling and false news stories. Moreover, the potential impact of deepfakes may be amplified by microtargeting. After all, the reason for Bail et al.'s (2019) 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 & De Vreese, 2010; p. 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, engage in greater message scrutiny than those who perceive a message as generic (Petty et al., 1995; Wheeler et al., 2005; Chang, 2006). Scrutinized messages are more likely to influence citizens (Petty & 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 PMT's effects on political behavior is scarce. Endres (2019; p. 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 US-context. Decreasing support for the citizen's 'own' candidate and increasing support for the opponent, as demonstrated by Endres (2019), is arguably impressive in a polarized two-party context such as the US (Abramowitz, 2013; Webster & Abramowitz, 2017). In a multi-party 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 & Rosema, 2019). Highly competitive districts such as those studied by Haenschen and Jennings (2019) are difficult to find in a multi-party 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 multi-party 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; p. 4). A prominent example of a gaffe is a recording of 2012 US presidential candidate Mitt Romney, who was covertly filmed when discounting 47% of Americans as entitled, dependent victims who will vote for Obama no matter what (Sheinheit & Bogard, 2016). According to Sheinheit and Bogard (2016), this gaffe correlated with a decrease in support for the 2012 US Republican candidate. A different example is the 'Dean scream', which correlated with the deterioration of the 2004 US campaign of Howard Dean (Kreiss, 2012). Deepfakes are essentially gaffes that never happened: inappropriate statements or behavior that aim to make people question the depicted politician judgment, ability or character (see for original definition on gaffes Frantzich, 2012; p. 4). 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 & Shapiro, 1989; Chanley et al., 2000; Maier, 2011; Praino et al., 2013).[14] Considering literature on gaffes and scandals, we expect that:

H1a: A deepfake meant to discredit a political candidate negatively affects people's attitudes toward the depicted politician.

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:

H1c: The effects of a deepfake meant to discredit a political candidate are stronger for the microtargeted group than the non-microtargeted 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 (Slothuus & De Vreese, 2010; Bolsen et al., 2014). 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 supporters' 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:

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

14 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) judgement of the corrupt politician's character as well.

Methods

Sample

The sample consisted of 278 participants. Participants were recruited by Kantar Lightspeed, a Dutch company specialized in recruitment for academic purposes, and were 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), 54.7% was female. 1% completed elementary school, 20% completed high school, 35% completed vocational school, and 43% held a bachelor's degree or higher.[15] 49.3% of the sample indicated to be Christian. We purposely oversampled Christians to get large enough groups, as the incidence rate of Christians in the Netherlands is only 31% (De Hart & Van Houwelingen, 2018).

Experimental design

We used a 2 (between subjects) by 2 (between subjects) factorial design, with *level of personalization* (microtargeted versus non-microtargeted), and *mode* (deepfake versus original) as between-subject factors (see Table 5.1).

Table 5.1. Experimental design.

	Deepfake	**Original**
Microtargeted	Group 1 (*N* = 72)	Group 3 (*N* = 66)
Non-targeted	Group 2 (*N* = 73)	Group 4 (*N* = 67)

Procedure

Participants were contacted by Kantar Lightspeed. 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. 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.

Independent variables

Deepfake

The deepfake stimulus is a 13 second subtitled video showing a leading politician of Dutch Christian Democrats 'CDA'. The first 8 seconds of the video calls for the attention of the participant and announces to the participants that they are going to see

15 Does not add up to 100 due to rounding.

a short video of [name politician], politician of the CDA. The following 5 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.[16]" Figure 1 shows a screenshot from the deepfake and a screenshot from the original video.

Figure 1 - Still from deepfake (above) and original video (below).

16 In Dutch: "Maar zoals Christus zou zeggen, pin mij er niet op vast."

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 AI-based lip synchronization techniques (Suwajanakorn, 2017).

To produce the silent video, we collected approximately 25 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 12 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, non-manipulated, subtitled version of the video of the politician that was also used for the deepfake. The first 8 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 5 seconds show the original version of the television interview that was manipulated in the experimental condition.

Microtargeted or non-microtargeted appeal

People received a *microtargeted* stimulus when they indicated in a filter question that they were Christian. People received a *non-targeted* stimulus when they indicated in the filter question that they were not religious. The stimulus is about Christianity, and therefore catered to the personal interest of the Christian participants –but not the non-religious participants. Hence, we speak of a microtargeted and a non-targeted 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 non-religious were randomly placed in either the experimental or control non-microtargeted 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 that pray once a week or more often (N = 81; 11% prayed once a week, 13% more than once a week, 76% prayed every day) and those that prayed between at least once a month (but not once a week or more) or who prayed less (N = 52; 21% prayed at least once a month, 10% prayed only on religious holidays, 12% prayed once a year, 58% prayed never). Three participants declined to volunteer how often they prayed and were considered missing. We dichotomized this variable for two reasons. First, because in reality citizens that are being profiled are often classified as either belonging to one subgroup (1) or not (0) (e.g., Briggs Meyers & Meyers, 2010). Two, because 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 5 years, ever cast their votes for the CDA (in European, national, provincial or local elections). Participants could answer either yes or no (217 no, 60 yes, 1 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 α = .93)

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. 0 stood for negative and 10 stood for positive (see Seltzer & 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 website Moreover, our experiment took place in a controlled environment and not during an election.

Manipulation check

Credibility

We measured the degree to which participants found the deepfake credible with two 7-point scale items: *I find the video authentic*, and*, I find the video credible (*a tweaked version of the scale used by Appelman and Sundar 2016; p. 76) (r = .77). 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): $t(274) = 3.08$, $p = 0.01$. 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 non-credible but rather had circumstantial issues, e.g: "*all politicians are non-credible and only care about their own interest and glory"* or "*because it is not in line with how I think and live my life"*. Of the 84 participants who found the deepfake (somewhat) non-credible, only 12 noted that the video was likely manipulated: e.g. *"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 score of the deepfake was close to the credibility of the original video, and because only 12 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 & Park, 2015), we decided to carry out our analyses with all participants, regardless of whether they perceived the deepfake as (somewhat non-)credible[17].

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 α = .79).

17 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 footnotes in the results section.

Comparing the experimental (M = 4.41, SD = 1.25, N = 143) and the control group (M = 4.53, SD = 1.27, N = 133) on the degree to which they scrutinized the stimulus, we found no significant differences between both groups ($t(274) = .74$, $p = .23$). A closer comparison between the religious and the non-religious participants in the experimental and control group yielded no significant differences between the four groups: $f(3, 272) = 1.88$, $p = .13$. 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: $t(65) = -2.51$, $p = .001$. Moreover, after comparing 'heavy prayers' with non-religious participants (M = 4.22, SD = 1.28, N = 74), the data showed that 'heavy prayers' also scrutinized the message significantly more elaborate than the non-religious 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 ($t(79) = -.21$, $p = .42$). 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.

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 ($t(276) = .07$, $p = .95$), 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 ($F(3, 261) = .51$, $p = .68$), gender ($F(3, 274) = .92$, $p = .43$), and education ($F(3, 273) = .55$, $p = .65$).

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 = .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 non-significant: $t(275) = 1.08$, $p = .14$.[18] This means that **H1b** is not supported.

18 In the credibility >3 sample, the M_exp (SD) was 4.90 (2.02), and the M_control (SD) was 5.30 (1.98). $T = 1.47$, $p = .07$.

Zooming in and comparing the four groups, using an ANOVA, we find a significant difference between the four groups regarding attitude toward the politician ($F(3, 273) = 3.21$, $p = .02$.[19] A Bonferroni post-hoc comparison (see Figure 2) showed that experimental group non-religious scored significantly ($p = .04$) lower ($M = 4.15$. $SD = 1.28$, $N = 73$) than the non-religious control group ($M = 4.62$, $SD = .89$, $N = 67$). 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.

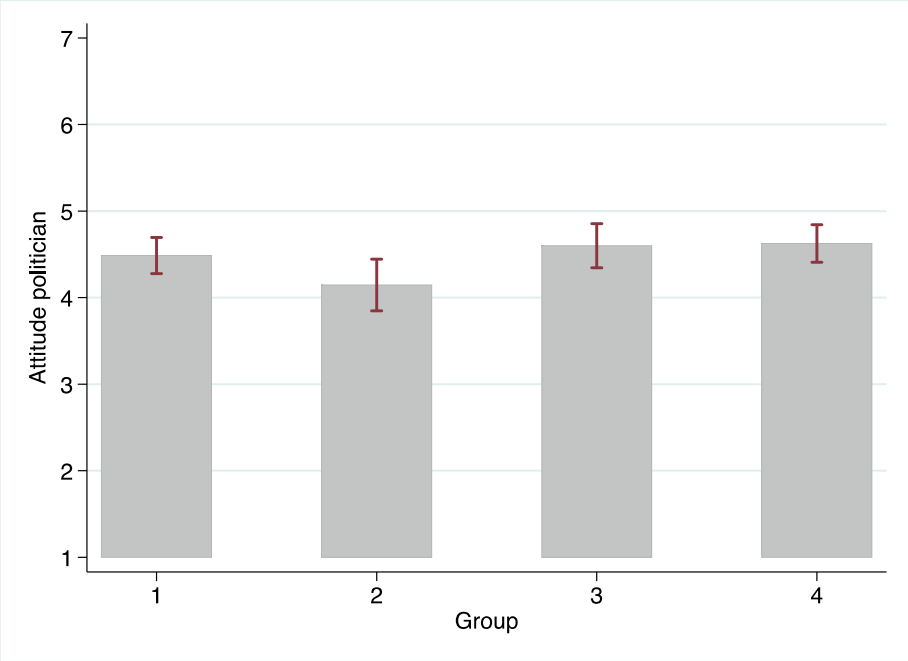

Figure 2 - Confidence interval plot attitude toward politician after exposure, per group. Group 1 = Christian experimental [95% CI 4.28 – 4.70], group 2 = non-religious experimental [95% CI 3.85 – 4.44], group 3 = Christian control [95% CI 4.35 – 4.85], group 4 = non-religious control [95% CI 4.41 – 4.84].

Microtargeting and attitude toward the depicted politician

For a closer examination of the role of microtargeting in moderating effects of deepfake exposure on attitudes toward the depicted politician, we looked at whether someone was a 'heavy prayer'. The microtargeted deepfake should be especially well-tailored to the subgroup of heavy prayers. An ANOVA showed neither a significant difference

19 In the credibility >3 sample, this difference was not significant.

between the mean attitude scores of the experimental group of 'light prayers' (once a month or less) (M = 4.31, SD = .78, N = 27), and the control group that prayed the same amount (M = 4.38, SD = 1.13, N = 25), nor between the experimental group of 'heavy prayers' (once a week or more) (M = 4.57, SD = .93, N = 41) and the control group that prayed the same amount (M = 4.75, SD = .97, N = 40). Finally, there was no difference between any of the four groups: $F(3, 129) = 1.39$, $p = .25$.

But as microtargeting in the context of this experiment is about reaching CDA voters who are also heavy prayers, we conducted a t-test to compare the scores of the experimental group of such participants (M = 4.72, SD = 1.04, N = 14) with their counterparts in the control condition (M = 5.43, SD = .77, N = 10): $t(22) = 1.84$, $p = .04$. Because of the small subsample, a Kolmogorov-Smirnov test was conducted. For the experimental group, the chi^2 value was .38, and for the control group the chi^2 value was .52: indicating normality for both groups. In addition to this, a Shapiro-Wilk test was conducted. The experimental group ($W = .90$, $p = .12$) as well as the control condition ($W = .89$, $p = .17$) were distributed normally, making the t-test a robust test even for these small subsamples.

Microtargeting and attitude toward the political party

Closer examining the effect of microtargeted deepfake exposure on the attitude toward the political party to which the depicted politician belongs, we find similar results. An ANOVA showed that the 'light prayers' in the experimental group (M = 4.63, SD = 2.27, N = 27) did not significantly score different than the 'light prayers' in the control group (M = 4.56, SD = 2.42, N = 25). There was also no significant difference between the heavy prayers in the experimental (M = 5.27, SD = 1.48, N = 41) or the control condition (M = 5.25, SD = 2.26, N = 40)[20]. Finally, the four groups did not significantly differ from each other: $F(3, 129) = 1.65$, $p = .18$.

Similar to the section above, when we zoomed in on the CDA voters who were heavy prayers, a t-test showed that the experimental group held significantly worse attitudes (M = 6.07, SD = .92, N = 14) toward the CDA than their counterparts in the control group (M = 7.30, SD = 1.64, N = 10) did: $t(22) = 2.35$, $p = 0.01$. Also similar to the section above, the Kolmogorov-Smirnov test indicated normality for the experimental (chi^2 .38) as well as the control condition (chi^2 .52). Concordantly, the Shapiro-Wilk test also indicated normal distribution for the experimental ($W = .99$, $p = 1.00$) as well as the control condition ($W = .95$, $p = .65$). Making the t-test above a robust test of the mean differences between these two small subsets of the sample. Moreover, this small subsample would not be enough to conclude the absence of an effect if the results were

20 In the credibility >3 sample, the heavy prayers in the experimental group scored significantly lower (M = 5.31, SD = 1.51) than the heavy prayers in the control group (M = 6.03, SD = 1.51), $t = 1.72$, $p = <.05$.

insignificant, quod non, but finding such a degree of significance with such low power, is enough to conclude the presence of an effect. When the deepfake was microtargeted, participants indeed held significantly and substantially lower attitudes toward the politician and his party: which supports **H1c**.

Motivated reasoning

Turning to the people who indicated to have voted for the CDA at least once in the previous 5 years, we find no significant differences in attitudes toward the politician between the experimental group of CDA voters (M = 4.89, SD = .97, N = 36) and the control group (M = 4.95, SD = .85, N = 24) (t(58) = .25, p = .40. Non CDA-voters in the experimental condition, however, held significantly worse attitudes toward the politician (M = 4.12, SD = 1.09, N = 107) than the non CDA-voters in the control condition (M = 4.54, SD = .97, N = 109) did (t(214) = 3.05, p = .001).

Examining the attitudes toward the CDA in relation with having voted for CDA in the previous 5 years, we see a similar pattern emerge. The data showed that for the experimental participants who voted for the CDA in the previous 5 years (M = 6.28, SD = 1.54, N = 36), their attitudes toward the CDA do not significantly differ from the people in the control group that voted for the CDA in the previous 5 years (M = 6.63, SD = 1.44, N = 24) (t(58) = .88, p = 19). But when we turn to the non CDA-voters, the picture is different: the experimental group holds significantly worse attitudes (M = 3.84, SD = 2.14, N = 107) than the control group (M = 4.34, SD = 2.35, N = 109): t(214) = 1.66, p = .05). These findings suggest motivated reasoning occurred for the partisans in the experimental groups and that motivated reasoning inoculated them from negative effects of the deepfake (answering RQ1).

Table 5.2 and table 5.3 give an overview of all the t test outcomes on the outcome variables attitude toward the politician and attitude toward his political party.

Table 5.2 - Means (SD) for attitude toward the politician as a function of exposure to the stimulus and a moderating variable (microtargeting/motivated reasoning)

Attitude politician				
Moderator	**M_exp (SD)**	**M_control (SD)**	**t**	**p**
None	4.31 (1.10)	4.62 (.96)	2.48	.01
Religious light	4.31 (.78)	4.38 (.78)	.26	.40
Religious heavy	4.57 (.93)	4.75 (.97)	.87	.19
Voted CDA (yes)	4.89 (.97)	4.95 (.85)	.25	.40
Voted CDA (no)	4.12 (1.09)	4.54 (.97)	3.05	.001
Voted CDA (yes) * Religious heavy	4.72 (1.04)	5.43 (.77)	1.84	.04

Table 5.3 - Means (SD) for attitude toward the political party as a function of exposure to the stimulus and a moderating variable (microtargeting/motivated reasoning).

Attitude party				
Moderator	**M_exp (SD)**	**M_control (SD)**	**t**	**p**
None	4.46 (2.26)	4.76 (2.38)	1.08	.14
Religious light	4.63 (2.27)	4.56 (2.42)	-.11	.54
Religious heavy	5.27 (1.48)	5.53 (2.26)	.60	.27
Voted CDA (yes)	6.28 (1.54)	6.63 (1.44)	.88	.19
Voted CDA (no)	3.84 (2.14)	4.35 (2.35)	1.66	.05
Voted CDA (yes) * Religious heavy	6.07 (.92)	7.30 (1.64)	2.35**	.01

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. 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 & Shapiro, 1989; Chanley et al., 2000; Maier, 2011; Praino et al., 2013), 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. (2019), 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. (2019).

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 CDA voters who were very religious, instead of all Christians. But why would other Christians that 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, CU and 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 heavily religious and voted CDA in the past five years.

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 (2013), 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. Motivated reasoning is sometimes considered negative in the face of truthful information (Richey [2012; p. 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 non-partisan 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 5 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 anti- nuclear 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): $t(138) = 2.27$, $p = 0.01$. 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 are more or less aligned with the idea of 'minimal effects' of political communication (see Bennett & Iyengar, 2008). The idea of minimal effects in political communication was later substantiated by Kalla and Broockman (2017), who in an important meta-analysis estimated that the persuasive effects of campaign contact and advertising on American voters was 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-US, multi-party, 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 14 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 believe 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

Deepfakes are new, and their effects on citizens have only been sporadically studied. More research is needed. Research should map the effects of deepfakes, and, more importantly, potential ways to counter those effects. A regulatory focus should be directed against this potential new frontier of disinformation warfare. The surprising low number of participants that 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.