It's nothing but a deepfake! The effects of misinformation and deepfake labels delegitimizing an authentic political speech

Hameleers, M.; Marquart, F.

Publication date
2023

Document Version
Final published version

Published in
International Journal of Communication : IJoC

International Journal of Communication 17(2023), 6291–6311 1932–8036/20230005

It's Nothing but a Deepfake! The Effects of Misinformation and Deepfake Labels Delegitimizing an Authentic Political Speech

MICHAEL HAMELEERS♦
University of Amsterdam, The Netherlands

FRANZISKA MARQUART
University of Copenhagen, Denmark

Mis- and disinformation labels are increasingly weaponized and used as delegitimizing accusations targeted at mainstream media and political opponents. To better understand how such accusations can affect the credibility of real information and policy preferences, we conducted a two-wave panel experiment (N_{wave2} = 788) to assess the longer-term effect of delegitimizing labels targeting an authentic video message. We find that exposure to an *accusation* of misinformation or disinformation lowered the perceived credibility of the video but did not affect policy preferences related to the content of the video. Furthermore, more extreme disinformation accusations were perceived as less credible than milder misinformation labels. The effects lasted over a period of three days and still occurred when there was a delay in the label attribution. These findings indicate that while mis- and disinformation labels might make authentic content less credible, they are themselves not always deemed credible and are less likely to change substantive policy preferences.

Keywords: credibility, misinformation, disinformation, deepfakes, fake news labels

Amidst the Russian invasion of Ukraine in March 2022, both sides of the conflict frequently accused each other of spreading mis- and disinformation. As an example, an authentic video of Vladimir Putin that showed a glitch when his hand moved toward the microphone was discredited as a deepfake, and Putin was consequentially falsely referred to as a hologram by Ukraine partisans (Marty, 2022). This example illustrates that journalists are facing novel challenges during a time when the term "fake news" has become ubiquitous and weaponized (Van Duyn & Collier, 2019; Waisbord, 2018). Hence, the disinformation order does not only relate to the dissemination of false and misleading information but also reflects political and societal challenges related to increasing distrust in established knowledge and information (e.g., Bennett & Livingston, 2018).

Michael Hameleers: m.hameleers@uva.nl
Franziska Marquart: fm@hum.ku.dk
Date submitted: 2022-11-11

Although most empirical research to date has focused on the content and consequences of actual mis- and disinformation (e.g., Vaccari & Chadwick, 2020), deceptive information occurs only seldomly in people's media diets (see Acerbi, Altay, & Mercier, 2022; Jungherr & Schroeder, 2021). Arguably, discussions and labels *about* mis- and disinformation are more prominent than false information itself, and disinformation is even considered a moral panic by some (Jungherr & Schroeder, 2021). These labels, in turn, may have a strong impact on people's trust in (legacy) media and empirical evidence (Egelhofer, Boyer, Lecheler, & Aaldering, 2022; van der Meer, Hameleers, & Ohme, 2023; Van Duyn & Collier, 2019). In addition, focusing on a relatively marginal phenomenon such as disinformation may overlook the structural driving forces that cause the wider epistemic crisis we are facing (Jungherr & Schroeder, 2021). However, we currently lack insights on the effects of different types of mis- and disinformation accusations and labels attacking authentic information, with a few exceptions (e.g., Egelhofer et al., 2022; Freeze et al., 2021). Against this backdrop, this article relies on a two-wave experiment to investigate the effects of two different delegitimizing labels attacking the credibility of authentic information—a misinformation and a deepfake accusation—on the credibility of an authentic political video and related policy preferences.

Investigating the effects of mis- and disinformation as a delegitimizing tactic is especially relevant to consider in the context of increasing concerns about deepfakes. Even though empirical evidence on the occurrence and effects of deepfakes is scarce (but see e.g., Dobber, Metoui, Trilling, Helberger, & de Vreese, 2020), the widespread debate and concerns about their impact may have an independent effect, making (false) deepfake accusations more credible and impactful. Hence, false accusations of untruthfulness (i.e., a deepfake label) may lower the credibility of authentic information (van der Meer et al., 2023). In this context, political actors may strategically label conflicting information as "fake news" or "deepfakes" to delegitimize it (Levy & Ross, 2021). In this study, we investigate the possibility of accusations of mis- or disinformation being used as a political tactic, the extent to which these labels may affect source credibility and policy preferences, and the durability of such effects. By investigating the temporal impact of false disinformation accusations on credibility and policy preferences, this article contributes to our understanding of the consequences of strategically used disinformation labels for the delegitimization of authentic news and political information.

Mis- and Disinformation Accusations as Delegitimizing Labels

Different actors can accuse the mainstream media, experts, or political opponents of spreading false information. To date, most empirical research has focused on the "fake news" label as a delegitimizing communication tactic by which the media or politicians are accused of misleading people by strategically hiding reality from them (e.g., Egelhofer & Lecheler, 2019; Egelhofer et al., 2022; Farhall, Carson, Wright, Gibbons, & Lukamto, 2019; Van Duyn & Collier, 2019). Such accusations mostly resonate with the ideas of mis- and disinformation. While misinformation is understood as any information that is false or inaccurate without being intentionally misleading, disinformation is the deliberate dissemination of manipulated or misleading information (Freelon & Wells, 2020; Jack, 2017; Tandoc, Lim, & Ling, 2018). Disinformation thus refers to the intentional dissemination of false or manipulated information (which may include a deliberate accusation of false information), whereas misinformation may either refer to false information in general or information that turned out to be incorrect without the intention to cause harm or gain profit (Wardle & Derakhshan, 2017).

Paradoxically, the delegitimizing accusations studied in this article are both a disinformation label *and* genre (Egelhofer & Lecheler, 2019). They function as a label as they accuse established, authentic information of being untrue or deliberately misleading (see also Freeze et al., 2021). At the same time, they function as a genre of false information as they contain factually incorrect statements that are used strategically to demobilize support for the discredited actor's issue positions. Hence, there is no empirical evidence or expert knowledge that can be used to substantiate the facticity of the labels. These labels are in line with a communication tactic in which established media sources and political elites are held responsible for distorting the truth and lying to the people (Egelhofer & Lecheler, 2019).

When (political) actors voice accusations of disinformation, they not only emphasize that the message is inaccurate but also explicitly assign blame for deliberate, goal-directed manipulation of content. In line with this intentional dimension, empirical research on the attribution of false information has mainly focused on accusations of disinformation (i.e., the "fake news" label, see Egelhofer & Lecheler, 2019; Tamul, Holz Ivory, Hotter, & Wolf, 2020). We know little about the effects of falsely labeling information as *misinformation*—that is, stating that correct information contains mistakes or inaccuracies without attributing these inaccuracies to deliberate manipulation of content (but see e.g., Tandoc & Seet, 2022). Although labeling information as misinformation is a less severe accusation than disinformation attributions in the form of a deepfake accusation, referring to true information as inaccurate could still undermine news consumers' trust in information, potentially increasing skepticism and decreasing message acceptance. In line with this, different (experimental) studies show that accusations of disinformation can reduce trust in the targeted media channel (e.g., Egelhofer et al., 2022; van der Meer et al., 2023). As misinformation labels are a more subtle accusation that stays closer to the truth, it may be a more credible strategy of delegitimization (also see Hameleers, Brosius, & de Vreese, 2022; Tandoc & Seet, 2022).

Yet, we do want to emphasize that misinformation and disinformation accusations (i.e., labeling true information as a deepfake) can both be considered as forms of disinformation. Hence, even though the accusation of misinformation is not targeted at intent, it may strategically undermine the credibility of factually accurate information. As labeling true information as false through the form of an inauthentic fact-checker is a deliberate attempt to delegitimize information, we consider it as disinformation.

The Effects of Mis- and Disinformation Accusations

In this experimental study, we use a scenario in which not the *original* video message but rather the message that *refutes* the video and classifies it as false is disinformation. Such "false flags" delegitimize and discredit real information, can create distrust in the source, and can decrease support for policies related to the original message (e.g., Egelhofer et al., 2022; Freeze et al., 2021; Tandoc & Seet, 2022). Yet, the question remains whether false flags themselves can still be effective and credible if the original video is deemed authentic. To comprehensively study the impact of mis- and disinformation accusations, we focus on three elements: The extent to which false flags affect the (1) perceived credibility of real audiovisual information, (2) support for policies discussed in the video message, and (3) the perceived credibility of the false flag itself.

To better understand the effects of mis- and disinformation accusations in response to an authentic political video message, we first consider the persuasiveness of audiovisual information. Literature on multimodal framing has argued that information that relies on a combination of textual and visual cues may be more persuasive than purely textual information (Geise & Baden, 2015; Grabe & Bucy, 2009; Powell, Boomgaarden, De Swert, & de Vreese, 2015). While text may more directly transmit meanings, visuals evoke stronger emotions, are more attention-grabbing (Garcia & Stark, 1991), and bear a closer resemblance to external reality than written texts (Messaris & Abraham, 2001). This quality of "indexicality" is especially relevant to consider in light of deepfakes (Vaccari & Chadwick, 2020): Visual disinformation might be perceived as more authentic and credible because it is seen as a direct index of reality and less likely (and more difficult) to be manipulated, whereas text can be created, altered, and manipulated by anyone that can process it. This quality could make deepfakes more believable and harmful (Diakopoulos & Johnson, 2021) than textual misinformation, which in turn has several implications for deepfake accusations.

We present the false flag in the form of a (fake) fact-check. Considering that fact-checks are typically regarded as trustworthy, malign actors may exploit their perceived credibility by refuting authentic information or political speeches in the form of a fact-check. To understand the impact of false fact-checks, we build further on the insights of (experimental) research on the effectiveness of fact-checking as a journalistic routine. While fact-checks are not always effective, in particular, because people may not accept counter-attitudinal refutations (Nyhan & Reifler, 2010; Thorson, 2016), empirical evidence indicates that they can lower the perceived accuracy of false claims (Nyhan, Porter, Reifler, & Wood, 2019) even when the fact-check is not real (Hameleers & van der Meer, 2020). In addition, exposure to delegitimizing labels can lower trust in authentic news (Van Duyn & Collier, 2019). Similar findings come from Freeze and colleagues (2021), who documented lowered credibility perceptions of original news articles when respondents were exposed to misinformation warnings about the said articles—irrespective of whether these warnings were true or false. Freeze and colleagues (2021) also demonstrate that false flags can negatively affect individuals' memory of accurate information by "contaminating" it, concluding that "misdirected and imprecise warnings may counter the positive influence of misinformation warnings on memory" (p. 1456; see also Carey, Chi, Flynn, Nyhan, & Zeitzoff, 2020). Combining insights on the effects of a deepfake, mis- and disinformation accusations, and corrective information, we thus hypothesize that an accusation of mis- or disinformation, similar to the effects of an actual deepfake (Vaccari & Chadwick, 2020) and fact-check (Hameleers & van der Meer, 2020), could create uncertainty and damage trust and specifically affects the perceived credibility of the real video.

H1a: Participants exposed to accusations of mis- and disinformation perceive the real video as less credible than participants who are not exposed to mis- or disinformation accusations.

Given the "malicious" nature of deepfakes, we expect that accusing a video of being a deepfake would have a stronger negative effect on the video's credibility than accusing it of containing misinformation, which could be seen as an "honest" mistake. Disinformation, in this case made possible by the creation of a deepfake, highlights that the sender of false information *deliberately* aimed to mislead the public by using artificial intelligence (AI) to fabricate a false video fragment (e.g., Hancock & Bailenson, 2021; Westerlund, 2019). This more severe accusation emphasizes that the creator of the video has deliberately altered

audiovisual material to mislead the public, whereas this accusation of manipulative intent is absent in the misinformation accusation.

Hence, the misinformation accusation merely states that some information depicted in the video is inaccurate or false, whereas the deepfake accusation labels the video as manipulated, doctored, or even completely fabricated content—in line with the central features of an actual deepfake (e.g., Dobber et al., 2020; Westerlund, 2019). Deepfake accusations, in contrast to misinformation labels, clearly emphasize that the communicator has intentionally deceived the recipients and aimed to manipulate their views on reality (e.g., Hancock & Bailenson, 2021). The centrality of intentional deception in the deepfake accusation and the absence thereof in the misinformation accusation (e.g., Freelon & Wells, 2020) should result in more severe effects on the credibility of the deepfake accusation. This is corroborated by recent empirical research. More specifically, using an online survey study in Singapore, Tandoc and Seet (2022) find that people are more likely to respond with perceived falsity and intentional deception when exposed to the label "fake news" compared with misinformation and disinformation accusations. This underlines that signaling a lack of facticity and malicious intentions through the popular fake news term may yield stronger delegitimizing effects than more neutral terms. We thus hypothesize the following:

H1b: Participants exposed to accusations of disinformation (deepfake) perceive the real video as less credible than participants exposed to accusations of misinformation.

Although a deepfake accusation is more severe than a misinformation label, it may also be seen as less credible. In our experiment, we accuse an authentic video of being a deepfake whereas such forms of audiovisual disinformation are still relatively rare and cost- or labor-intensive to create (Dobber et al., 2020). Next to deepfakes not being very prominent in general (also see Brennen et al., 2021, fake news perceptions are only salient among a small group of populist citizens, whereas misinformation beliefs are widespread in society, e.g., Hameleers et al., 2022). Additionally, since it is much more labor-intensive to create a fake video than a fake (textual) fact-check, it is possible that respondents may rather question the credibility of the manipulated fact-check than the video itself. News consumers may have critical media literacy skills at their disposal that help them to resist disinformation—and thus also the false accusation of a deepfake video (Vraga & Tully, 2019). However, it may be more difficult to detect more subtle and non-intentional deviations from facticity. While an accusation of misinformation only challenges the veracity of the information made in the video, the deepfake accusation challenges the authenticity and intent of the news message itself. We therefore expect that participants perceive a deepfake accusation as less credible than an accusation of misinformation.

H2: False attributions of misinformation are seen as more authentic and accurate than false accusations of disinformation (i.e., a deepfake accusation).

Effects of Mis- and Disinformation Accusations on Policy Preferences

The effects of falsely labeling authentic information as mis- or disinformation could go beyond reducing credibility perceptions. An important political goal of disinformation may be to steer public opinion and influence the policy preferences of citizens (Bennett & Livingston, 2018). Similar to the rationale

underlying H1a and H1b, we expect an effect of false refutations on policy preferences related to the issue positions forwarded in the video. Fact-checking literature has shown that, beyond lowering the perceived accuracy of misinformation, corrective information can result in less issue agreement with the statements of the message flagged as false (e.g., Nyhan et al., 2019), and exposure to fact-checks can depolarize opinions about debated issues (Hameleers & van der Meer, 2020). Hence, policy positions may receive less support once their factual basis is discredited by mis- and disinformation. Yet, other research has found that political evaluations—at least in the partisan setting of the United States—are harder to correct (Nyhan et al., 2019).

Applied to the less polarized setting of European politics in general and the Dutch multiparty system more specifically, we expect that policy preferences could be changed by mis- and disinformation labels. Especially as we are looking at a low-salient policy position (i.e., the European Union's [EU] Green Deal) in the context of our study, we believe that the accusation forwarded by the mis- or disinformation label could influence people's positions on the issue at hand. We therefore expect that exposure to a deceptive and false fact-check that labels the authentic video message on the Green Deal as misinformation or a deepfake will lower respondents' support for the policies proposed in the Green Deal.

H3a: Participants exposed to accusations of mis- and disinformation are less likely to support related policies than participants who are not exposed to such a rebuttal.

Just like a disinformation label may have stronger effects on message credibility than a misinformation label, we expect that a deepfake accusation has a stronger impact on policy preferences than a misinformation attribution. Specifically, the disinformation accusation emphasizes that the video has deliberately been doctored with the intent to deceive citizens about the policies discussed in the video, which may create a more substantial level of cynicism and distrust in the policies that are discussed. Hence, although pointing out factual inaccuracies in the message may make citizens more skeptical and critical, the deepfake label may result in the systematic rejection of the policies as they are deemed illegitimate and manipulative.

H3b: The effects of false fact-checks on policy support are stronger for the deepfake attribution than the misinformation attribution.

These effects are likely to be moderated by the perceived relevance of the issue that is labeled as mis- or disinformation—in the case of this study, climate change. Research on the effectiveness of fact-checks has indicated that people are most likely to adjust their beliefs and issue agreement in line with the fact-check when their prior attitudes and ideological orientations do not resonate strongly with the false information (Nyhan & Reifler, 2010; Thorson, 2016). This can be explained as the result of motivated reasoning (Kunda, 1990; Taber & Lodge, 2006). To maintain an internally consistent and positive self-esteem, individuals are more likely to (uncritically) accept information that reassures their prior beliefs and criticize or reject information that challenges their beliefs. When people do not perceive climate change as an important issue, the mis- or disinformation accusations *against* the Green Deal video message resonate with their prior beliefs, which should make the mis- and disinformation label more effective. In line with this, we expect the following:

H4: The effects of misinformation and deepfake accusations on (a) perceived credibility and (b) policy support are stronger when participants believe that climate change is not an urgent issue.

The Duration of Effects of Mis- and Disinformation Accusations

Finally, we note that extant research on fact-checking has mainly studied the short-term effects of rebuttals (Nyhan et al., 2019; Thorson, 2016). In real-life information settings, original information and rebuttals that flag information as false do not always follow one another immediately. To simulate this realistic information setting, and in line with other research on the duration of media effects (e.g., Iyengar & Kinder, 1987; Lecheler & de Vreese, 2011; Mutz & Reeves, 2005), we incorporated time as a central component in our experimental design: Some of the participants saw the mis- or disinformation accusation right after the authentic video, whereas others saw it after three days. We measured the perceived credibility of the video and policy support in both waves. This design allowed us to assess (1) whether mis- and disinformation attributions were still effective when the response to the authentic video was delayed by a few days and (2) if the effects of refutations lasted when people saw a refutation only in the first wave. Based on the findings of Lecheler and de Vreese (2011), we expected that the effect of rebutted information would be also present in the second wave—although in a weaker form. Applied to corrective information more specifically, Brashier, Pennycook, Berinsky, and Rand (2021) found that exposure to fact-checks after seeing headlines enhances the likelihood that people can discern truth from false information, even after a week. In our experiment, we compared conditions in which the false flag directly followed the news item with a condition in which it was only presented to people in the second wave (three days later). We expected mis- and disinformation accusations to have the strongest effects when they directly followed the authentic video message. Yet, at the same time, the continued influence effect presupposed that fact-checks presented (directly after) exposure had a lasting effect on truth discernment (Brashier et al., 2021). We thus hypothesized the following:

H5: The effects of mis- and disinformation attributions on the (a) perceived accuracy and authenticity of the real news item and (b) policy support are weaker but still present when delaying the rebuttal.

H6: The effects of a mis- or disinformation accusation that directly follows the authentic video in the first wave is expected to have a persisting effect on the (a) perceived accuracy and authenticity of the real news item and (b) policy support in the second wave.

Context

We conducted the experiment in the Netherlands, a country with a multiparty political system, relatively high levels of media trust, and low levels of polarization. This setting was chosen as it offered a relatively resilient context for disinformation, considering that most people in the country trust established information sources (Humprecht, Esser, & Van Aelst, 2020). In addition, it was expected that the multiparty setting in the Netherlands would make the polarizing attacks of disinformation campaigns less effective across the board. However, in the Dutch political landscape, right-wing populist accusations of disinformation and the explicit use of the fake news label are prominent, and the discourse around disinformation could appeal to voters on the fringes of the political spectrum. Therefore, we

expected that the mis- and disinformation accusations studied in this article would resonate well with the actual discourses around false information in the Dutch setting.

In this study, we used a video of the European Green Deal published on YouTube. In the video, Frans Timmermans voiced general statements on the importance of acting together as the EU in the fight against climate change. The video was watched only 3,000 times and received just 59 likes. Although Dutch participants may have been familiar with the ideas of the European Green Deal, and the overall agenda of reducing emissions, Timmermans and the specific statements made on behalf of the European Commission were not central in the media and public discourse in the Netherlands. This was confirmed by the low levels of engagement in the video, and the fact that participants in the study did not recognize this video.

Considering that our focus was on a low-salient political actor and issue position, we also believed that the delegitimizing attack had the potential to steer policy evaluations by casting doubt on the epistemic foundations of a real political speech. As we did not select an expert-driven or authoritative message for the experiment, it could be expected that deceptive fact-checks would offer an even stronger delegitimizing narrative when responding to official and evidence-based reporting. Finally, it should be noted that the mis- and disinformation accusations have targeted climate change policies, which are surrounded by polarized debates in the Netherlands and beyond.

Method

We conducted two pretests to test the credibility of the fake rebuttals. In the first pretest (N = 108), we found barely any differences between how respondents in the mis- and disinformation groups rated the credibility of the original video (mean 3.94 vs. 3.95 on a 7-point scale, nonsignificant differences based on *t*-tests) and fact-checks (3.73 vs. 3.69, also nonsignificant differences). As a consequence of that, we included a control group to be able to compare the effects of being exposed to a fact-check with a condition with no fact-check. We also worded the deepfake claim more strongly. In addition, we included several attention checks in the design. In the second pretest (N = 108), the group that saw the video and no fact-check rated the video most credible (M = 4.37). The group that saw a *misinformation* fact-check rated the video as significantly less credible (M = 3.84, $p < .001$), but the group that saw a disinformation fact-check rated the video as least credible (M = 3.78). The misinformation label itself (4.04) was rated as more credible than the disinformation label (M = 3.30, $p < .001$).

The hypotheses of this study were preregistered[1] (along with the hypotheses of a second experiment dealing with the Green Deal), and although their wording was slightly adjusted to match the text flow in this article, the hypothesized effects as such were not changed. Both the pretests as well as the main study were reviewed and approved by the University of Amsterdam's ethical review board. The respondents were debriefed at length at the end of the experiment to ensure that there would be no misunderstanding about which information was true or false.

[1] https://aspredicted.org/blind.php?x=mc9zb9.

Design

The real video that was flagged as mis- or disinformation was constant across conditions and concerned a message about the EU's Green Deal (European Commission, 2019). In two of the three experimental groups in Wave 1 (W1), this video clip was followed by a fabricated online news article that flagged the video as mis- or disinformation (false flag). There were two variations of this accusation: A false accusation of misinformation (there was an alleged honest mistake in the video message) and the attribution of disinformation (the video was labeled a deepfake). In the third group of W1, no fact-check was included. Next to the three levels of the rebuttal, the experiment included a time component. Some participants saw the rebuttal directly after the video clip (W1), and others were only exposed to it in the second wave after three days (W2). The study's design is summarized in Figure 1a. Figure 1b presents a flowchart of the experimental design across the waves.

Group number	*n*	Wave 1 Fake news label	Wave 2 Fake news label
1	275	Disinformation	-
2	247	Misinformation	-
3	97	-	Disinformation
4	89	-	Misinformation
5	80	-	-

Figure 1a. Study design.

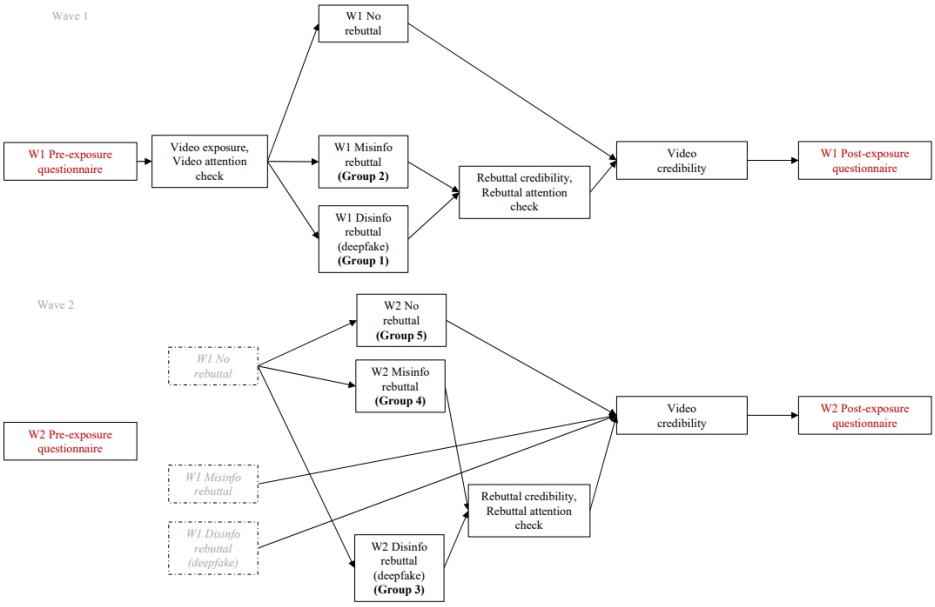

Figure 1b. Flowchart of the experimental design across both waves.

Sample

Respondents were recruited through a polling company (*PanelInzicht*) and were incentivized by the pollster after they took part in both waves of the online study. No hard quotas were enforced, but we aimed for a diverse sample that generally reflected the Dutch population. The sample consisted of 788 Dutch respondents,[2] of which 413 (52.4%) are women (compared with 50.3% in the general Dutch population in 2020). Our sample was slightly older than the population average (42.2 years) since we only included respondents above the age of 18 years. Their mean age was 47.9 years (*SD* = 12.9), ranging from 18 to 65; most respondents had a medium level of education and considered themselves to have a medium level of income.

Independent Variables

First, all participants were exposed to the authentic short video (European Commission, 2019). The video did not have source cues and was accompanied by Dutch subtitles. After this video was watched by all the participants, the sample was split into three groups: In the control group, participants did not see the second message; in the misinformation condition, participants saw a fabricated fact-check accusing the authentic video of containing false information. In the disinformation condition, they

[2] In this study, we only considered participants who took part in both waves of the experiment. Incomplete answers (i.e., because of dropouts after Wave 1) were not included in further analyses.

were shown a fabricated fact-check accusing the message of being a deepfake (including an explanation of what a deepfake is). The stimuli are included in the online appendix files (see: https://surfdrive.surf.nl/files/index.php/s/k8ONLks7ydpChie). The misinformation rebuttal made claims in the form of "This video contains claims that are not accurate" and "We can't verify the empirical basis of these claims." The deepfake rebuttal implied that the video was a deliberate attempt to manipulate the public: "The video is a deepfake. A fabricated video in which the makers used AI to make Timmermans say things he never said," and "The climate lobby manipulated this video to hide the truth about climate change."

As an attention check, respondents were asked what period of time Frans Timmermans had mentioned in the video in reference to climate change (80.58% gave correct answers). As an attention check for the fact-check, respondents were asked to identify a statement that described the content of the fact-check (W1, 78.74% correct; W2, 85.54% correct). In line with studies such as that by Aronow, Baron, and Pinson (2019), we did not exclude respondents who did not pass the attention check; however, we took the results as an indication that respondents generally paid attention to the stimuli.

Dependent Variables

Means and standard deviations for all dependent variables and the moderator, as well as scale reliability, can be found in Table 1. The credibility of the video message and the fact-check were measured as the average agreement with five statements, each on a scale ranging from 1 (completely disagree) to 7 (completely agree). Policy support for the EU Green Deal was measured using an agreement scale ranging from 1 (completely disagree) to 7 (completely agree), pertaining to five statements, of which three were favorable and two were unfavorable to the policy (reverse coded). Climate change urgency attitudes were measured with six statements (two reverse coded) using an agreement scale (from 1 = completely disagree to 7 = completely agree), with higher scores implying agreement with the notion that climate change is an urgent issue. The translated items for all scales are included in Appendix B of the online appendix file.

Table 1. Descriptive Statistics for Dependent Variables.

Statistic	*N*	Mean	*SD*.	Min	Max	Cronbach's alpha
Credibility fact-check Wave 1	522	3.65	1.33	1	7	.89
Credibility fact-check Wave 2	186	3.57	1.21	1	7	.87
Credibility video Wave 1	788	4.21	1.42	1	7	.92
Credibility video Wave 2	788	3.92	1.28	1	7	.91
Green Deal support Wave 1	788	4.74	1.35	1	7	.89
Green Deal support Wave 2	788	4.60	1.29	1	7	.87
Climate change urgency attitudes	788	4.81	1.50	1	7	.92

Results

We reported all results for the five experimental groups separately. Even though there was no treatment difference between Groups 3, 4, and 5 in Wave 1, this approach enabled us to compare over-time developments between Wave 1 and 2 for these groups in a more conservative way.

The Effects of False Flags on Credibility Perceptions of the Original Message

The mean values for the credibility of the authentic video and fake fact-check as well as support for the EU Green Deal across all five conditions are displayed in Figure 2 for the two waves. We first tested H1a, which stated that *participants who were exposed to the mis- or disinformation accusation would find the video less credible than participants who did not see such an accusation*. An analysis of variance showed that there were significant differences among the five experimental groups (F = 10.23; df = 4; p < .01). Bonferroni post hoc tests (see mean scores in Figure 2) showed that the perceived credibility of the authentic video was significantly lower in Group 1 (the deepfake condition) than in Groups 4 (p < .01) and 5 (p < .01; control groups). The same applies to the misinformation accusations: Participants in Group 2 (misinformation condition) rated the video as significantly less credible than participants in Groups 3 (p = .02), 4 (p < .01), and 5 (p < .01; control groups).[3] That means that respondents who saw any version of the mis- or disinformation accusation rated the video as significantly less credible than respondents who did not see either, which confirms H1a.

Hypothesis 1b stated that *participants exposed to accusations of disinformation (deepfake) would perceive the real video as less credible than participants exposed to accusations of misinformation*. According to Bonferroni post hoc tests, the differences in video credibility between Group 1 and 2 (W1) were not significant (p = 1.00), meaning that the negative effect on the video's credibility did not depend on whether the fact-check made an accusation of mis- or disinformation. This was also replicated for Groups 3 and 4 (p = 1.00), who were exposed to the false flag only in Wave 2. We therefore reject H1b.

We assumed, in H2, that *false attributions of misinformation would be seen as more authentic and accurate than false accusations of disinformation*. Mean differences across conditions were significant for the credibility of the false flag in Wave 1 (F = 9.79, df = 1, p = .002) and Wave 2 (F = 4.50; df = 1; p = .04). In both cases, the disinformation (i.e., deepfake) version of the fact-check was considered less credible than the misinformation version. This confirms H2. Our findings thus indicate that the more severe attack on the authenticity and intentions of the real video is less credible than labeling true information as erroneous without the intention to mislead the audience.

The Effects of False Flags on Policy Support

Hypothesis 3a stated that *participants exposed to accusations of mis- and disinformation were less likely to support related policies than participants who were not exposed to such a rebuttal*, and we further

[3] It is noteworthy that there was also a significant difference between Groups 3 and 5 even though they did not see different versions of the stimulus material (i.e., both functioned as control groups for Wave 1).

assumed, in H3b, that this *negative effect on policy support would be even stronger for the deepfake than the misinformation attribution.* We found no significant differences in support for the EU Green Deal in either Wave 1 (F = 1.48, df = 4; p = .21) or Wave 2 (F = 1.55, df = 4; p = .19) between respondents who were exposed to a mis- or disinformation attribution versus respondents in the control group. Therefore, we reject H3a and H3b.

Moderation Effects

Hypothesis 4 stated that *the effects of misinformation and deepfake accusations on (a) perceived credibility and (b) policy support were stronger when participants believed that climate change was not an urgent issue.* Regarding H4a, there was no significant interaction effect of urgency attitudes and the experimental manipulation on video credibility in Wave 2 (F = 1.65, df = 4; p = .16), but there was a statistically significant effect in Wave 1 (F = 3.72, df = 4; p = .005). However, the effect was the opposite of what we expected: Group differences following the fact-check were larger for those who thought climate change was an important issue. The more people's prior beliefs aligned with the authentic claims made in the real video, the stronger the impact of fake accusations on (lowering) the credibility of the authentic information. This interaction is visualized in Appendix C. There is thus no support for H4a.

While there was a strong correlation between urgency attitudes and support for the EU Green Deal (r = .80 in Wave 1), urgency attitudes did not moderate the (non)-effects of the experimental manipulation on Green Deal support in Waves 1 (F = .58, df = 4; p = .68) or 2 (F = .17, df = 4; p = .98), providing no support for H4b.

Time Delay

We further assumed, in H5, that the *effects of mis- and disinformation attributions on the (a) perceived accuracy and authenticity of the real news item and (b) policy support were weaker but still present when delaying the rebuttal*. Comparisons of average credibility perceptions of the respective dis- (Groups 1 and 3) and misinformation conditions (Groups 2 and 4; see Figure 2) in W1 and W2 show that respondents in the second wave rated the original message as even less credible than participants who had seen the rebuttal immediately after being exposed to the video in W1. We thus reject H5a. While the differences in policy support between W1 and W2 rebuttal conditions are less pronounced, participants who only saw the rebuttal after three days still expressed less support for the Green Deal. Hypothesis 5b is also rejected.

Finally, H6 stated that *the effects of a mis- or disinformation accusation that directly followed the authentic video in the first wave would have a persisting effect on the (a) perceived accuracy and authenticity of the real news item and (b) policy support in the second wave.* There were significant differences in the mean credibility of the video in Wave 2 (F = 5.45; df = 4; p < .01). In Wave 2, video credibility for Group 1 (W1 deepfake) and Group 2 (W1 misinformation) was quite consistent and did not change much in comparison with Wave 1. Average video credibility also somewhat decreased for the W2 control group (Group 5). However, Groups 1 (W1 deepfake, p < .01) and 2 (W1 misinformation, p < .01) were still significantly lower in their perceived video credibility than the control group. This offers

support for H6a and shows that mis- and disinformation accusations have a persistent effect in lowering the credibility of authentic videos, which lasts for three days. Support for the Green Deal did not differ significantly in Wave 2 (F = 1.55; df = 4; p = .19) and was slightly lower for all groups compared with Wave 1. These results indicate that mis- and disinformation attributions persistently affect policy support over time.

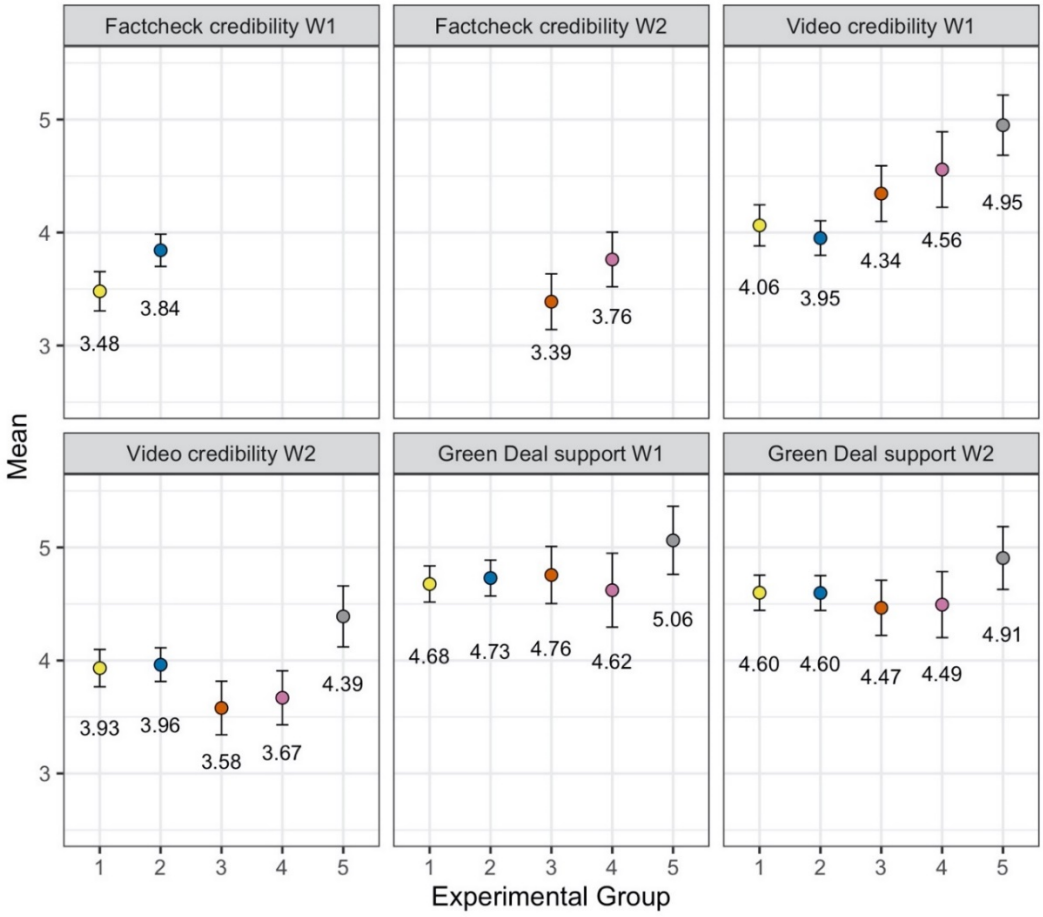

Figure 2. Group means across conditions and waves.

Discussion

The present study investigated the effects of false accusations of misinformation or disinformation (deepfake labels) with regard to an authentic video about the EU's Green Deal—and the duration of the

effects on credibility and policy support. Our main findings indicate that false accusations of both misinformation and disinformation (in this study, an alleged deepfake) leveled by some agents can lower the credibility of an authentic political video and offer empirical evidence for the effectiveness of the mis- and disinformation label, which is used strategically to attack political opponents, media outlets, and political positions that are incongruent with their political agenda (e.g., Egelhofer & Lecheler, 2019; Farhall et al., 2019; Tandoc & Seet, 2022). In line with empirical research that indicates that authentic fact-checks can reduce the credibility of mis- and disinformation (Nyhan et al., 2019; Porter & Wood, 2022), our findings show how using the legitimacy of these formats can also discredit authentic information (Egelhofer et al., 2022; Freeze et al., 2021). Agents of disinformation can thus make use of the authenticity of fact-checking formats when aiming to delegitimize established truths that are incongruent with their issue positions.

Labeling the authentic video as a deepfake or as containing misinformation had similar negative effects on how credible users thought the video was. This was against our expectation that the deepfake accusation would have a stronger effect than a less-severe misinformation accusation. The finding that deepfake accusations are less credible than misinformation accusations while having similar effects on lowering the credibility of authentic information could be explained by regarding the accusations as a trigger event that signaled suspicion about the content (also see Van der Meer et al., 2023). Irrespective of the severity of the accusation, mis- and disinformation labels may motivate people to deviate from the truth bias (Levine, 2014), which lets them critically reconsider whether information can be deemed authentic. The deepfake accusation may, however, be subject to more doubt as it does not only attack the truthfulness of the statements of the political speech but also points out that the video is synthetic and fabricated from scratch. It can also be argued that the misinformation accusation blames the speaker of the message (Timmermans), whereas the deepfake accusation attributes blame to the climate lobby. Both messages may motivate people to reconsider the credibility of the original message, whereas the more extreme nature of the accusation in the deepfake condition may also trigger doubt related to the label.

These findings show that it is relatively easy to discredit authentic audiovisual information with relatively mild accusations of misinformation and disinformation even when these are *not* deemed credible. In our study, and contrary to prior research (Nyhan et al., 2019), this effect was not conditional on the resonance of the fake accusations with people's prior attitudes toward climate change. If anything, the delegitimizing effects of false flags were strongest among people who were *most* likely to agree with the claims of authentic information. This may be explained by higher levels of issue importance and accuracy motivations among people supporting the authentic video: They may be more invested in the issue and more open to corrective information that points them to inconsistencies (also see Hameleers & van der Meer, 2020). People who tend to distrust the message in the first place may be less open to the novel information presented in the correction: They already hold the belief that the message is inaccurate and do not need the fact-check to further persuade them.

We found that the negative impact on the authentic video's credibility persisted over time. Although previous research on the impact of disinformation and fact-checking has mainly looked at the direct impact of disinformation or rebuttals (e.g., Thorson, 2016), corrective information typically responds to false information after some time elapsed. Likewise, it may be argued that disinformation only poses a real threat to democracy when its delegitimizing impact lasts longer than within the time frame of a lab experiment.

We show that fake accusations of mis- and disinformation continue to lower the perceived credibility of authentic videos after three days have passed. These findings illustrate the real-life implications of disinformation targeted at legacy journalism and authentic news: Delegitimizing labels can have a lasting influence on people's reality perceptions.

However, our findings also give reason for optimism. First, the lower credibility ratings for the deepfake flag compared with ratings for the misinformation accusation show that respondents generally have critical news media literacy skills they use to judge the quality of news media information. However, this also indicates that audiovisual information is generally perceived as credible: Accusations of a deepfake manipulation are disregarded possibly because individuals deem it unlikely that videos can be (convincingly) manipulated. In other words, people think that videos may contain false information (misinformation) but not that they are entirely fabricated (disinformation). Second, the false flag had no reducing effect beyond the video's credibility: While mis- and disinformation accusations can negatively affect the credibility of the video itself, they do not influence policy support. This could be seen as good news from a democratic point of view as it points to the limitations of applying delegitimizing labels of false information to authentic information as a strategy to attack political opponents and demobilize support for their policies. In sum, although disinformation agents may succeed in generating confusion about factual information (see also Vaccari & Chadwick, 2020), our findings indicate that there could be limits to the political impact of this strategy. Of course, this conclusion must be viewed within the specific constraints of this study.

Despite offering insights into the impact of mis- and disinformation accusations targeted at authentic content, this study has limitations that may be remedied in future research. First, the lack of findings for policy preferences may be due to the fact that attitudes toward climate change policies are relatively stable and established—particularly in the Netherlands. The topic has been the subject of public debate for several decades, and citizens are likely to have relatively set opinions on their support for climate policies. More volatile attitudes and policy preferences may be more malleable and could be more vulnerable to false misinformation or deepfake accusations. In addition, EU politics is generally more removed from citizens than local and national politics, which sets the bar even higher for changing policy preferences in the EU context. We therefore suggest that future research look at mis- and disinformation accusations in response to a more diverse set of more- and less-polarizing and national issues.

Although this study is the first to look at the longer-term impact of disinformation in the form of false flags, we only measured our dependent variables twice over a period of three days. Future research may track the impact of disinformation over an extended time frame. Furthermore, our stimulus was a single, short video clip, which is relatively easy to manipulate or take out of context. News shows (e.g., on TV) are seen in a different context, contain more content, and are accompanied by more source cues, which would likely make them more difficult to delegitimize. As with most experiments, the experimental setup can lead to issues of external validity.

We also consider the presentation of the mis- and disinformation accusations in the same format as legitimate fact-checks as a potential limitation of this article. In reality, accusations of mis- and disinformation or explicit "fake news" labels are often expressed in the direct communication of (populist) politicians or communicated by alternative media outlets that attack the mainstream (e.g., Egelhofer &

Lecheler, 2019). Although these accusations can come in the form of misleading fact-checks, future research may experiment with the context of accusations of mis- and disinformation to enhance the external validity of mapping the effects of delegitimizing attacks on information's credibility and authenticity. In line with this, future research may explore the effects of deceptive fact-checks related to different polarizing issues, such as the war in Ukraine, immigration, or health communication. Importantly, the deceptive use of false fact-checkers to delegitimize established information may be regarded as an important tool to harm opponents or issue positions that are incongruent with the views of the attacking party.

These limitations notwithstanding, this article offers insights into the direct and longer-term impact of different types of mis- and disinformation accusations, which have become ubiquitous in the current digital information age. We show that these accusations may have the intended delegitimizing impact on the credibility of authentic information but, on a more optimistic note, also indicate that attacks on legitimate information cannot demobilize citizens' support for important global and highly politicized issues such as climate change. These findings have implications for media policy and practice and highlight the importance of stimulating critical news media literacy skills among news consumers, which allow them to more reliably identify mis- and disinformation. Journalists, news organizations, and platforms should continue correcting false information while also sensitizing news consumers to the potential of (mis)using delegitimizing mis- and disinformation accusations as a political strategy, given that such accusations have become a key feature of today's digital journalism and news environment.

References

Acerbi, A., Altay, S., & Mercier, H. (2022). Research note: Fighting misinformation or fighting for information? *Harvard Kennedy School Misinformation Review, 3*(1), 1–15. doi:10.37016/mr-2020-87

Aronow, P. M., Baron, J., & Pinson, L. (2019). A note on dropping experimental subjects who fail a manipulation check. *Political Analysis, 27*(4), 572–589. doi:10.1017/pan.2019.5

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication, 33*(2), 122–139. doi:10.1177/0267323118760317

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences of the United States of America, 118*(5), 1–3. doi:10.1073/pnas.2020043118

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

Carey, J. M., Chi, V., Flynn, D. J., Nyhan, B., & Zeitzoff, T. (2020). The effects of corrective information about disease epidemics and outbreaks: Evidence from Zika and yellow fever in Brazil. *Science Advances, 6*(5), 1–11. doi:10.1126/sciadv.aaw7449

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

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

Egelhofer, J. L., Boyer, M., Lecheler, S., & Aaldering, L. (2022). Populist attitudes and politicians' disinformation accusations: Effects on perceptions of media and politicians. *Journal of Communication, 72*(6), 619–632. doi:10.1093/joc/jqac031

Egelhofer, J. L., & 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. doi:10.1080/23808985.2019.1602782

European Commission. (2019, December 1). *#VdLcommission: Presentation message by Frans Timmermans* [Video file]. YouTube. Retrieved from https://www.youtube.com/watch?v=GSgKu6yds3k

Farhall, K., Carson, A., Wright, S., Gibbons, A., & Lukamto, W. (2019). Political elites' use of fake news discourse across communications platforms. *International Journal of Communication, 13*, 4353–4375.

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

Freeze, M., Baumgartner, M., Bruno, P., Gunderson, J. R., Olin, J., Ross, M. Q., & Szafran, J. (2021). Fake claims of fake news: Political misinformation, warnings, and the tainted truth effect. *Political Behavior, 43*, 1433–1465. doi:10.1007/s11109-020-09597-3

Garcia, M., & Stark, P. (1991). *Eyes on the news*. St. Petersburg, FL: Poynter Institute for Media Studies.

Geise, S., & Baden, C. (2015). Putting the image back into the frame: Modelling the linkage between visual communication and frame-processing theory. *Communication Theory, 25*(1), 46–69. doi:10.1111/comt.12048

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

Hameleers, M., Brosius, A., & de Vreese, C. H. (2022). Whom to trust? Media exposure patterns of citizens with perceptions of mis- and disinformation related to the news media. *European Journal of Communication, 37*(3), 237–268. doi:10.1177/02673231211072667

Hameleers, M., & van der Meer, T. G. L. A. (2020). Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research, 47*(2), 227–250. doi:10.1177/0093650218819671

Hancock, J. T., & Bailenson, J. N. (2021). The social impact of deepfakes. *Cyberpsychology, Behavior and Social Networking, 23*(4), 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

Iyengar, S., & Kinder, D. R. (1987). *News that matters: Television and American opinion*. Chicago, IL: University of Chicago Press.

Jack, C. (2017). *Lexicon of lies: Terms for problematic information*. Data & Society. Retrieved from https://datasociety.net/library/lexicon-of-lies/

Jungherr, A., & Schroeder, R. (2021). Disinformation and the structural transformations of the public arena: Addressing the actual challenges to democracy. *Social Media+ Society, 7*(1), 1–13. doi:10.1177/2056305121988928

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

Lecheler, S., & de Vreese, C. H. (2011). Getting real: The duration of framing effects. *Journal of Communication, 61*(5), 959–983. doi:10.1111/j.1460-2466.2011.01580.x

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

Levy, N., & Ross, R. M. (2021). The cognitive science of fake news. In M. Hannon & J. de Ridder (Eds.), *The Routledge handbook of political epistemology* (pp. 181–191). London, UK: Routledge.

Marty, M. (2022, March 22). *Did Putin use a green screen to fake an Aeroflot meeting?* Goosed. Retrieved from https://goosed.ie/news/putin-fake-green-screen-video/

Messaris, P., & Abraham, L. (2001). The role of images in framing news stories. In S. D. Reese, O. H. Gandy, & A. E. Grant (Eds.), *Framing public life* (pp. 215–226). Mahwah, NJ: Erlbaum.

Mutz, D. C., & Reeves, B. (2005). The new videomalaise: Effects of televised incivility on political trust. *American Political Science Review, 99*(1), 1–15. doi:10.1017/S0003055405051452

Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2019). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior, 41*(1), 939–960. doi:10.1007/s11109-019-09528-x

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior, 32*(2), 303–330. doi:10.1007/s11109-010-9112-2

Porter, E., & Wood, T. J. (2022). Political misinformation and factual corrections on the Facebook news feed: Experimental evidence. *The Journal of Politics*, *84*(3), 1812–1817. doi:10.1086/719271

Powell, T. E., Boomgaarden, H. G., De Swert, K., & de Vreese, C. H. (2015). A clearer picture: The contribution of visuals and text to framing effects. *Journal of Communication, 65*(6), 997–1017. doi:10.1111/jcom.12184

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science, 50*(3), 755–769. doi:10.1111/j.1540-5907.2006.00214.x

Tamul, D. J., Holz Ivory, A., Hotter, J., & Wolf, J. (2020). All the president's tweets: Effects of exposure to Trump's "fake news" accusations on perceptions of journalists, news stories, and issue evaluation. *Mass Communication and Society, 23*(3), 301–330. doi:10.1080/15205436.2019.1652760

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news": A typology of scholarly definitions. *Digital Journalism, 6*(2), 137–153. doi:10.1080/21670811.2017.1360143

Tandoc, E. C., & Seet, S. K. (2022). War of the words: How individuals respond to "fake news," "misinformation," "disinformation," and "online falsehoods." *Journalism Practice,* 1–17. doi:10.1080/17512786.2022.2110929

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 + Society, 6*(1), 1–13. doi:10.1177/2056305120903408

Van der Meer, T. G., Hameleers, M., & Ohme, J. (2023). Can fighting misinformation have a negative spillover effect? How warnings for the threat of misinformation can decrease general news credibility. *Journalism Studies,* 1–21. Advance online publication. doi:10.1080/1461670X.2023.2187652

Van Duyn, E., & 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. doi:10.1080/15205436.2018.1511807

Vraga, E. K., & Tully, M. (2019). News literacy, social media behaviors, and skepticism toward information on social media. *Information, Communication & Society, 24*(2), 1–17. doi:10.1080/1369118X.2019.1637445

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

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking (Council of Europe Report)*. Retrieved from https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

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