International Journal of Communication 19(2025), 180–198 1932–8036/20250005

How Far Can Political Deepfakes Credibly Deviate from Reality? Responses to Political Deepfakes with Varying Degrees of Deception

MICHAEL HAMELEERS
TONI VAN DER MEER
TOM DOBBER
University of Amsterdam, The Netherlands

Political deepfakes have been regarded as potentially harmful for democracy. However, to date, we lack a clear understanding of how people respond to deepfakes, especially when they differ in plausibility and the extremity of the deceptive message. Against this backdrop, this study relies on a qualitative and quantitative analysis of open-ended questions (N = 601) asked after exposure to 3 different political deepfakes. These deepfakes either contained a plausible message or implausible messages with counterattitudinal statements. Distrust in the plausible deepfake was mostly attributed to the insincerity of the depicted politician. The implausible deepfakes, in contrast, were also doubted because of perceived issues with the authenticity of the footage. As a main contribution to the (AI-powered) disinformation literature, we show that the credibility of disinformation is understood in different ways depending on the context of deception, ranging from the plausibility of content to the authenticity of the presentation of synthetic media.

Keywords: deepfakes, disinformation, misinformation, political communication, qualitative analysis

Deepfakes, which we define as synthetic multimedia disinformation generated with AI, have caused widespread societal concern. Because deepfakes offer a strong link to reality, they are associated with a lower likelihood of deception detection than text-based disinformation (Sundar, Molina, & Cho, 2021). Hence, people may be more likely to believe something that they see as compared with written abstractions of events or phenomena. Moreover, as deep-learning techniques are getting more sophisticated in generating realistic synthetic media, it may become more difficult for citizens to detect the differences between authentic and synthetic media (Weikmann & Lecheler, 2023; Westerlund, 2019). Because of their alleged persuasiveness, deliberately targeted deepfakes may be strategically used to harm political opponents, blackmail individuals, disseminate propaganda, or increase distrust in conventional journalistic news coverage and the established order (e.g., Diakopoulos & Johnson, 2021; Maras & Alexandrou, 2019).

Michael Hameleers: m.hameleers@uva.nl
Toni Van der Meer: g.l.a.vandermeer@uva.nl
Tom Dobber: t.dobber@uva.nl
Date submitted: 2024-01-15

Despite these concerns, empirical evidence on both the prevalence of deepfakes (e.g., Brennen, Simon, & Nielsen, 2021) and their persuasiveness (e.g., Hameleers, van der Meer, & Dobber, 2022) is scarce. Although political deepfakes may indirectly result in distrust and confusion (Vaccari & Chadwick, 2020), they may not be more persuasive and credible than textual disinformation. To date, however, empirical research on the impact of deepfakes has mostly relied on a quantitative assessment of credibility, trustworthiness, misperceptions, or beliefs in response to deepfakes. Existing qualitative evidence on the reception of deepfakes suggests that people are likely to doubt deepfakes when the arguments forwarded in the audiovisual disinformation do not closely match the political positions and prior speeches of the depicted actor (Hameleers, van der Meer, & Dobber, 2024). But how may doubts and responses to deepfakes differ across different levels of deception?

Crucially, although extant research shows that deepfakes may in some conditions be regarded as credible and doubted in other settings, we currently lack an understanding of the reasons why people doubt deepfakes across levels of content manipulation and extremity of the content. To understand the political consequences of emerging technologies in AI and deepfakes, we need to arrive at a more comprehensive and detailed overview of users' perceptions of deepfakes and the reasons why they accept or resist the deceptive message.

This study makes different contributions to the literature on disinformation and deepfakes in particular. First, moving beyond research that quantitatively assessed the credibility of disinformation and deepfakes (e.g., Hameleers et al., 2022), our qualitative assessment of unprimed responses to deepfake videos allows us to assess whether people find deepfakes (in)credible based on features of the content, source, ideological bias, or presentation mode. Second, deepfakes may manipulate or fabricate information to different extents and may deviate more or less from familiar issue positions to deceive recipients. Yet, to date, most experimental studies have only investigated one manipulation that strongly deviates from familiar positions (e.g., Vaccari & Chadwick, 2020). Offering more insights into people's unprimed responses to different degrees of content manipulation contributes to a more comprehensive mapping of the boundary conditions of the acceptance of AI-powered disinformation.

To offer in-depth insights into users' responses to different political deepfakes, the following research question was central in this study: How are different degrees of deception in deepfakes perceived by recipients? We specifically rely on an online experiment in which participants were randomly exposed to one of three different political deepfakes that varied in the level of plausibility and manipulated partisan bias. We use an unprimed, open-ended measure of responses to these different deepfakes (i.e., thought listing), which we recoded into different categories of deepfake acceptance or rejection in subsequent quantitative analyses. As a major contribution to the disinformation literature, we offer an inventory of the different unprimed responses yielded by deepfakes that vary in extremity—which can be used to offer a better understanding of the vulnerability, resilience, and political consequences of deepfakes in a global digital setting.

Political Deepfakes as Partisan Disinformation

Different from misinformation that refers to unmotivated errors, disinformation involves the intentional manipulation and dissemination of false information to achieve certain preconceived goals (e.g.,

Bennett & Livingston, 2018; Freelon & Wells, 2020). Disinformation can be disseminated for various reasons, such as financial gain (i.e., in health disinformation, clickbait coverage), increasing cynicism in foreign democracies, or seeking support for conspiracies and misperceptions aligning with (partisan) political agendas. In this study, we focus on political disinformation that is created to make an existing established political actor look bad by delegitimizing their credibility, loyalty, and trustworthiness. This type of disinformation is in line with the prevalence of many radical right-wing disinformation campaigns that seek to attack and harm established political actors to increase polarized cleavages and create momentum for alternative radical movements opposed to the established order (e.g., Bennett & Livingston, 2018).

In this study, we specifically investigate the effects of political disinformation that is generated using artificial intelligence (AI) and deep-learning techniques: so-called deepfakes (also see e.g., Westerlund, 2019). Generally speaking, deepfakes can be regarded as synthetic multimedia content generated and disseminated with the intention to deceive (Hancock & Bailenson, 2021). Deepfakes can be based on visual, audio, or video manipulations. As video-based deepfakes are regarded as most threatening and representing the most prevalent modes of manipulation, this study focuses on the reception of deepfakes based on synthetic videos in which both the audio and the visual components are manipulated using deep-learning techniques (also see e.g., Dobber, Metoui, Trilling, Helberger, & de Vreese, 2020; Lee & Shin, 2021).

Deepfakes have been associated with far-reaching consequences for society and democracy (e.g., Weikmann & Lecheler, 2023). Hence, as they offer seemingly realistic depictions of real people saying or doing things, deepfakes may result in misperceptions among recipients. Yet, to date, the strong concerns about deepfakes in the political sphere have not been backed up by strong empirical evidence (see e.g., Dobber et al., 2020; Vaccari & Chadwick, 2020). Although some studies have found that deepfakes and video-based disinformation in general can be more credible than textual disinformation (see e.g., Lee & Shin, 2021), others found that deepfakes in the political realm do not deceive recipients directly (Vaccari & Chadwick, 2020).

However, given that extant research mainly relied on a quantitative assessment of credibility, issue agreement, or political evaluations, we know little about why news users may accept or reject deepfakes' deceptive messages. Based on the open-ended analysis of responses to one single deepfake in the Netherlands, Hameleers et al. (2024) found that people may be likely to base their acceptance or rejection of a deepfake on the congruence between the disinformation narrative and the profile of the political actor. In their study, most people doubted the deepfakes because they deviated strongly from the prior statements and political profiles of the depicted actors. By differentiating between the assessment of the credibility of the content, source, and presentation of deepfakes, this study contributes to a more comprehensive understanding of the multifaceted nature of deepfakes' acceptance or rejection. Such a more profound understanding can help us to better understand why people may be vulnerable to AI-generated disinformation and can inform media literacy interventions tailored to the specific motivations to fall for or resist disinformation.

In the current digital information ecology where citizens are presented with different truth claims that compete for legitimacy (e.g., Bennett & Livingston, 2018; Waisbord, 2018) citizens may be overloaded, which impedes their capacity to process all information critically and systematically (Lang, 2000). Hence, to navigate the digital media ecology, media users have to take shortcuts, as they do not have sufficient

cognitive capacity to pay attention to the content and arguments of all the information they see. Video-based disinformation may especially cause cognitive overload, as the rich mode of presentation may take up many cognitive resources, which comes at the cost of systematically processing message arguments (Sundar, 2008). To deal with this overload, news users can use a mental shortcut referred to as the realism heuristic (Sundar, 2008). Basically, this shortcut can be understood as "seeing is believing": Information that uses rich and vivid audiovisual depictions of reality is more likely to be deemed as trustworthy than written abstractions of it. Because deepfakes, more than textual disinformation, directly show events and speeches, they may offer a cue for realism that can sidestep the critical evaluation of the content of disinformation statements (also see Vaccari & Chadwick, 2020).

The effectiveness of this heuristic can also be understood in light of the truth-default theory (TDT; Levine, 2014). This theory postulates that people are, on average, more likely to accept the trustworthiness of information than to doubt it, unless suspicion is actively triggered by the information or the context in which it is presented. Triggers of suspicion may, among other things, consist of a strong discrepancy between known beliefs and novel information, or the extremity of content. As video-based disinformation such as deepfakes motivate heuristic processing via the realism heuristic, suspicion may be sidestepped. Hence, because of the direct depiction of reality, people may be less likely to pay attention or critically evaluate the deviation between the political reality and the content of the manipulated speech.

In mapping responses to different deepfakes that vary in the level of content plausibility and political extremity, we are primarily interested in the extent to which people trust or distrust the deepfake, and what reasons they give for resisting or accepting the different manipulated messages. Considering that asking for credibility, truthfulness, accuracy, or related concepts directly may trigger suspicion, we believe that open-ended and unprimed questions may best capture the likelihood of and reasons for deception detection. In line with research on distinctions between mis- and disinformation, there may be different perceived causes for untruthfulness (e.g., Bennett & Livingston, 2018). Given the previously identified distinction between unmotivated errors (misinformation) and intentional deception (disinformation), recipients of deepfakes may distinguish between these different responses (e.g., Hameleers et al., 2022). Hence, some recipients may deem deepfakes as untrustworthy because they lack in facticity, whereas others may emphasize that the message is deceptive and created with harmful intent. Yet, to date, extant research has not explored the multifaceted nature of the credibility of AI-powered disinformation, such as deepfakes. Against this backdrop, we forward the following exploratory research questions:

RQ1: To what extent and how do the responses to deepfakes reflect acceptance or rejection of the deepfakes' authenticity and trustworthiness?

RQ2: What causes and reasons do recipients identify for the acceptance or rejection of deepfakes?

The Role of Content Plausibility and Deception Detection

In this study, we expose people to deepfakes that vary on the level of content plausibility and partisan biases. More specifically, in the bipartisan setting of the United States, we have created a more substantive plausible deepfake (Nancy Pelosi in the United States talks about how Republicans and

Democrats should come closer together) that we contrast to two less plausible deepfakes (Nancy Pelosi is shown attacking her own party or praising opposed partisans). The plausible deepfake resembles statements that Pelosi has actually voiced over the past years. Therefore, these statements are in line with the associations that people may have stored about this well-known political figure. Based on the truth-default theory and deception detection literature (Levine, 2014), we expect that the more plausible deepfake is less likely to trigger suspicion, which should motivate heuristic processing according to the "realism heuristic" offered by the video-based presentation (Sundar, 2008).

The less plausible deepfakes—in contrast—may cause more suspicion, as they present conflicting partisan information that delegitimizes a well-known political actor. Therefore, recipients may deviate from the truth-default state and process the implausible arguments more critically. Hence, when people are confronted with highly unlikely statements that run against their prior associations and encounters with speeches from the political actor, deepfakes may not be able to motivate processing via a realism heuristic. To test the role of plausibility, we will recode the open-ended responses to the acceptance/rejection of the deepfake's plausibility and authenticity and assess the extent to which participants exposed to implausible deepfakes are more likely to doubt its plausibility than participants exposed to the plausible deepfake:

H1: Implausible deepfakes are more likely to trigger suspicion than a plausible deepfake.

Motivated Reasoning: Accuracy and Defensive Motivations

The acceptance of disinformation as well as perceptions of untruthfulness may be highly contingent on motivated reasoning biases (e.g., Schaewitz, Kluck, Klösters, & Krämer, 2020). Moreover, people are more likely to reject corrections of misinformation when their partisan identities align with the false claims that are refuted (e.g., Thorson, 2016). This can be explained in light of partisan motivated reasoning: Evidence from congenial sources is more likely to be trusted and accepted than information from uncongenial sources (e.g., Bolsen, Druckman, & Cook, 2014).

This tendency can be motivated by the need to defend prior beliefs and identities when they are under attack, which has been understood as defensive motivated reasoning (Kunda, 1990). As people are psychologically wired to avoid the discomfort caused when being confronted with discrepant views, people may reject or counterargue discrepant views and accept congenial information (Festinger, 1957). Even though evidence on pervasive partisan motivated reasoning is mixed (see e.g., Garrett, 2009), there are strong indications that people are more likely to support false information or conspiracies that attack the opposed party, whereas deception targeting in-group members is more likely to be rejected.

Directional motivated reasoning may not be driven only by the need to be consistent and to confirm prior beliefs but can also be directed by the need to hold accurate beliefs (Kunda, 1990). Hence, people may strive toward the most accurate beliefs and cognitions, and process information in a way to confirm beliefs of accuracy (even if this is a perception of accuracy instead of facticity). In line with this alternative explanation of biased processing, Pennycook and Rand (2019) found that, instead of defensive motivated reasoning, disinformation has the strongest effects among people who do not engage in analytical reasoning.

In line with these alternative explanations of the role of motivated reasoning, we will explore the extent to which responses to deepfakes in terms of acceptance and rejection are driven by accuracy versus defensive motivations. We will explore this using both a qualitative and quantitative approach. First, the open-ended responses to the deepfakes will be analyzed qualitatively for indications of defensive and accuracy motivations, answering the following question:

RQ3: To what extent and how do responses to deepfakes reflect accuracy and defensive motivations?

As a second approach, we will rely on a recoding of the open-ended responses in terms of the acceptance versus rejection of the deepfakes. Based on preexposure measures of partisanship and analytical thinking, we will quantitatively assess the role of both indicators of motivated reasoning. Based on the findings of Pennycook and Rand (2019) and previous research on the role of partisan motivated reasoning (e.g., Thorson, 2016), we forward the following hypotheses:

H2: Deepfakes are more likely to be accepted and regarded as trustworthy when they resonate with recipients' partisan identities.

H3: Participants with higher accuracy motivations are more likely to reject a deepfake than participants with lower levels of analytical thinking.

Method

Design and Manipulations

To explore recipients' unprimed responses to plausible and implausible deepfakes, we collaborated with a visual effects and computer vision specialist that created three different deepfakes for this project. Using deep-learning techniques, three synthetic videos of Nancy Pelosi (speaker of the United States House of Representatives at that time) were created and contrasted with one authentic video of the same political actor (the control condition). In line with literature conceptualizing deepfakes as synthetic AI-generated multimedia in which a real person is made to say or do things that this person has not said or done in reality (e.g., Westerlund, 2019), we used fictitious narratives that were manipulated to make it seem as if Pelosi had expressed them. In line with literature on political disinformation, we used a delegitimizing narrative that could intentionally cause harm by delegitimizing the depicted politician (e.g., Dobber et al., 2020).

With this experiment, we wanted to explore how implausible the deepfake narrative can be to be credible and trustworthy. As deepfake techniques may motivate heuristic processing by signaling realism (Sundar, 2008), malign actors may get away with rather extreme manipulations. At the same time, recipients who are familiar with the political setting of the disinformation campaign may become suspicious when narratives deviate too far from the stored associations related to well-known political figures. To test the conditions under which suspicion is triggered, we specifically contrasted more implausible deepfakes (with an anti-Democrat or anti-Republican narrative) to a more plausible deepfake in which Pelosi expressed

nonpartisan views that did not delegitimize her own party or praised the opposing party. The scripts of the stimuli are stored online.[1]

For video-based stimuli, it is challenging to isolate the independent variables and make sure that all other factors remain stable across conditions. For this reason, we made sure that the background, tone, valence, length, and other factors of the speech were constant across conditions (i.e., all videos showed Pelosi in the same background using the same tone of voice). Hence, as the deepfakes were based on the same real footage, the control and experimental conditions did not differ in the presentation of the actor, background, speech, and other factors. We can also confirm that the stimuli were rated as similar in negativity, quality, and seen as comparable with political videos participants normally come across in daily life. In post hoc tests, there were no differences in perceived negativity/positive valence, emotionality, and similarity to information encountered in (digital) news.

Because of ethical considerations, data protection, and copyright concerns, we cannot share the actual deepfakes created for this experiment here. Hence, the risk that deepfakes of a political actor would circulate outside of the experimental setup had to be prevented. Specifically, although the impact of isolated misinformation messages may be limited, the association of an existing and well-known political figure with deceptive and delegitimizing statements outside of the context of a controlled experiment and without a clear debriefing being present comes with various risks. That being said, the study received ethical approval from the University of Amsterdam. We further relied on an extensive controlled debriefing procedure that ensured participants did not leave the survey with misperceptions related to the manipulated information.

Dependent Variable: Responses to the Deepfakes

After exposure to the deepfake, we simply asked participants to write down their thoughts on the video: "In general, what are your thoughts about the video you just saw? Please write down your thoughts in a few sentences." We intentionally did not refer to credibility, trustworthiness, or believability, as we aimed to circumvent triggering suspicion by the formulation of the item. Responses to the open-ended question were analyzed in two different steps. First, a three-step qualitative coding approach was taken in line with the data-reduction procedures of the Grounded Theory (GT) approach (e.g., Charmaz, 2006). We followed a "lite" approach to GT as we used only the open, focused, and axial coding steps instead of relying on the wider epistemological framework of the approach (including line-by-line coding, cyclic-iterative data collection, saturation, and theory formation).

During open coding, we highlighted relevant fragments of participants' responses and labeled them in line with the sensitizing concepts of trust, credibility, authenticity, manipulation, and (dis)agreement. During the next step of focused coding, we reduced the long list of open codes and labels to more overarching themes and categories that offered more analytic insights into the different ways in which credibility and authenticity were perceived. Finally, during axial coding, we looked at connections

[1]https://www.dropbox.com/scl/fi/pz0logldnk3qxesjaqd82/Appendices-A_deepfakes.docx?rlkey=x586bd15ihdnwaye7t4zwt7c6&dl=0

between themes and categories in light of the research questions (i.e., how agreement evaluations connected to trust in the authenticity of information). Peer debriefing was used to assess the trustworthiness of the qualitative coding process, which meant that a second independent coder inspected the raw materials, open codes, and data-reduction procedures, and evaluated the process in terms of construct validity and traceability. Although traditional inter-coder-reliability (ICR) indices are not suitable for the nature of the qualitative data, we did subject the recoding of qualitative data into categories to stricter ICR tests. Specifically, 10% of all items were manually coded by two coders (percentage agreement 95%, Krippendorff's alpha .88).

To test the hypotheses and to arrive at a quantitative assessment of the responses to the deepfake, the open-ended responses were finally recoded into categories of credibility and trustworthiness. Based on the final themes of the qualitative analyses, a codebook with four major categories was used: (1) the acceptance of the authenticity and truth value of the video; (2) doubts in the authenticity and truth value of the video; (3) rejecting the authenticity and truth value of the video, and (4) responses that could not be classified as authenticity or credibility evaluations (i.e., a rejection of the politician, hostile comments, or disagreement with the arguments voiced). As a second variable, we coded for the reasons why people accepted/doubted or distrusted the video. Again, inspired by the themes of the qualitative analyses, we looked for the difference between doubts related to the perceived artificiality of the message, the dishonesty of the political actor, the discrepancy between the known statements of the political actor, and doubts related to disagreement with or the opposition to the depicted politician.

Moderators

Partisan identities were measured with the question "Generally speaking, do you think of yourself as a Republican, a Democrat, or an Independent?" (1 = Strong Republican, 2 = Republican, 3 = Independent closer to the Republican Party, 4 = Independent, 5 = Independent closer to the Democrat Party, 6 = Democrat, 7 = Strong Democrat). We considered scores between 1 and 3 as Republican (40.4%) and scores 5 through 7 as Democrat (36.5%) for the regression models used to test H3. Independents (23.2%) were not used for these analyses.

Accuracy motivations were measured using a four-item scale (all items were measured on a 7-point scale, ranging from (1) completely disagree to (7) completely agree). The statements were formulated as follows: (1) It is important to find information that corrects misconceptions that I have; (2) I seek information to see both sides of an issue; (3) Information helps me to determine what is true and what is not true; (4) I often find myself trying to decide whether information is accurate or not (M = 5.39, SD = 1.18, Cronbach's alpha = .861). We used media literacy, education, and the strength of people's political beliefs as additional covariates in robustness checks.

Ethical Considerations and Procedures

Two main considerations were central in the ethical procedures: (1) participants needed to be carefully debriefed about the synthetic nature of the video and fabricated statements and (2) the deepfakes needed to be presented only in the controlled setting of the experiment and were not shared, downloaded,

or stored in a nonencrypted way. These main considerations were included in the ethical documentation that received approval from the ethical research board of the University of Amsterdam.

About the first issue, all participants received a very extensive debriefing message only a few minutes after seeing the deepfake, which directly corrected misperceptions and negative evaluations that could result from the video. This debriefing fact-checked all statements, clearly emphasized that the video was a deepfake, and offered instructions on how participants could detect and identify deepfakes. A follow-up question ensured that people recognized the video as a deepfake and did not consider it as a real political speech. Participants were also presented with additional information on deepfake detection. About the second issue, the deepfakes were presented only within the secured environment of the experimental modules and could not be downloaded or shared outside of this setting.

Sample

Data collection was conducted by the international research agency Kantar that fielded the survey among a representative sample of U.S. citizens in July 2022. We achieved 601 valid completes (completion rate 85.0%). Of these completes, 49.4% identified as female. Higher-educated and lower-educated participants were equally represented, with 23.3% and 29.0% respectively (47.8% had a moderate level of education). Of the total sample, 33.9% identified as Republican, and 39.8% as Democrat (26.3% as Independent). The mean age of participants was 52.51 years (SD = 16.91). By and large, the diversity captured in our sample represents the distribution on key demographics in U.S. society. Hence, the deviations from census data are minimal, and key political background variables are distributed according to the main variety in U.S. society. Although representativeness was not the aim of this experiment, we have conducted robustness checks with weights to fully match the sample with the actual distributions in the population. No differences were found.

Manipulation Checks

At the end of the survey, participants were asked to remember the content of the videos they were exposed to, which allowed us to differentiate between the statements of the different deepfakes and the control conditions (i.e., they had to identify whether the video was attacking Democrats, expressed sympathy toward Republicans, or whether the video contained a message saying that the different sides should come closer together). We specifically asked whether people accurately remembered the narrative of the (1) plausible depolarizing deepfake ($\chi^2(3) = 161.74$, $p < .001$); (2) the implausible in-group attack ($\chi^2(3) = 237.81$, $p < .001$) and the (3) implausible out-group sympathy deepfake ($\chi^2(3) = 201.34$, $p < .001$). The chi-square tests indicate a significant and strong association between identifying the correct narrative in the respective conditions participants were assigned to. More specifically, 71.7% of all participants in the depolarizing plausible deepfake condition correctly identified the narrative, 73.2% of all participants exposed to the implausible in-group attack identified the narrative, and 74.8% of the participants exposed to the implausible out-group sympathy deepfake correctly identified the narrative. Overall, most participants could correctly retrieve the narrative of the deepfake they saw.

Results

Before discussing recipients' responses to the various deepfakes in detail, we will offer some descriptive statistics on the rating of the deepfakes here. More specifically, after recoding the open-ended responses into four categories—(1) accepting the truthfulness and authenticity of the video, (2) doubting the video's veracity and authenticity, (3) (deep)fake classification, and (4) other responses—we can conclude that 38.5% of all participants exposed to implausible deepfakes accepted the authenticity of the message. For the plausible deepfake, however, most recipients accepted the authenticity and honesty of the synthetic video (69.6%).

Looking at the reasons why people doubted the deepfakes (RQ1), the lack of sincerity and honesty of Pelosi was mentioned frequently by participants exposed to the plausible deepfake (40.0%), but less likely to be mentioned by participants exposed to an implausible deepfake (29.6%). However, participants exposed to an implausible deepfake more frequently referred to the allegedly artificial and seemingly manipulated nature of the video (41.8%) than participants exposed to the plausible deepfake (15.6%). Finally, although the perceived discrepancy between the viewpoints expressed in the synthetic video and the actual opinions and positions of Pelosi were rarely mentioned in response to a plausible deepfake (5.8%), this was a likely reason for doubt when participants were exposed to an implausible deepfake (47.1%).

Table 1 offers an overview of the main categories of doubt identified in response to the deepfakes. The summarized findings show three core causes of doubt: (1) the message does not reflect the true beliefs of the politician, although it is not manipulated; (2) the message seems artificial and manipulated, for example, based on nonmatching audio and video; (3) the message seems fake as the viewpoints expressed are at odds with the viewpoints of the politician. Only the last two reasons relate to perceived manipulation and artificiality of the video.

Table 1. Overview of the Main Responses to Deepfakes of Participants Expressing Doubt in the Deepfake.

Reason for doubt	Associated condition	In-vivo quote
Lack of sincerity and honesty of Pelosi as a politician (no actual manipulation)	Plausible deepfake	"It felt like she was acting these views instead of being truthful about what the Democratic Party believes in."
The artificial and seemingly manipulated nature of the video (manipulation)	Implausible deepfake	"Absolutely not a true and accurate recording of Nancy Pelosi, so it must be a deep fake."
Perceived discrepancy between the viewpoints expressed in the synthetic video and the actual opinions and positions of Pelosi (manipulation)	Implausible deepfake	"I don't believe for a second that she would say something like this. She has no respect for Donald Trump or our country. She is all about money."

Responding to Implausible Deepfakes

Importantly, there are no substantial differences in responses to the two different implausible deepfakes, which either reflect support for out-party members or disapproval of in-party members. The main reason to reject the deepfake or to doubt its authenticity and credibility is the perceived gap between Pelosi's prior viewpoints and speeches and the implausible deepfakes' attack on in-group partisans or the approval of out-group partisans.

Although some participants expressed certainty when expressing distrust toward the authenticity of the video they saw, many were less certain and expressed doubts related to the credibility and authenticity. This illustrates that the deepfake may cause suspicion, although recipients were unsure about its authenticity: "It seemed fake to me. I doubted that was really Nancy Pelosi." Other participants expressed uncertainty as the viewpoints seemed out of touch with reality and different from all the things the politician said before: "I doubt very much that she felt that way by what she has said in the past. This is a complete reversal of what she has said and done before, so I even doubt it was her." Other doubting participants clearly expressed the need for further investigation and verification: "I can't help but wonder if she really expressed those views or if this is a doctored video. I would question it if I saw it online and maybe do some more research to see if it's legit." Doubt was not always related to the authenticity of the video or its artificiality but also related to the sincerity and honesty of Pelosi's expression (also see Table 1). Thus, even considering that some participants doubted whether the views matched the politician's beliefs, they still accepted the authenticity of the video.

The main reason participants highlighted for rejecting the video's trustworthiness related to a perceived discrepancy between the conventional viewpoints of the depicted politician and the manipulated statements in the deepfake. As one recipient of the implausible deepfake mentioned: "I don't believe for a second that she would say something like this. She has no respect for Donald Trump or our country. She is all about money."

When participants accurately detected the synthetic nature of the video, they mostly mentioned that the voice or movements of the depicted actor were not authentic, resulting in the judgement that the video should be a deepfake: "Absolutely not a true and accurate recording of Nancy Pelosi, so it must be a deep fake." Even though some participants pointed to the high realism of the manipulation, they still recognized the synthetic nature: "That wasn't Pelosi speaking; pretty good fake, though." Thus, even though recipients did not directly see flaws in the manipulation, some of them could clearly identify that the depicted speaker of the video was not the original political actor: "The facial characteristic employed by the actor are similar in nature to Pelosi's and the actor's face is similar, but again the person in the video is NOT Nancy Pelosi."

We can also identify a group of participants that did not express doubts or uncertainty related to the implausible deepfakes. They either explicitly mentioned that the video was true or expressed their opinions about the depicted statements as if they were genuinely coming from the depicted politician. For example, one participant mentioned the veracity of the statements: "The speech was very true in its context, and she conveyed the facts." Other participants explicitly expressed approval of the statements expressed

by Pelosi: "I think it is remarkable that she does this. I honestly believe her. You need to show understanding for segments of the population that feel left behind. Very strong statement of Nancy!"

Responding to a Plausible Deepfake

Substantially different from participants' responses to the implausible deepfake, only a few participants expressed doubts in the veracity or authenticity of the plausible deepfake. Only 7.4% doubted the authenticity of the deepfake, whereas 3.7% claimed to be sure that the video was inauthentic or fake (for the implausible deepfakes, 18.7% of participants doubted its authenticity and 20.5% were sure that the video was fake). A logistic regression analysis confirmed these findings: Participants exposed to an implausible deepfake were significantly and substantially more likely to doubt the veracity of the message than participants exposed to a plausible deepfake (B = 1.63, SE = .20, p < .001, odds ratio = 5.13, 95% CI [3.45, 7.63]. These findings are in line with the first hypothesis.

Our in-depth qualitative analyses reveal that most responses to the plausible deepfake did not refer to the level of authenticity of the actual video, but they did express doubts about the sincerity of Pelosi and whether the viewpoints expressed were actually reflecting her positions. This might be a result of that; even though the statement was plausible, it still slightly deviated from Pelosi's normal statements. As one participant formulated it:

> I feel like Nancy Pelosi is delivering a message that she does not truly believe in. There is a disconnect between politicians and the people who chose them to serve this country. I believe the message though, because if we (Democrats and Republicans) can't see eye to eye on decisions then nothing will get done to better society as a whole.

In a similar vein, some participants felt that the message delivered by the deepfake was true, and even agreed with it, but still experienced a disconnect between the political actor's actual opinions and viewpoints and the speech delivered. As one participant said: "This comes directly from a psychology book, even though the text is completely right." Other participants mentioned that it felt as if Pelosi was reading out a "script," that the text was "rehearsed," or even that she was "forced" to deliver this speech that did not reflect her true opinions. As one participant expressed: "She looked a little weird, like she was saying this under duress." Another respondent put it even more bluntly: "I thought that she seemed like she was forced to make that video with a gun." Hence, different from the implausible deepfakes that used conflicting partisan cues, recipients of the plausible deepfakes mostly agreed with the message and believed it was part of an authentic speech, but they still doubted the honesty of the depicted political actor.

To conclude, when the deepfake is more plausible, the comments mostly do not refer to the level of trustworthiness, authenticity, or accuracy of the video but rather doubt the sincerity of the opinions voiced by Pelosi, or reflect recipients' (lack of) agreement with the statements voiced. When people are exposed to a more plausible deepfake, they are more likely to cast doubt on the sincerity of the statements voiced by the politician than the authenticity of the video and the manipulation involved. The level of suspicion triggered by the video is therefore more likely to be related to sincerity than authenticity: When the deviation

from facticity is moderate and when the deepfake does not include implausible partisan cues, recipients may not recognize the video as a deepfake and regard the video as an authentic speech in which a real political actor voices things they do not really mean—for example, to gain politically or because he or she is forced to say things he or she does not mean. These findings offer an important contribution to the disinformation literature by revealing the complexities of credibility assessments: They can relate to the authenticity of the manipulation, as well as the match between the political reality and the statements of the speech or the sincerity of the depicted actor.

Motivated Reasoning Reflected in Responses to Deepfakes

Many participants explained that the speech was not authentic because the statements did not accurately reflect the depicted politician's profile. Even though participants expressed uncertainty, they clearly articulated the reasons why they believed that the video was inaccurate. Participants also deconstructed various elements of the message—such as the accuracy of the voice and lip movements—to arrive at a verdict of the authenticity of the speech: "There were times in the video when the lips moving did not accurately reflect the words heard. Trying to get in good graces with people who don't like her." As this quote further illustrates, some participants also aimed to reveal the motives for manipulation, trying to arrive at a justification why the deepfake was constructed.

Although such accuracy motivations prevailed, some participants also rejected the message because they did not like or disagreed with the depicted political actor: "I am not the biggest fan of Nancy Pelosi. She flip flops on the issues." In rejecting the deepfakes, some participants clearly expressed their partisan identities and disapproval of Pelosi as an opposed political actor:

> Pelosi is too far removed from the people who elected her to be effective. She got elected by lying to her constituents. She only does what she wants and does not listen to the voices of those who elected her.

Others phrased their motives more directly: "I don't think about what she says. I don't like her." Some participants expressed their rejection in emotional partisan tones, defending their in-group partisans while delegitimizing the depicted politician: "This woman that tore up President Trumps speech on national TV means nothing to me!" Some responses were even uncivil in tone:

> I think it is all bull shit, she is a disgrace to this country. Foreign countries laugh at her. I can't stand to listen to a LIAR!" and "Everything that comes out of that clown's mouth is a joke. She is full of ****.

Based on an analysis of the evaluations, it can be concluded that accuracy motivations (i.e., the rejection of the credibility or truth value of deepfakes based on systematic argumentations on veracity and authenticity) outweighed defensive motivations (i.e., rejecting the deepfake based on disliking or disapproving of the depicted political actor). Hence, whereas 36.2% of all responses to the deepfakes reflected accuracy motivations, only 4.1% of the responses reflected defensive motivations. To look at the role of motivated reasoning in a more indirect way, we explored the role of partisan

identities (H2) and accuracy motivations (H3) in the acceptance and rejection of deepfakes. Findings from a binary logistic regression model in which the rating of the video as authentic/true (versus inauthentic/false) was inserted as the outcome variable first indicated that participants were significantly less likely to classify a deepfake than an authentic video as authentic or true ($B = -2.66$, $SE = .25$, $p < .001$, odds ratio = .07, 95% CI [.04, 14]). Partisanship did not play a role in this classification: Democrats were not more likely to classify the video as a deepfake than Republicans ($B = 1.02$, $SE = .82$, $p = .214$), and the correct identification of a deepfake as inauthentic was not moderated by partisan identity ($B = -1.19$, $SE = .84$, $p = .156$).

Turning to H3, we find that the higher people's accuracy motivations, the less likely they were to classify the video as a deepfake ($B = -.14$, $SE = .07$, $p = .049$, odds ratio = .87, 95% CI [.76, 99]). Yet, the nonsignificant interaction effect ($B = .39$, $SE = .39$, $p = .325$) between exposure to a deepfake versus an authentic video and accuracy motivations on authenticity ratings indicate that higher levels of accuracy motivations do not result in a better ability to discern a deepfake from an authentic video. If we use a different proxy for analytical thinking—the need for cognition scale—the findings are identical. We thus also have to reject H3.

Although participants were randomly assigned to the conditions (post hoc randomization checks confirmed that this succeeded), we additionally controlled for other factors related to the credibility of deepfakes: education, media literacy, familiarity with the political actor, and strength of political beliefs. Adding these variables as covariates did not impact the findings reported here. Overall, we do not find evidence to support H2 or H3.

Discussion

Although political deepfakes are surrounded by popular concerns on their detrimental consequences, empirical evidence on their impact is scarce. We specifically lack in-depth insights into people's responses to different expressions of political deepfakes and the extent to which the level of content plausibility of the manipulation matters for its acceptance. Against this background, this study uses an experiment in which people were exposed to three different deepfakes that varied on the level of plausibility and the direction of the partisan attack manipulated. Our main findings indicate that plausible deepfakes (i.e., reflecting issue positions that are not contradicting the partisan identity of the depicted politician) are significantly and substantially more likely to be accepted than implausible deepfakes (i.e., reflecting viewpoints opposed to the politician's partisan identity). Whereas many recipients pointed toward the discrepancy between the actual viewpoints previously articulated by the depicted politician and the deceptive narrative of the deepfake in the implausible conditions, most recipients accepted the plausible deepfake.

These findings extent beyond the qualitative evidence provided by Hameleers and colleagues (2024), who found that people may resist deepfakes because they recognize the discrepancy between the political profile and the narrative of disinformation. In this study, we show that more credible forms of manipulation are less likely to be detected. As an important theoretical contribution, it can be argued that the strong concerns on the effects of deepfakes voiced in the literature should be contextualized. Although

deepfakes may be harmful when they use subtle manipulations that delegitimize political actors by taking their statements out of context, deepfakes may not be able to credibly associate politicians with extreme political claims that counter their political profile.

This highlights that there are boundary conditions to the acceptance of AI-generated disinformation. These boundary conditions may reflect the different varieties of audiovisual disinformation identified in extant literature (e.g., Weikmann & Lecheler, 2023). This emphasizes that it is crucial to look at the "gray areas" of disinformation and move beyond the most likely cases of extreme disinformation narratives: The decontextualization of information to deceive may be more effective and credible than blatant lies. Indeed, as most visual disinformation is more likely to be based on decontextualized information than the fabrication of narratives (e.g., Brennen et al., 2021), future research may need to more comprehensively explore the effects of different degrees of deception in (visual) disinformation.

These main findings support the deception detection framework articulated in the truth-default-theory (Levine, 2014). In line with this, people are intrinsically wired to accept the honesty and authenticity of information, unless suspicion is actively primed. The implausible deepfake may prime suspicion by presenting people with a counterfactual narrative, motivating them to more critically scrutinize the message. When such a trigger of suspicion is absent, as was the case for the more plausible deepfake, people may not look for faults in the message and may not detect the synthetic nature of the video. As video-based disinformation may overburden systematic processing while offering a mental shortcut that signals authenticity, people may be less likely to spot the deception. Our findings specify the conditions for deception detection in response to video-based disinformation. More specifically, the deviation from actual viewpoints and familiar statements voiced by the depicted actor should not exceed a plausible threshold.

As a key implication, then, our findings may offer first insights into the boundary conditions of plausible deepfake manipulations. The much-held assumption that deepfakes can be used to "make everyone say anything" is inaccurate in the sense that the manipulation still has to overcome the detection of deception. Even if people do not identify the artificiality of the actual AI-generated videos, they still rely on their previous encounters with a well-known depicted actor to arrive at an assessment of the message's trustworthiness. No matter how good AI developments become in the future, the actual content and message arguments may trigger suspicion when they do not resonate with prior held beliefs. The real danger of deepfakes may therefore be found in the more subtle manipulations that gradually delegitimize a (political) actor by distorting or decontextualizing the truth.

This has important implications for media literacy education and fact checking. News users should not only be warned about highly deceptive claims and extreme statements in disinformation but also be made aware of how subtle manipulations may be used in disinformation, for example, related to the decontextualization of authentic information and political claims (Weikmann & Lecheler, 2023). Fact checkers, in turn, may warn news users about the subtle techniques of manipulation used in audiovisual disinformation, which could help to prebunk deepfakes that rely on hard-to-detect deviations from reality (also see e.g., Çömlekçi, 2022). For different interventions, it is crucial that news users are made aware of

the context in which deceptive statements are used. Even though manipulations may be subtle and close to reality, knowing why deceptive claims were made and how these resonate with the political goals of disseminators may instill more resilience to subtle forms of manipulation.

Contrary to our expectations, our findings did not indicate accuracy or defensive motivated biases in the classification of deepfakes. Republicans and Democrats were equally likely to accept or reject deepfakes, and higher levels of need for cognition or accuracy motivations did not result in a better capability of distinguishing a deepfake from an authentic political speech. This goes against the findings of Pennycook and Rand (2019), who found that a lack of reasoning makes people more susceptible to misperceptions in a textual disinformation context. We could explain this discrepancy as the consequence of the different processing routes of textual versus video-based disinformation (also see Sundar, 2008). Deepfakes, as a form of video-based disinformation, may motivate heuristic processing because of the realism index they offer (Sundar et al., 2021). Irrespective of people's partisan biases and motivated reasoning, the seemingly authentic depiction of an unfiltered reality may be deemed as accurate—unless the deviation from familiar viewpoints gets too extreme.

This study has several limitations. First, we were able to use only limited disinformation narratives. We contrasted a rather plausible deepfake to very implausible deepfakes in a bipartisan setting, and it could be argued that we missed fewer extreme manipulations in between these categories. In a similar vein, we made a very well-known political figure express ideas that may clearly not reflect her positions. We suggest future research to explore the role of different levels of facticity and plausibility, as well as the extent to which the level of familiarity with the statements voiced by different actors matters for the detection of deception. For example, are people less likely to detect deception when they are exposed to implausible statements voiced by a less well-known political figure or a political actor who has less consistent opinions on different issues than the Democrat we used in this experiment? Another limitation of the study is the high mean age of participants. Arguably, younger generations have more technical and digital information literacy, which may enable them to better spot the differences between deepfakes and authentic information. To what extent is the detection of deepfakes contingent on the age of participants and their level of digital (or AI-related) information literacy?

Next to answering these questions, we believe that future research may apply comparative research to understand how deepfakes are received across countries with different levels of resilience to disinformation (Humprecht et al., 2020). Arguably, high-trust settings in which bipartisan cleavages are less pronounced may offer more room for the association of deceptive statements with different political actors. By repeating the experiment in different national settings, it can also be explored how ideology and existing political beliefs moderate the impact of deepfakes that vary in their ideological bias. As a final limitation, we did not contrast deepfakes to other modes of deception, such as cheapfakes or AI-driven manipulations of speech (e.g., Hameleers, 2024). We suggest future research to unpack the role of different modalities and techniques that can be used to trigger authenticity and circumvent the detection of deception.

References

Bennett, L. W., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication, 33*(2), 22–139. doi:10.1177/0267323118760317

Bolsen, T., Druckman, J. N., & Cook, F. L. (2014). The influence of partisan motivated reasoning on public opinion. *Political Behavior, 36*(1), 235–262. doi:10.1007/s11109-013-9238-0

Brennen, J. S., Simon, F. M., & Nielsen, R. K. (2021). Beyond (mis)representation: Visuals in COVID-19 misinformation. *The International Journal of Press/Politics, 26*(1), 277–299. doi:10.1177/1940161220964780

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. London, UK: SAGE Publication.

Çömlekçi, M. F. (2022). Why do fact-checking organizations go beyond fact-checking? A leap toward media and information literacy education. *International Journal of Communication, 16*, 4563–4583.

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. H. (2020). Do (microtargeted) deepfakes have real effects on political attitudes? *International Journal of Press/Politics, 26*(1), 69–91. doi:10.1177/1940161220944364

Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society, 23*(7), 2072–2098. doi:10.1177/1461444820925811

Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson.

Freelon, D., & Wells, C. (2020). Disinformation as political communication. *Political Communication, 37*(2), 145–156. doi:10.1080/10584609.2020.1723755

Garrett, R. K. (2009). Politically motivated reinforcement seeking: Reframing the selective exposure debate. *Journal of Communication, 59*(4), 676–699. doi:10.1111/j.1460-2466.2009.01452.x

Hameleers, M. (2024). Cheap versus deep manipulation: The effects of cheapfakes versus deepfakes in a political setting. *International Journal of Public Opinion Research*, *36*(1), 1–9. doi:10.1093/ijpor/edae004

Hameleers, M., van der Meer, T. G. L. A., & 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), 1–12. doi:10.1177/20563051221116346

Hameleers, M., van der Meer, T. G., & Dobber, T. (2024). They would never say anything like this! Reasons to doubt political deepfakes. *European Journal of Communication, 39*(1), 56–70. doi:10.1177/02673231231184703

Hancock, J. T., & Bailenson, J. N. (2021). The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 149–152. doi:10.1089/cyber.2021.29208.jth

Humprecht, E., Esser, F., & Van Aelst, P. (2020). Resilience to online disinformation: A framework for cross-national comparative research. *The International Journal of Press/Politics, 25*(3), 493–516. doi:10.1177/1940161219900126

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498. doi:10.1037/0033-2909.108.3.480

Lang, A. (2000). The limited capacity model of mediated message processing. *Journal of Communication, 50*(1), 46–70. doi:10.1111/j.1460-2466.2000.tb02833.x

Lee, J., & Shin, S-Y. (2021). 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. doi:10.1080/15213269.2021.2007489

Levine, T. R. (2014). Truth-Default Theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology, 33*(4), 378–392. doi:10.1177/0261927X14535916

Maras, M. H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. *The International Journal of Evidence & Proof, 23*(3), 255–262. doi:10.1177/1365712718807226

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188*(1), 39–50. doi:10.1016/j.cognition.2018.06.011

Schaewitz, L., Kluck, J. P., Klösters, L., & Krämer, N. C. (2020). When is disinformation (in) credible? Experimental findings on message characteristics and individual differences. *Mass Communication & Society, 23*(4), 484–509. doi:10.1080/15205436.2020.1716983

Sundar, S. S. (2008). *The MAIN model: A heuristic approach to understanding technology effects on credibility*. Cambridge, MA: MacArthur Foundation Digital Media and Learning Initiative.

Sundar, S. S., Molina, M. D., & Cho, E. (2021). Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication, 26*(6), 301–319. doi:10.1093/jcmc/zmab010

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication, 33*(3), 460–480. doi:10.1080/10584609.2015.1102187

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media and Society, 6*(1), 1–13. doi:10.1177/2056305120903408

Waisbord, S. (2018). Truth is what happens to news: On journalism, fake news, and post-truth. *Journalism Studies, 19*(13), 1866–1878. doi:10.1080/1461670X.2018.1492881

Weikmann, T., & Lecheler, S. (2023). Visual disinformation in a digital age: A literature synthesis and research agenda. *New Media & Society, 25*(12), 3696–3713. doi:10.1177/14614448221141648

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 39–52. doi:10.22215/timreview/1282