They would never say anything like this! Reasons to doubt political deepfakes

Hameleers, M.; van der Meer, T.G.L.A.; Dobber, T.

DOI
10.1177/02673231231184703

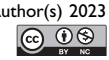
Publication date
2024

Document Version
Final published version

Published in
European Journal of Communication

Full Length Article

European Journal of Communication
2024, Vol. 39(1) 56–70


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02673231231184703
journals.sagepub.com/home/ejc


They Would Never Say Anything Like This! Reasons To Doubt Political Deepfakes

Michael Hameleers, Toni G. L. A. van der Meer, and Tom Dobber
University of Amsterdam, Netherlands

Abstract

Although deepfakes are conventionally regarded as dangerous, we know little about how deepfakes are perceived, and which potential motivations drive doubt in the believability of deepfakes versus authentic videos. To better understand the audience's perceptions of deepfakes, we ran an online experiment ($N = 829$) in which participants were randomly exposed to a politician's textual or audio-visual authentic speech or a textual or audio-visual manipulation (a deepfake) where this politician's speech was forged to include a radical right-wing populist narrative. In response to both textual disinformation and deepfakes, we inductively assessed (1) the perceived motivations for expressed doubt and uncertainty in response to disinformation and (2) the accuracy of such judgments. Key findings show that participants have a hard time distinguishing a deepfake from a related authentic video, and that the deepfake's content distance from reality is a more likely cause for doubt than perceived technological glitches. Together, we offer new insights into news users' abilities to distinguish deepfakes from authentic news, which may inform (targeted) media literacy interventions promoting accurate verification skills among the audience.

Keywords

Deepfakes, experiment, deception detection, disinformation, media literacy, misinformation, verification

Concerns about the democratic consequences of deepfakes – synthetic videos created with deep learning techniques to make authentic persons say or do inauthentic things – have increased considerably in recent years (e.g., Dan et al., 2021; Iacobucci et al., 2021; Paris and Donovan, 2020). As a form of disinformation, deepfakes are created

Corresponding author:
Michael Hameleers, University of Amsterdam, Amsterdam, Netherlands.
Email: m.hameleers@uva.nl

purposively to deceive recipients, and to make an impact on their political beliefs (e.g., Dan et al., 2021; Hancock and Bailenson, 2021; Vaccari and Chadwick, 2020). Although recent empirical research has started to map the effects of deepfakes in the political realm (Barari et al., 2021; Dobber et al., 2020; Vaccari and Chadwick, 2020), we lack a more comprehensive understanding of the societal impact of deepfakes (Hancock and Bailenson, 2021), and the extent to which people are able to correctly detect the deceptive nature of deepfakes more specifically.

It may be difficult for citizens to accurately detect deception, especially when news users are confronted with highly realistic modes of deception alongside accusations of disinformation and fake news labels targeting factually accurate information and actual falsehoods (e.g., Egelhofer and Lecheler, 2019; Van Duyn and Collier, 2019). As deepfakes offer a realism heuristic and a seemingly authentic and close depiction of reality (Sundar et al., 2021), such audiovisual manipulations may circumvent the detection of deception. Thus, it may be hard for news users to distinguish a realistic deepfake from an authentic video in a digital information ecology characterized by information overload and fragmentation. At the same time, however, empirical research has not offered convincing evidence for strong disruptive effects of political deepfakes (e.g., Dobber et al., 2020; Vaccari and Chadwick, 2020). It thus remains an open question how political deepfakes are perceived by citizens.

Against this backdrop, the purpose of this study is (1) to assess the believability of a deepfake and to (2) explore the underlying considerations of news users who doubt deepfakes versus authentic videos. To do so, we rely on an experimental study in which we created a state-of-the art political deepfake that makes a Dutch politician express radical issue-positions – in line with the delegitimizing strategies typically associated with disinformation campaigns in Western democracies.

This paper aims to make important theoretical and empirical contributions to the disinformation literature. Although extant research has mapped disinformation's effects on misperceptions, political judgements and news credibility (e.g., Schaewitz et al., 2020), we know markedly little about the impact of deepfakes (but see e.g., Barari et al., 2021; Dobber et al., 2020; Hameleers and van der Meer, 2020; Iacobucci et al., 2021; Lee and Shin, 2022; Vaccari and Chadwick, 2020). At the very least, we lack an empirical test of claims associating deepfakes in the political realm with dystopian ramifications – assuming that news users are unable or unwilling to separate fake from real videos. By directly comparing believability and doubts expressed in response to synthetic versus related authentic information, we can better position the allegedly worrisome consequences of deepfakes into the disinformation literature.

The believability of deepfakes

In this paper, we focus on disinformation: Information that deviates from facticity intentionally, and is created, fabricated, manipulated and disseminated to achieve political aims (e.g., Bennett and Livingston, 2018; Freelon and Wells, 2020; Wardle and Derakhshan, 2017). We consider deepfakes as a form of deception, for which the agent of disinformation deliberately aims to cultivate false beliefs among receivers (e.g., Hancock and Bailenson, 2021). This is different from misinformation, which

refers to false information in general, or information that is inaccurate without the intention to deceive (e.g., Freelon and Wells, 2020). In this paper, we specifically focus on deepfakes as political disinformation: Intentionally deceptive synthetic videos that aim to offer a direct index or 'heuristic' of reality which may circumvent the activation of suspicion among recipients.

Deepfakes can technically be understood as synthetic videos in which real actors (i.e., politicians, celebrities) are made to express or perform inauthentic things (e.g., Dobber et al., 2020; Paris and Donovan, 2020; Vaccari and Chadwick, 2020). Deepfakes are made using the affordances of Artificial Intelligence (AI). Specifically, deep learning is used to transpose manipulated audio on authentic speeches, or swap the inauthentic narratives and behaviors of a (voice and/or motion) actor on the moving images of a depicted (political) actor used as the subject of the deepfake (e.g., Dobber et al., 2020; Hancock and Bailenson, 2021).

Based on multimodal framing literature (e.g., Geise and Baden, 2014; Powell et al., 2018), we expect that deepfakes may have an advantage over textual disinformation when it comes to believability and the perceived credibility of information. In line with this, the findings of experimental research by Sundar et al. (2021) indicate that video-based disinformation is rated as more realistic than text- or audio-based disinformation, but only among participants that are less involved in the issue. Applied to deepfakes in particular, Lee and Shin (2022) and Hwang et al. (2021) found that deepfakes are seen as slightly more vivid and credible than textual disinformation, but differences are relatively small.

Deepfakes may, however, not directly deceive recipients. Vaccari and Chadwick (2020) found that exposure to a political deepfake mostly affected the trust people have in (online) media: People are more likely to become uncertain than to adjust their political evaluations after seeing a deepfake, which could indirectly harm news trust. Barari et al. (2021) further indicate that, although deepfakes can be believable, they are not more persuasive than other forms of mis- or disinformation (also see Hameleers et al., 2022).

In assessing responses to disinformation, we acknowledge that trust can be regarded as a multidimensional concept, for example, as it consists of trust in the selection of issues, trust in the selectivity of factual information, trust in the accuracy of reality depictions, and the assessment of the journalist as unbiased and trustworthy (Kohring and Matthes, 2007). Just like trust, credibility may have different components, such as perceived authenticity and objectivity (Kohring and Matthes, 2007). Engelke et al. (2019) further reflect on the differentiation between trust and distrust, which may particularly be useful for our endeavor where we explore perceived deception and disinformation, including the potential perceived manipulative intent of the communicator. As we rely on the analysis of open-ended responses, we aim to incorporate the different dimensions of trust, credibility, and distrust in our analysis. Here, we refer to the expression of doubt as a proxy for a lack of trust and credibility, or distrust in the message (i.e., as it may be perceived as fake or manipulated).

Based on the reviewed literature, we can conclude that, to date, evidence on a strong persuasive advantage of deepfakes compared to other forms of disinformation and authentic content in the political realm is lacking. In this setting, we ask: How believable, trustworthy and credible are deepfakes compared to authentic information and textual disinformation? (RQ_1).

Exploring the perceived causes of deception

As a key aim of this paper, we aim to inductively explore the *perceived* causes of doubt in both authentic information and deepfakes, and herewith assess the extent to which news users are able to correctly identify the deceptive nature of deepfakes, and the causes they perceive to be underlying its deception. Doubts in believability can apply to various aspects of the message, such as the message, the source, or the presentation of the message. These doubts may reflect a lack of trust or credibility, or specific distrust related to the source or content of the message. Importantly, although close-ended questions typically used in mis- and disinformation studies may activate suspicion by asking people to rate the credibility of (mis)information (Levine, 2021), we measure responses to the deepfake and authentic video in an open-ended manner. Using this approach, we are able to arrive at an unprimed assessment of responses to deepfakes, revealing the reasons of doubt and uncertainty in response to deepfakes and other forms of disinformation.

Using an inductive approach, we specifically aim to explore the variety of motives that people express for doubting authentic and inauthentic speeches, hereby also exploring 'false negatives' in deception detection. As Vaccari and Chadwick (2020) found that deepfakes may result in uncertainty, confusion, and eventually lower news trust, it is important to reveal the specific reasons identified by news users to doubt the veracity of factually accurate and (intentionally) false information. The following research question is introduced: What causes for doubt are identified by skeptical news users, and to what extent are these perceived causes reflecting substantial/argument-based versus technological reasons? (RQ_2).

Method

To test our exploratory research questions on the perception of deepfakes, we use an experiment in which we exposed participants to a political deepfake versus textual disinformation and authentic videos. In this deepfake, a real mainstream political actor in the Netherlands was made to express an inauthentic and uncivil speech: A radical right-wing de-legitimizing issue position that has been associated with prevalent disinformation campaigns in Western democracies and Europe in particular (e.g., Bennett and Livingston, 2018; Marwick and Lewis, 2017).

Design

We randomly exposed participants to (1) an unrelated authentic news message (control); (2) a related authentic news message or (3) disinformation. As a second factor, the modality of the message was varied: Participants either saw a textual stimulus or a video (a deepfake in the disinformation condition). The design thus had six between-subjects conditions: Participants either saw one of the two factually accurate messages or disinformation, and either watched a video or read a short online article with exactly the same script (see Appendix A). In the survey-embedded online experimental environment, participants first answered questions on their background (including political interest, ideology, demographics) before they saw the stimulus material on their computer. After seeing the

video or text (they could not skip this task), they answered questions on the dependent variable. The survey task (including the questionnaires at the start and end) lasted for an average of 15 min.

The topic and depicted politician were held constant: All participants were exposed to a speech of the former leader of the Dutch Christian-Democrats who expressed his views on norms, traditions and values in a short (50 s) video. This actor is a mainstream politician who does not typically express hostile or radical right-wing issue positions, although he can 'flirt' with right-wing populist elements in his speeches (this enhances the credibility of the deepfake). Although we kept as many factors as possible constant across the different conditions, the manipulation of a deepfake as a form of deception required us to deviate from the arguments of the authentic speech – the message was fabricated and intentionally deceptive to reflect real cases of disinformation as close as possible.

However, the overall topic and political statements were similar across the conditions, and matched for the related authentic condition that serves as a benchmark to assess the perceived deceptiveness of the deepfake. We also found that the different stimuli scored similarly on perceived emotionality, negativity, and ideological bias. However, the differences between the unrelated (more neutral) authentic speech and the deepfake were more substantial than the differences between the related authentic speech and the deepfake (which we also included as an additional control condition as it more closely reflects the argumentation and language of the deepfake).

Stimuli

Original deepfakes and disinformation narratives were developed for this research project. To create the stimuli, we used deep learning techniques to create a synthetic video in which we made an established politician voice a deceptive political message. We specifically relied on a voice actor that expressed the targeted manipulated speech, which we transferred onto real footage of the delegitimized political actor. The mouth movements and gestures of this depicted political actor were manipulated to match the speech. A professional computational visual artist was hired to achieve the most realistic outcome at the time of data collection.

As baseline for the stimuli, we rely on two authentic political speeches with low salience at the time of data collection (we controlled for prior exposure and familiarity with the speech). We selected a former political leader with relatively conservative views: This should make the targeted deepfake – a right-wing populist message – relatively plausible, and at least not completely incompatible with the views of the depicted politician.

For the unrelated message, we used an excerpt of the depicted politician's speech (recorded on video in 2019), where some general statements on the country's progress and the erosion of a forward-looking culture are discussed. For the related authentic speech, we use an excerpt of the same speech, but use a fragment in which statements similar to the disinformation message are expressed. In this excerpt, the politician speaks about national values and norms, the ordinary Dutch people, and the experience of being alienated due to migrants entering the country. In other words, the fragment reflected a right-wing conservative agenda and flirts with right-wing populist rhetoric. This authentic message was thus ideationally closely related to the targeted deepfake speech.

To construct the deepfake, authentic footage of the political actor was used as training material for algorithms that imitated the facial movements of the targeted politician. After training these algorithms, the existing footage was manipulated by adjusting the facial movements to the recorded (deceptive) political speech. Among other things, lip-sync techniques were used to make the recorded voice reflect the movements of the depicted politician. The scripts of all stimuli are available in the supplemental materials file (Appendix A).

Ethics

As participants were deceived as part of the experiment, ethical considerations are crucial to consider. The data collection and design of the experiment received ethical approval from the University of Amsterdam's ethical review board. In collaboration with the ethics board, extensive debriefing procedures were developed.

Sample

Data collection was outsourced to a large international research company (Kantar) who relies on mixed resources to approach nationally representative samples. In the Dutch population, right-wing parties are electorally more successful than left-wing parties. The average level of education is relatively high. According to the central bureau for statistics, only 10% of the population has obtained primary education or less, whereas 30% is higher educated (university degree or higher). As we aimed for a sample that approached these statistics, we enforced quota during data collection. In the end, 829 valid responses were retained for the analyses. The mean age of participants was 49.75 years ($SD = 14.65$). 42.6% was male. Regarding the highest level of obtained education, 20.7% was lower, and 34.1% was higher educated, by and large matching national statistics on these demographics. Political interest as well as left-right ideological self-placement were normally distributed. In our sample, 28.8% of all participants self-identified as (somewhat) left-wing (7.1% scored the two most extreme values) and 34.3% identified as (somewhat) right-wing (here we also see that 7.1% scored the most extreme values on the scale). This distribution is representative of the voting behavior of Dutch citizens, which enhances the generalizability of our findings. In our sample, 55.2% was (very) interested in politics, whereas only 23.0% was not (at all) interested. Arguably, this does not reflect the level of political interest in Dutch society, which is lower compared to the distribution on this variable in our sample. Finally, we assessed the mean approval rate of the depicted political actor in the deepfake (the former leader of the Christian Democrats). His average approval rate ($M = 36.32$, $SD = 23.90$) was very similar to other politicians.

Here, we would like to stress that the different demographic and socio-political variables were randomly distributed across the conditions of the experiment. We conducted post-hoc randomization checks that confirmed that age, level of education, left-right-wing self-placement, political interest, gender, and prior support for the politician were present in equal proportions across all groups. Controlling for these variables in the analysis did not yield any other findings than presented in the results section. However, looking across

conditions, we can see that participants who disapproved of the depicted politician and his political views were more likely to doubt the message (irrespective of veracity). At the same time, the correct classification of disinformation as a deepfake was not predicted by prior levels of support for the political actor or congruent issue attitudes.

Based on the informed consent procedure, a basic pre-treatment attention check and quota, we excluded 2.5% of the participants who did not agree with the terms and conditions for participation. We excluded 11.1% due to failing to comply with the attention check. The sample size was determined prior to data collection based on power analyses. Based on previous research on the effects of disinformation, we expected small effect sizes (<.20). We also have means and standard deviations available for our key dependent measures – which we used to calculate the appropriate sample size needed to obtain a power of .80.

Dependent Variable

To measure believability, and the different components central to trust, distrust and credibility, we used an open-ended answer field (essay format) where participants were asked to write down the reasons why they doubted different elements of the message, source, and presentation. This procedure inevitably entails the potential priming of suspicion – as we asked respondents to think about reasons for doubt. However, we only posed the question for participants with skeptical view regarding the message, and we used a question wording that was fairly neutral regarding the causes for doubt (we did not directly want to prime distrust or perceived manipulation). In addition, as exactly the same question was used for the authentic messages and the manipulated disinformation conditions, the potentially priming role of asking this question is similar across conditions.

The open-ended question was thus only completed for people with (moderate levels of) doubts related to the believability of the message, which we defined as scoring 4 or lower on at least one of the following three items measured on 7-point scales: "How believable were the following elements of the political speech? (the way it was presented, the content of the speech, the political actor)." Thus, even if people only scored low on one of these items, they were considered as having (some) doubts in the deepfake (the measures correlate strongly at $r = .88$). We also assessed whether separate indicators of doubt (i.e., in the presentation, the message and the source) yielded different findings related to the causes of doubt. The findings are the same when we distinguish between these elements of doubt.

The open-ended question following this assessment was formulated as follows: "You just indicated that you doubted (elements of) the believability of the item you just saw. Could you please indicate why you had some doubts?" We additionally used alternative re-coding procedures and classifications in pilot testing phases, but this threshold was found to generate the most valid and meaningful results.

Answers to the open-ended questions were coded in two subsequent steps. First, we rely on a line-by-line three-step coding procedure that follows the steps of Grounded Theory coding (Charmaz, 2006). First, open coding was employed: All text was highlighted and assigned descriptive labels to code the perceived reasons of doubt. We looked for reasons or perceived causes of doubt that were either related to *what* was said (statement or content-wise distrust and discrepancies) or *how* it was said (perceived manipulation of the

audio-visual stimuli, deepfake classifications and related technological causes). Other reasons were also coded to not limit the analytical focus by pre-defined schemes.

During the next step, focused coding was conducted: Individual codes were merged, grouped or raised to a higher order by taking them out of the specific context, which offered insights into the main motives, reasons and causes of participants' doubts. Here, we paid attention to the multidimensional measurement of trust (Kohring and Matthes, 2007) and the differentiation between trust and distrust (e.g., Engelke et al., 2019). Finally, axial coding was used – were we looked for the connection between themes based on overlaps and discrepancies. Peer debriefing on all coding procedures was conducted, which means that a second researcher assessed all the data reduction steps, offered feedback, and independently reconstructed themes based on the raw lists of codes until full agreement was reached.

Manipulation checks

We assessed whether people correctly remembered the arguments that set the deepfake apart from the authentic speeches. We first of all confirm that participants were more likely to associate the deepfake's argument that 'Dutch norms and values should be protected against foreign influences' with the deepfake ($M = 5.52$, $SD = 1.51$) than the authentic speeches ($M = 4.58$, $SD = 1.65$; $t(828) = 10.39$, $p < 001$). The same applies to the statement that 'the speech argues that immigrants and refugees are allowed to influence our culture with backwards ideas, traditions and religions' (authentic: $M = 4.38$, $SD = 1.65$; deepfake: $M = 5.36$, $SD = 1.65$; $t(828) = 10.62$, $p < 001$). The one statement that more clearly distinguished the related authentic speech from the deepfake was 'People from backwards societies who we welcome in our country are likely to commit violent crimes such as robberies and rapes' was most clearly differentiated: People were substantially more likely to associate it with the deepfake ($M = 5.48$, $SD = 1.57$) than the authentic messages ($M = 3.94$, $SD = 1.71$; $t(828) = 16.33$, $p < 001$). The differences become more pronounced when we compare the deepfake to the unrelated control condition, which confirms that the manipulations were perceived as intended. Hence, the related authentic video matches the deepfake on the line of argumentation much closer than the unrelated authentic video.

As an additional check for the realism of the deepfake compared to the authentic videos, we assessed the extent to which participants in different experimental groups agreed with the statement that the 'message is similar to the information I come across on a daily basis' – we found non-significant differences between the authentic unrelated ($M = 3.77$, $SD = 1.55$), authentic related ($M = 3.62$, $SD = 1.91$) and deepfake conditions ($M = 3.52$, $SD = 1.66$).

Results

The relative believability of a deepfake

As a first exploratory step (RQ_1), we aimed to assess the relative believability of the political deepfake. We rely on a simple one-way ANOVA in which we compare the mean believability scores across the conditions, and use (Bonferroni corrected) pairwise mean score comparisons to map the difference between authentic versus

deepfake speeches and deepfakes versus textual disinformation. The results are depicted in Table 1.

On average, deepfakes are quite likely to be regarded as believable: The mean credibility score approaches the midpoint of the 7-point aggregate scale, although the overall believability is significantly lower than 4 ($M = 3.44$, $SD = 1.45$, $p < .001$). However, deepfakes are not significantly more or less believable than textual disinformation. If anything, textual disinformation is slightly more believable than a deepfake ($M = 3.56$, $SD = 1.31$), albeit not significantly so. If we look at the comparison of deepfakes to related and unrelated authentic political speeches using the same modality, we see that a deepfake is significantly and substantially less believable than an unrelated authentic speech ($\Delta M = -1.21$, $SE = .15$, $p < .001$). This difference is substantially smaller when comparing the believability of a deepfake to a related authentic speech ($\Delta M = -.51$, $SE = .10$, $p < .001$). Based on these findings, we can arrive at quite optimistic conclusions on the believability of a political deepfake. However, the difference in believability is less substantial when we compare a synthetic video (the deepfake) to an authentic video that is related in terms of the arguments manipulated in the deepfake video.

Perceived causes of doubt

Why do people doubt the believability of deepfakes, and are these perceived causes associated with the argument versus technological aspects of the deepfake? (RQ_2). Based on the line-by-line analyses and complete coding of all open-ended responses, we can distinguish the following perceived causes for participants' distrust in the message they saw: (1) a content-wise discrepancy between the political reality and the statements voiced in the message (30.6%); (2) perceived manipulation and doctoring of the audiovisual stimuli (i.e., explicit references to inauthentic audio fragments and manipulation of the stimuli) (12.1%); (3) broader levels of political distrust and opinion-based disagreement (15.9%); (4) lack of factual evidence, sourcing and argumentation (10.3%) and (5) self-perceived media literacy (2.7%). 28.5% gave an answer that could not be categorized as a substantive reason or theme (i.e., don't know answers). As robustness checks, we distinguished between different degrees of doubt (scores ranging from 1–2 indicate more severe doubts than 3–4). The main findings are similar across scores, although

Table 1. Mean score comparison of believability across modalities and facticity.

	Unrelated authentic information	Related authentic information	Disinformation
Textual	$4.08_{a,a}$ (1.11)	$3.67_{b,b}$ (.99)	$3.56_{b,b}$ (1.31)
Video	$4.65_{b,a}$ (.96)	$3.95_{b,b}$ (1.20)	$3.44_{b,c}$ (1.45)
F (5,1294)	17.73***		
partial η^2	.064		
N	829		

*** $p < .001$

Note. Means with different first subscripts indicate significant mean differences within columns, means with different second subscripts indicate significant mean differences within rows (based on corrected pairwise t-tests ($p < 0.05$)).

political distrust and opinion-based disagreement are more salient in the severe doubt (scores 1–2 on doubt) compared to the moderate doubt group.

Regarding the first and most salient theme, most participants who distrusted the speech pointed to a discrepancy between the politician's actual statements and issue positions and the manipulated (radical right-wing) issue positions: "He would never say anything like this. They are not as radical as the PVV, FvD or JA21 [right-wing populist parties in the Netherlands]. It is very confusing, this is not what Buma [the depicted politician] says normally." This theme relates to the perceived cause of the extremity of the manipulated statements, which was associated with Fake News: "I believe there are a lot of extreme statements in this message. I think this is Fake News." For most participants, then, the speech was seen as incredible as the political statements did not match the known values and political ideology of the party that was allegedly making the claims.

Another common theme was also related to the content-level: Participants frequently pointed to the lack of facts and empirical evidence offered to support the political statements: "Because he does not give any foundations or explanations. No facts, numbers or evidence. He speaks on behalf of the people without saying how he arrived at these conclusions." Another perceived cause not directly related to the content of the message itself was the systematic distrust people had in political elites and politicians, their dishonesty and deliberate use of manipulation for political gains: "Politicians use some facts and place these in another context for their own political gain." Another related cause was political cynicism resulting in low believability: "Politicians can't be trusted as they distort the truth and fail to live up to their promises." Participants thus emphasized that they did not believe the speech because of an overall lack of trust in politicians and the elites that govern the country.

Although the actual recognition of the deepfake based on technical manipulations was far less salient than perceived content-wise discrepancies, some participants explicitly pointed out how the deepfake was created and fabricated: "Because we see the politician, but we hear the voice of someone else, which has been synchronized with his mouth movements. This is Fake News!" Some participants identified the combination of content-wise discrepancies and perceived technological manipulation and doctoring as the main cause for their distrust: "Text and mouth were not completely synchronized." Some also explicitly pointed out that the speech was a deepfake: "This is not real. It's a deepfake."

Some participants indicated that they did not trust the speech because they personally disagreed with the statements voiced: "Because I completely disagree with the content of the message." Others more explicitly indicated their disagreement with arguments: "I seriously doubt whether the elite treats migrants better than the Dutch population." Hence, next to pointing out the discrepancy between reality and the speech (based on perceived fabrication/tempering or content-wise discrepancies), participants mentioned that the discrepancy between their own (political) views and the speech as a reason for their lower levels of believability.

Finally, media literacy was mentioned as a reason for lower believability by some participants. In this case, participants emphasized that not all content should be (uncritically) accepted, and that it was important to doubt information that does not seem to be accurate: "Because it is published on social media, and I do not accept everything directly as

being truthful." Some also emphasized that one (unclear) source could not be trusted without consulting more information: "I think it is crucial to consult multiple sources. This is just one source, I cannot know for sure if this is true or not." This theme relates to the first two themes in the sense that participants critically assess the veracity of the statements and stimuli to arrive at their verdict of untrustworthiness.

In the next step, we quantitatively coded the open-ended responses to match these main categories. Using this new variable, we were able to assess which perceived causes mentioned are most likely to correspond with participants' (in)correct classifications. To do so, we recoded people's believability assessment into a binary variable: A correct classification of untruthfulness versus an incorrect assessment. Table 2 shows the proportion of incorrect versus correct classifications across all five perceived causes of unbelievability. A binary logistic regression model revealed the centrality of content-wise discrepancies and technologically motivated perceived deception in correct classifications. Specifically, participants perceiving the content or statement-wise discrepancy between the depicted political actor and his party and the political speech as the cause of their doubt were significantly more likely to arrive at a correct than incorrect judgement ($B = .79$, $SE = .19$, $OR = 2.21$ $p < 0.001$, 90% CI OR [1.54, 3.17]). The same was found for participants basing their doubt on technological causes (i.e., the manipulation and fabrication of audiovisual materials to create a deepfake) ($B = 1.42$, $SE = .26$, $OR = 4.12$, $p < 0.001$, 90% CI OR [2.49, 6.82]). For the other causes, the effects of perceived caused of deception on correct versus incorrect classifications of disinformation were not significant.

Discussion

Although alarming conclusions and dystopian narratives on the destabilizing impact of deepfakes have been voiced, we lack empirical evidence on the effects of exposure to political deepfakes (but see Dobber et al., 2020; Vaccari and Chadwick, 2020). To better understand the extent to which news users doubt deepfakes, and to comprehend why

Table 2. Perceived reasons for unbelievability of materials specified for incorrect versus correct classifications of untruthfulness.

	Perceived causes for unbelievability					
	Content-wise discrepancy	Perceived manipulation and fabrication of stimuli	Overall distrust	Lack of factual evidence and sourcing	Media literacy	Total
Incorrect	44.5%	30.0%	67.4%	62.4%	54.5%	53.9%
Correct	55.5%	70.0%	32.6%	36.7%	45.5%	46.1%
Total	30.6%	12.1%	15.9%	10.3%	2.7%	
χ^2 (5)	53.07***					
N	829					

***$p < 0.001$

they potentially doubt it, we rely on an experiment in which we exposed participants to authentic speeches versus disinformation and asked them to rate the messages' believability, and express their reasons for doubting the content, source, and presentation of the message.

Overall, we found that deepfakes are seen as relatively believable, but substantially less so than authentic videos that more closely reflect the issue positions of the depicted politician. Most participants who doubted the speech based their judgment on the inconsistency between the (extreme radical right-wing) arguments that were voiced in the political speech and the known profile of the depicted politician, who is not likely to express such extreme viewpoints in real life. Other participants, although substantially less than those who pointed to a content-wise discrepancy, connected their doubt to perceived manipulation or doctoring of the audio-visual materials itself, for example, as the voice did not match the politicians' actual voice, or the moving images that did not always closely match the audio.

Does this mean that we can be optimistic about citizens' abilities to distinguish facts from fiction? Unfortunately, the answer is more complex than that. First of all, less than half of all participants that expressed doubts were correct: Over 50% of all participants doubted a message that was authentic. Such 'false negatives' were most likely to occur for people expressing overall levels of distrust in the political elite and the established order as their reasons for doubt. This reflects the overall climate of distrust and the (radical right-wing) cultivation of cynicism and distrust in many Western democracies (e.g., Egelhofer and Lecheler, 2019; Van Aelst et al., 2017; Waisbord, 2018). This climate of factual relativism arguably contributes to overall distrust and doubt among news users, who are not always able or willing to accurately separate facts from fiction.

Yet, our findings also have more optimistic implications for news media literacy in times of heightened concerns on deepfakes. Media literate citizens should theoretically be better able to distinguish false information from authentic content (Flynn et al., 2017), even when quality deepfakes cannot be distinguished from real. Our findings confirm that when people base their doubts on argument-based or stimuli-based causes, they are more likely to arrive at an accurate verdict on the false and deceptive nature of disinformation. In line with the theoretical notion of media literacy as the critical skills needed to understand biases in the production and consumption of media content, and the (political) motivations driving media coverage (Jones-Jang et al., 2021), we show that media literacy can help news users to correctly detect deepfakes, which could minimize its potentially harmful impact on democracy. In addition, following Hobbs' (2021) understanding of critical media literacy as a process of learning that involves asking questions to the media, we can also regard the expression of uncertainty and doubt as a form of inquiry. Hence, when people doubt the authenticity or honesty of the content they are exposed to, they may analyze the causes, consequences and content of information more deeply, and gain control by learning about information. Yet, our findings do indicate that self-perceived media literacy that is not directly related to the identification of argument-based or stimuli-based causes may harm the correct identification of disinformation. There may thus be a difference in media literacy as inquiry and asking questions that spark further verification (Hobbs, 2021) and overconfidence in one's own skills to make decisions about the trustworthiness of information.

Based on our findings, we can arrive at practical recommendations for the design of successful media literacy interventions that can help news users to resist deepfakes whilst trusting authentic information. First and foremost, news users need to be stimulated to reflect on the plausibility and (ideological) fit of (political) statements in the context of depicted actors, parties or institutions. A basic question that needs to be answered is: Does the content of the message reflect the positions of this politician, and are the statements not too extreme for the depicted actor? Second, although deepfakes may get more realistic over time, news users may look for glitches and technological inconsistencies in the content: Does the audio reflect the real voice of the actor, and are the moving images authentic? Third, media literacy interventions need to restore overall trust in authentic content, explaining which sources can be trusted and how citizens themselves may verify (official) information when in doubt.

Despite offering important insights into news users' susceptibility to deepfakes and their media literacy in recognizing them, this paper has a number of limitations. First, although we used the most sophisticated techniques of deep learning and visual effects available at the time of data collection, techniques are improving rapidly. In addition, the evidence is based on a single case study, with a single right-wing populist messages connected to a mainstream politician. Although we believe that most of the findings are transferable to other national settings with similar political landscapes characterized by relatively high levels of distrust in the political and media system and successful right-wing populism, future research may need to enlarge the scope by including different politicians, political viewpoints and countries.

Nevertheless, we believe that this study has offered important insights into news users' susceptibility and resistance to the political consequences of deepfakes. Although we show that deepfakes can be regarded as credible and trustworthy, they are not uncritically accepted by recipients. By revealing the specific considerations underlying people's detection of deepfakes, news media literacy programs in the digital society can be optimized to protect news users from the potentially harmful effects of multimodal doctoring and manipulation.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by a Facebook Policy Research Initiative (an unrestricted gift).

ORCID iDs

Michael Hameleers https://orcid.org/0000-0002-8038-5005
Tom Dobber https://orcid.org/0000-0002-6657-4037

Supplemental material

Supplemental material for this article is available online.

References

Barari S, Lucas C and Munger K (2021) Political deepfake videos misinform the public, but no more than other fake media. *OSF Preprints*. https://doi.org/10.31219/osf.io/cdfh3

Bennett LW and Livingston S (2018) The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*. https://doi.org/10.1177/0267323118760317.

Charmaz K (2006) *Constructing Grounded Theory*. London: Sage.

Dan V, Paris B, Donovan J, et al. (2021) Visual mis- and disinformation, social media, and democracy. *Journalism & Mass Communication Quarterly* 98(3): 641–664.

Dobber T, Metoui N, Trilling D, et al. (2020) Do (microtargeted) deepfakes have real effects on political attitudes? *International Journal of Press/Politics* 26(1): 69–91.

Egelhofer JL and Lecheler S (2019) Fake news as a two-dimensional phenomenon: A framework and research agenda. *Annals of the International Communication Association* 43(2): 97–116.

Engelke KM, Hase V and Wintterlin F (2019) On measuring trust and distrust in journalism: Reflection of the status quo and suggestions for the road ahead. *Journal of Trust Research* 9(1): 66–86.

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

Freelon D and Wells C (2020) Disinformation as political communication. *Political Communication* 37: 145–156.

Geise S and Baden C (2014) Putting the image back into the frame: Modeling the linkage between visual communication and frame-processing theory. *Communication Theory* 25(1): 46–69.

Hameleers M and van der Meer T (2020) Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research* 47(2): 227–250.

Hameleers M, van der Meer TGLA and Dobber T (2022) You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media. *Social Media+ Society* 8(3). https://doi.org/10.1177/20563051221116346.

Hancock JT and Bailenson JN (2021) The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking* 23(4): 149–152.

Hobbs R (2021) *Media Literacy in Action: Questioning the Media*. New York: Rowman & Littlefield Publishers.

Hwang Y, Ryu JY and Jeong SH (2021) Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking* 24(3): 188–193.

Iacobucci S, De Cicco R, Michetti F, et al. (2021) Deepfakes unmasked: The effects of information priming and bullshit receptivity on deepfake recognition and sharing intention. *Cyberpsychology, Behavior, and Social Networking*. https://doi.org/10.1089/cyber.2020.0149.

Jones-Jang SM, Mortensen T and Liu J (2021) Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *American Behavioral Scientist* 65(2): 371–388.

Kohring M and Matthes J (2007) Trust in news media: Development and validation of a multidimensional scale. *Communication Research* 34(2): 231–252.

Lee J and Shin S-Y (2022) Something that they never said: Multimodal disinformation and source vividness in understanding the power of AI-enabled deepfake news. *Media Psychology* 25(4): 531–546. https://doi.org/10.1080/15213269.2021.2007489.

Levine T (2021) Distrust, false cues, and below-chance deception detection accuracy: Commentary on stel et al. (2020) and further reflections on (un)conscious lie detection from the perspective of truth-default theory. *Frontiers in Psychology* 12. doi: 10.3389/fpsyg.2021.642359.

Marwick A and Lewis R (2017, May 15) *Media Manipulation and Disinformation Online*. New York, NY: Data & Society Research Institute. https://datasociety.net/output/media-manipulation-and-disinfo-online/

Paris B and Donovan J (2020) Deepfakes and cheapfakes: The manipulation of audio and visual evidence. *Data & Society Report*. https://datasociety.net/library/deepfakes-and-cheap-fakes/

Powell TE, Boomgaarden HG, De Swert K, et al. (2018) Video killed the news article? Comparing multimodal framing effects in news videos and articles. *Journal of Broadcasting & Electronic Media* 62(4): 578–596.

Schaewitz L, Kluck JP, Klösters L, et al. (2020) When is disinformation (in) credible? Experimental findings on message characteristics and individual differences. *Mass Communication & Society*. https://doi.org/10.1080/15205436.2020.1716983.

Sundar SS, Molina MD and Cho E (2021) Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication*. https://doi.org/10.1093/jcmc/zmab010

Vaccari C and Chadwick A (2020) Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society* 6(1): 1–13.

Van Aelst P, Strömbäck J, Aalberg T, et al. (2017) Political communication in a high-choice media environment: A challenge for democracy? *Annals of the International Communication Association* 4: 3–27.

Van Duyn E and Collier J (2019) Priming and fake news: The effects of elite discourse on evaluations of news media. *Mass Communication and Society* 22(1): 29–48.

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

Wardle C and Derakhshan H (2017) Information disorder: Toward an interdisciplinary framework for research and policymaking, *Council of Europe report*. http://tverezo.info/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-desinformation-A4-BAT.pdf