Media and Communication (ISSN: 2183–2439)
2023, Volume 11, Issue 2, Pages 5–14
https://doi.org/10.17645/mac.v11i2.6301

Article

The (Un)Intended Consequences of Emphasizing the Threats of Mis- and Disinformation

Michael Hameleers

Amsterdam School of Communication Research, University of Amsterdam, The Netherlands; m.hameleers@uva.nl

Submitted: 21 October 2022 | Accepted: 6 February 2023 | Published: 28 April 2023

Abstract
The mis- and disinformation order does not only consist of the dissemination of deceptive content but also involves using fake news as a blame-shifting label in politics and society. The salience of this label on social media and in political discourse, and the frequent discussions held about the threats of fake news in public opinion, may result in a systematic overestimation of mis- and disinformation's presence. Even more so, these primed perceptions about false information may affect people's evaluations of factually accurate information. In this article, we offer a theoretical account of how the public's and media's attention to mis- and disinformation, fake news labels, and the threats of mis- and disinformation may have a negative impact on people's trust in factually accurate information and authentic news. In addition, relying on an experimental case study of pre-bunking interventions, we illustrate the extent to which tools intended to increase media literacy in the face of mis- and disinformation may also have ramifications for trust in reliable information. Based on this, we propose a forward-looking perspective and recommendations on how interventions can circumvent unintended consequences of flagging false information.

Keywords
credibility; disinformation; fake news; media literacy; misinformation; truth bias

Issue
This article is part of the issue "Fakespotting: (Dis)Information Literacy as Key Tool to Defend Democracy" edited by José Antonio Muñiz-Velázquez (Universidad Loyola Andalucía) and Claudio Paolucci (University of Bologna).

1. Introduction

The spread of mis- and disinformation has been regarded as a severe threat to democracies across the globe (e.g., Bennett & Livingston, 2018). We define disinformation as the covert and deliberate dissemination of deceptive information (e.g., Chadwick & Stanyer, 2022; Freelon & Wells, 2020). Misinformation refers to false information disseminated without the intention to deceive (e.g., Wardle, 2017). Although most empirical and theoretical accounts of mis- and disinformation's effects have focused on the consequences of mis- or disinformation as a genre of deceptive information, disinformation also exists as a discursive issue (e.g., Egelhofer & Lecheler, 2019). To offer an example, different politicians have weaponized the label fake news and exploited the opinion climate of factual relativism by accusing their opponents of disseminating false information. Such fake labels can have real consequences. As illustrated by Van Duyn and Collier (2019), using mis- or disinformation labels in elite discourse negatively influences people's trust in authentic news, and lowers the accuracy of discerning false from real statements. Against this backdrop, we have to shift our focus from mis- and disinformation as a purely informational crisis to a more holistic understanding of the threats associated with discussions surrounding the concept.

In line with the severe public concerns about mis- and disinformation, numerous interventions with the mission statement to combat and fight mis- and disinformation have been launched, such as fact-checks, media literacy interventions, and inoculation strategies (e.g., Roozenbeek & van der Linden, 2019; Tully et al., 2020). Although these interventions are effective (see Walter et al., 2020, for a meta-analysis on the effectiveness of fact-checks), they may also contribute to heightened, or even disproportionate, public concerns about the dissemination of mis- and disinformation. Hence, recent empirical evidence has demonstrated that mis- and disinformation are extremely rare phenomena in people's media diets: Less than 2% of people's information diet in the US is estimated to contain mis- or disinformation (e.g., Acerbi et al., 2022). This low number does not seem to be reflected in the global response to mis- and disinformation, or the perceived prevalence of deceptive information. Based on a large-scale comparative survey, more than half of all participants (54%) are concerned about false information and its detection (Newman et al., 2022). In line with these concerns, different platforms, initiatives, committees, and working groups are introduced to warn people about disinformation, or promote resilience to the threats associated with mis- and disinformation.

Considering that more than half of the global population is concerned about mis- or disinformation whereas it may only make up less than 2% of their media diet, this article focuses on the effects of talking about the threats of mis- and disinformation. More specifically, this article offers a theoretical account of how the public's and media's attention to mis- and disinformation, fake news labels, and warnings may have a negative impact on people's trust in factually accurate information and authentic news. To further illustrate the unintended consequences of mis- or disinformation discussions, we will rely on an empirical example to test the potential backfire effect of media literacy interventions that warn people about mis- and disinformation. As a main contribution, we highlight the need to approach mis- and disinformation as a context-bound disorder, that involves discourses of fake news and societal responses to the threat. In a post-truth information ecology, talking about mis- and disinformation—be it in a malicious or well-intended manner—may have negative consequences for society as it may contribute to higher levels of distrust, doubt and factual relativism.

2. Theory

2.1. Mis- and Disinformation as Informational, Perceptual, and Discursive Phenomena

Before mapping the different components of the mis- and disinformation order, we need a comprehensive definition of the concept. Here, we follow extant literature that has distinguished misinformation from disinformation (see e.g., Wardle & Derakhshan, 2017). Misinformation is either used generally as an umbrella term to refer to false information that is not based on relevant expert knowledge or evidence (Vraga & Bode, 2020), or specifically to indicate that false information is disseminated without the intention to deceive recipients (Wardle, 2017). Disinformation, however, refers to the covert and goal-directed manipulation, decontextualization, or fabrication of information with the intention to cause harm or deceive (see e.g., Chadwick & Stanyer, 2022; Freelon & Wells, 2020). Mis- and disinformation can be disseminated for various reasons, such as the cultivation of cynicism or distrust, the reinforcement of societal cleavages, financial profit, or the legitimization of identity-congruent ideologies (e.g., Hameleers, 2022; Marwick & Lewis, 2017). Yet, the intentional component of disinformation is often difficult to distinguish without a full understanding of the context of deception (Hameleers, 2022). In any case, mis- and disinformation both refer to information that is false and potentially deceptive. As false information can be misinformation in one context and disinformation in another, we refer to the terms mis- and disinformation interchangeably in this article. By referring to mis- and disinformation, we aim to capture the broader concept of false information and warnings about its presence—which may refer to both the intentional dimension reflecting disinformation or the general scope of false information (misinformation).

Mis- and disinformation may not only relate to the accuracy of information, but can also be used as blame-shifting labels. More specifically, mis- and disinformation can be used to delegitimize information, sources, or actors, for example, by referring to them as "fake news" (see e.g., Egelhofer & Lecheler, 2019). In line with this, mis- and disinformation can also be regarded as discursive phenomena in which information is labeled as false and/or deceptive, herewith lowering the credibility of opposed information sources or communicators. Egelhofer and Lecheler (2019) have argued that fake news as a label specifically can exist as a delegitimizing master frame used mostly by right-wing populist actors to blame the established press for spreading false information. This application of disinformation as a label implies malicious intent: Disinformation is discursively used to delegitimize (political) opponents and to negatively impact the public's evaluation of the targets addressed by the label (Van Duyn & Collier, 2019). This way, fake news may be used to create the impression that incongruent information is false—a strategy that mostly fits right-wing populists' anti-establishment narrative (e.g., Waisbord, 2018). Based on the findings of an experiment, it can be argued that such attacks can be effective: Exposure to fake news discourses used by political elites can lower people's trust in real information, and harm their resilience against disinformation (Van Duyn & Collier, 2019).

Discourses of mis- and disinformation are not exclusively based on de-legitimizing narratives and accusations disseminated by malign actors. Hence, in line with increased public concerns about mis- and disinformation (Newman et al., 2022), the news media, journalists, educators, and governmental organizations are frequently warning the public about the dangers of disinformation and fake news in society. These discussions, albeit well-intended, may raise fear and cause overall levels of doubt among news audiences that no longer know whom they can trust. In this article, we specifically argue that measures used to prevent or counter the impact of mis- and disinformation—such as media literacy interventions, inoculation strategies, and fact-checks—may unintentionally cause suspicion and lower the credibility and trustworthiness of authentic information.

Despite these potentially harmful side effects, interventions such as fact-checking, media literacy messages, and inoculation strategies are proven to be effective in lowering the acceptance and credibility of mis- and disinformation. Inoculation interventions, such as a misinformation game that exposes people to a small dose of false information to make them more resilient to actual misinformation, have been found to lower misperceptions (Roozenbeek & van der Linden, 2019). Such interventions may work as they offer users practical suggestions on how to resist and recognize misinformation, whilst the actual confrontation with false information helps them to understand its mechanisms. In a similar vein, media literacy messages that present news users with a list of suggestions on how to recognize and resist misinformation are found to be effective in lowering the credibility of misinformation in experimental research settings (e.g., Hameleers, 2022; Tully et al., 2020). There is even more research on the effectiveness of fact-checking messages (e.g., Chan et al., 2017; Walter et al., 2020). Fact-checks typically offer a short, fact-based verdict on the truthfulness of suspicious or highly prominent claims (e.g., Uscinski & Butler, 2013). Although these refutations of false information are mostly seen as neutral, independent, and free of political biases, they are often accused of demonstrating a Liberal bias in the US (Shin & Thorson, 2017). Although some studies have found a backfire effect of fact-checks, meaning that such interventions increase rather than decrease the acceptance of misinformation (e.g., Thorson, 2016), more recent literature and meta-analyses have generally indicated that exposure to fact-checking information lowers the acceptance of misinformation and corrects misperceptions (e.g., Walter et al., 2020). Yet, talking about disinformation and its threats may not only have positive ramifications. In line with this, accusations of disinformation are found to result in lower levels of trust in the media or factually accurate information (e.g., Egelhofer et al., 2022; Van Duyn & Collier, 2019), which illustrates that mis- and disinformation used as delegitimizing labels may negatively impact trust in factually accurate information.

As warning labels and media literacy interventions are both based on the same general strategy of pre-bunking mis- and disinformation (Hameleers, 2022), we integrate these approaches to explore the (un)intended effects of discussing the harms of mis- and disinformation *before* people are exposed to either factually accurate or false information. More specifically, many interventions used to warn people about false information contain both an explanation of the critical skills that are needed to verify information (e.g., Moore & Hancock, 2022) and a warning about the severity of mis- and disinformation and its presence in online news settings. As this article aims to map how the combination of teaching critical media literacy skills and a warning about the omnipresence of falsehoods—popular means of governmental interventions used to warn people about mis- and disinformation—may backfire by lowering the acceptance of factually accurate information, we focus on a combination of media literacy skills and a warning about mis- and disinformation in people's newsfeeds.

Despite the optimistic findings of previous experimental research on the effects of pre-bunking techniques, interventions that warn about harmful and false information may also lower the perceived authenticity of factually accurate information. Hence, such messages may trigger suspicion by cultivating the perception that false information is a highly salient issue. Even though media literacy interventions can increase the accurate discernment between false and real headlines, exposure to such information has also been found to reduce the perceived credibility of real news (Guess et al., 2020). In support of this, Modirrousta-Galian and Higham (2022) re-analyzed five experimental studies and found that gamified inoculation techniques that pre-bunk mis- and disinformation are less effective than assumed, as they trigger skepticism related to both false and real news. Thus, to comprehensively measure the effectiveness of warning labels and media literacy interventions, it is important to distinguish between the acceptance of misinformed claims and factually accurate information. Considering that experimental research found that being exposed to a warning label before seeing factually accurate information can also lower the trustworthiness of true information (Hameleers, 2022; Modirrousta-Galian & Higham, 2022), this article explores how the presentation of a warning label combined with a media literacy intervention placed before factually accurate information and mis- or disinformation influences the rating of both types of information.

Based on research on the effects of malign fake news labels and discussions or interventions targeting mis- and disinformation, it can be argued that talking about mis- and disinformation can have negative effects on the credibility and trust of real news. Therefore, it is crucial to turn our attention to disinformation as more than an informational crisis. Deceptive information, albeit problematic, may only take up a marginal proportion of today's newsfeed (Acerbi et al., 2022). Although exposure to false information can have real effects on people's beliefs, evidence on the impact of mis- and disinformation is mostly based on findings collected in the short-term and in controlled lab experiments—where people are forcefully exposed to mis- and disinformation that they may not select in real life (see e.g., Dobber et al., 2020; Hameleers et al., 2020; Zimmermann & Kohring, 2020). This begs the question of whether current interventions, media discourses, and political discussions about the alleged uncontrolled dissemination and impact of mis- and disinformation are legitimized or disproportionate. Is the treatment emphasizing the dangers of disinformation coming with more severe side effects threatening the credibility of real information? Considering that the large majority of all information is factually accurate (e.g., Acerbi et al., 2022), such side effects are worrisome, and can potentially amplify the increasingly more relative status of factual knowledge and evidence. Before reporting on a case study to test these side effects in the context of a media literacy intervention, we will consider the wider processing biases and mechanisms that may explain the problematic consequences of mis- and disinformation as a discursive phenomenon.

2.2. The Potential Shift From Truth-Default to Deception Default in a Post-Truth Ecology

In line with the skewed distribution between false and truthful information in people's newsfeeds, people have a tendency to rate incoming information as honest and accurate (Levine, 2014). This tendency favoring honesty over deception has been formalized in the Truth-Default-Theory (TDT) (Levine, 2014). Being biased toward truthfulness serves as a heuristic that enables people to deal with the overload of information in their information ecology. As people do not have the resources to fact-check the truthfulness of all information that they encounter, the truth-default state serves as a filtering mechanism that helps people to cope with information overload. The TDT postulates that people accept the honesty of incoming information by default, and more or less heuristically, unless the communication context triggers suspicion.

People can deviate from the truth-default state when certain trigger events are causing suspicion of deception (Levine, 2014). These trigger events may, for example, comprise a lack of coherency within the argument structure of new information, or a discrepancy of new information with the external reality and existing information (Luo et al., 2022). Moreover, and relevant in the context of mis- and disinformation, perceived deception motives, dishonest demeanor, and logical inconsistencies of information with known facts or within narratives may all be considered as potential triggers of suspicion (Clare & Levine, 2019). Although interventions intended to fight mis- and disinformation may serve as trigger events that help people to detect deceptive information, they may also cause suspicion that is consequentially applied to real news. Hence, when media literacy interventions tell people to look out for suspicious content and be aware of deceptive news, their overall levels of cynicism and doubt may be triggered, which also increases the likelihood that they perceive real news as false.

On a more general level, we should consider whether the (potentially disproportionate) attention to mis- and disinformation, the weaponization of fake news, and the wide accessibility of counter-epistemic communities online (Waisbord, 2018) are causing a shift from truth-default to deception-default in veracity judgments. In contrast to the TDT, literature on the deception bias and interpersonal deception theory postulate that people are monitoring their information environment for deception while being sensitive to truthful information (Bond et al., 2005; Burgoon, 2015). Considering survey data illustrating that less than half of all respondents (42%) across 46 countries trust the news media most of the time (Newman et al., 2022), it could be argued that people are not very likely to be biased toward the truth.

To summarize, we argue that, above and beyond deceptive information, warning people about the negative impact of mis- and disinformation could be considered as a trigger event that causes suspicion and motivates people to deviate from the truth-default state. Considering that people are very concerned about false information, whereas they are likely to distrust the news (Newman et al., 2022), the salience of mis- and disinformation discussions in media discourse and public opinion may promote a deception bias in a post-factual information era. As warning labels and media literacy messages may enhance general levels of cynicism and trigger suspicion applied to both factually accurate information and mis- or disinformation (e.g., Modirrousta-Galian & Higham, 2022), we expect that exposure to warnings about mis- and disinformation enhances general distrust to all subsequent information. In our study, we specifically investigate the effects of exposure to a pre-bunking message that primes suspicion by emphasizing the threats of mis- and disinformation. As this can be considered a trigger event for perceived deception (Levine, 2014), we introduce the following central hypothesis: Exposure to a pre-bunking message that warns recipients about the dangers of mis- and disinformation lowers trust in both factually accurate information and misinformation.

3. Case Study: How Interventions May Backfire as a Consequence of Priming Deception

3.1. Data Collection

To assess the (unintended) effects of media literacy interventions, we conducted an experiment in the US and the Netherlands ($N = 377$). In the first part of the experiment, after asking pre-treatment questions on age, gender, education, ideology, media use, and perceptions, 50% of the sample was exposed to a media literacy intervention. The other 50% did not see such an intervention. In the next step, participants were randomly exposed to mis- or disinformation with an anti-immigration stance or an evidence-based message that was factually accurate (for the stimuli, see Appendix A of the Supplementary File). This topic was chosen as it is central in disinformation campaigns in both Europe and the US (e.g., Bennett & Livingston, 2018; Marwick & Lewis, 2017). As it strongly reflects a delegitimizing and politicized anti-immigration narrative, the stimulus could be regarded as intentionally false (disinformation). However, the message may also be disseminated without harmful intentions (i.e., by ordinary citizens). For this reason, the false message central in this experiment can be both mis- and disinformation, depending on the communicator and the context of dissemination.

We compared the US and the Netherlands as these countries have been regarded as contrasting cases when it comes to resilience toward disinformation (Humprecht et al., 2020). The Netherlands, as a less polarized multiparty context with relatively high levels of trust in established media, could be considered relatively resilient: Citizens may be more likely to trust established and verified sources than counter-epistemic platforms, and the setting of different political parties makes it more difficult to raise cynicism or reinforce existing cleavages in society. In the US, however, the extremely low level of trust in the news may correspond to a more vulnerable context for disinformation—only 26% of all citizens trust the media most of the time according to the 2022 Reuters Institute Digital News Report (Newman et al., 2022). The high level of polarization further offers a discursive opportunity structure for malign actors to reinforce bi-partisan cleavages (Humprecht et al., 2020). The overall goal of the comparison is to explore whether similar effects of disinformation discussions—in the form of a media literacy message warning about disinformation—can be observed in settings with different levels of resilience to disinformation. We do not formulate directional hypotheses about the (un)intended effects of interventions between these two settings as levels of resilience could both dampen or enhance the effects of interventions. More specifically, lower levels of resilience in the US compared to the Netherlands could mean that there is more room for media literacy interventions to improve the discernment between false and accurate information. Alternatively, however, lower resilience could also imply that interventions are less likely to be accepted by citizens with higher levels of institutional distrust, such as the case in the US. Against this backdrop, we leave it an open question whether differential levels of assumed resilience to mis- and disinformation also translates into different (un)intended effects of warning labels.

3.2. Sample

We achieved 377 completes (188 US participants; the completion rate was 80.8%). Of these completes, 46.9% identified as female. The final sample closely reflects the populations' distribution on educational level, and shows a non-skewed distribution of the different levels of education. More specifically, 47.2% had a moderate level of education, whereas 24.1% and 28.6% had a lower and higher level of education, respectively. Ideology was equally distributed across the left-right divide: 39.5% identified as (somewhat) left-wing and 49.1% as (somewhat) right-wing (11.4% did not know or did not want to say). These distributions reflect the voting behavior of the different populations included in the study. Finally, the mean age of the participants was 43.30 years (*SD* = 14.31).

3.3. Exposure to a Media Literacy Intervention

Based on conventional interventions used in both countries, we designed a media literacy intervention that warned people about mis- and disinformation, whilst offering concrete suggestions on how to detect falsehoods (see Appendix A in the Supplementary File). As we based ourselves on templates that have been used by media literacy organizations in different contexts, we consider the intervention as externally valid. In the intervention, three key suggestions for media users were foregrounded: (a) It is important to check the source of messages; (b) it is important to look for factual information, and critically assess whether these facts match reality; and (c) it is suggested to look for logical argumentation styles. This way, the intervention also follows the theoretical premises of the trigger events of deception detection postulated in the TDT. The intervention is similar to media literacy messages tested in previous experimental research (e.g., Guess et al., 2020; Tully et al., 2020). Similar to the approach taken by Tully et al. (2020), the media literacy message emphasized the need to "spot fake news," for example by verifying the validity of the message's source. To enhance ecological validity, the intervention we used aimed to mimic existing infographics, warning messages, and online suggestions often encountered as close as possible. The intervention was, for example, based on the Dutch' governments public service announcements on how fake news should be spotted (the governmental platform of the national media literacy organization was used as inspiration). The intervention was also based on online lists of suggestions forward by fact-checking organizations such as factcheck.org. Although the interventions were similar in terms of argument structure, style, and other features, they were tailored to the different national settings (i.e., the US intervention talked about mis- and disinformation in the US).

3.4. Measures: The Perceived Credibility of (Dis)Information

After seeing the disinformation article or the fact-based information, participants evaluated the perceived credibility of the message based on the following statements (all measured on 7-point disagree-agree scales): (a) the message is truthful; (b) the message is accurate; (c) the message is based on false assumptions (reverse-coded); and (d) the message tried to deceive me (reverse-coded). To equally prime a truth and deception bias, we used a 50:50 mixture of items emphasizing the truthfulness versus dishonesty/lack of facticity of the messages. The items formed a reliable average scale of perceived credibility (*M* = 3.94, *SD* = 1.26, Cronbach's alpha = .849).

3.5. Procedures and Manipulation Checks

The data collection for the experiment was completed by the international research agency Dynata in March 2020, which has a large mixed-resources database of diverse panelists across countries. Potential participants were recruited via e-mail. Upon accepting the invitation from the company, participants were forwarded to the online survey via a redirect link. Upon entering, participants first of all completed questions on socio-demographics and background variables that could be used as controls (i.e., left-right ideological self-placement). Then, they were randomly allocated to the media literacy intervention or a filler survey block (control condition). In this control condition, they read information about a recipe (a non-political message that scored low on arousal and political content but equal in length). After that, participants were randomly assigned to the factual information or mis- and disinformation condition (equal group sizes). After reading the (dis)information, participants evaluated the perceived credibility of the statements they were exposed to (the main dependent variable).

In the final survey block, participants answered a series of manipulation check questions that asked them to remember the statements of the mis- and disinformation treatment and the media literacy interventions. The manipulations succeeded: Participants in the mis- and disinformation condition were significantly and substantially more likely to associate the message with falsehoods emphasizing increasing crime rates caused by migrants (M = 5.17, SD = 1.54) compared to participants exposed to fact-based information (M = 3.55, SD = 1.67, p < .001). Likewise, participants who were not exposed to a media literacy intervention were overall not likely to perceive to be warned about mis- or disinformation and its effects (M = 2.98, SD = 1.80), whereas participants exposed to the media literacy intervention remembered the media literacy's warning about the threats of disinformation (M = 5.25, SD = 1.57, p < .001).

Randomization checks confirmed that there were no differences between groups in terms of gender, age, level of education or ideological preferences. These factors were thus equally distributed across groups. Controlling for these factors did not influence any of the findings reported in this article.

Participants were carefully debriefed in the final step of the survey. As part of this procedure, they were presented with a fact-check that outlined the (un)truthfulness of the message they were exposed to. Participants were informed about why they were deceived, and the fictionality of the intervention was also emphasized. In addition to this, links with further reading on misinformation, as well as existing media literacy initiatives, were offered. All participants were confronted with this debriefing information, even if they left the survey earlier (i.e., a re-direct was implemented).

4. Results

To compare the perceived credibility of factually accurate information versus mis- or disinformation pre-bunked with a media literacy message, we ran a one-way ANOVA. The four conditions were included as a categorical independent variable, and the mean credibility scale was included as a dependent variable (see Table 1 for the mean scores across conditions). The results of the ANOVA indicate that there are significant differences in the perceived credibility of the messages shown in the different conditions (F(3, 373) = 2.81, p = .039, partial η^2 = .022). Inspecting the pairwise mean score comparisons, we only see only one significant between-conditions difference in the perceived credibility of the information participants were exposed to. More specifically, participants exposed to factually accurate information without a media literacy intervention are more likely to perceive the message as credible (M = 4.21, SD = 1.16) compared to participants exposed to mis- or disinformation that was preceded by a media literacy intervention (M = 3.72, SD = 1.16). Without a media literacy intervention, however, there was no significant difference in the credibility of factually accurate information (M = 4.21, SD = 1.16) and disinformation (M = 3.93, SD = 1.27). Only when the media literacy message was present (compared to absent), participants rated disinformation as significantly less credible than factually accurate information (ΔM = .49, SE = .16, p = .002). Yet, the differences between mis- or disinformation with and without pre-bunker are very small. Table 1 shows that these findings are most pronounced in the US—as the differences are not demonstrated in the Netherlands.

Table 1. Mean score differences across all four conditions.

	No pre-bunking message			Pre-bunking message		
	Total *M* (*SD*)	US *M* (*SD*)	NL *M* (*SD*)	Total *M* (*SD*)	US *M* (*SD*)	NL *M* (*SD*)
Mis- or disinformation	3.93^b (1.27)	3.50^a (1.25)	4.33^b (1.17)	3.72^a (1.16)	3.47^a (.97)	4.02^a (1.30)
Factually accurate information	4.21^b (1.16)	4.33^b (1.34)	4.06^b (.91)	3.95^a (.98)	3.87^a (1.01)	4.02^a (.95)
F, Df *(3, 377)*	2.81*					
Partial η^2	.022					

Notes: Means with different subscripts (^{a}, ^{b}) within columns are significantly different (p < .05); * p < .05.

More relevant for our main research question, we also compared the effect of exposure to a media literacy message in response to both real information and disinformation. Comparing the credibility ratings of factually accurate information with and without this intervention, we see no significant difference in the perceived credibility between exposure to a media literacy intervention presented before factually accurate information and the absence of such a message (ΔM = .26, SE = .17, p = .71). Based on this finding, we can conclude that the media literacy intervention did not significantly lower the credibility of factually accurate information.

We find a similar effect for the media literacy intervention preceding mis- or disinformation: The perceived credibility of mis- and disinformation is not significantly different for participants exposed to a media literacy intervention than participants that did not see such a forewarning message (ΔM = .21, SE = .17, p = 1). Although our findings suggest a difference in perceived credibility between factually accurate information that is not preceded by a media literacy message and pre-bunked mis- or disinformation, presenting a media literacy message before factually accurate information diminishes the potentially positive effect of media literacy interventions. On average, the credibility of both mis- or disinformation and real information is lowered by the intervention so that the difference between pre-bunked mis- or disinformation and factually accurate information is not significant.

In the final step, we repeated the analyses for both country cases separately (see Table 1). The country-specific analyses show that the mean score differences are most pronounced in the US. Although fact-based information is more credible than corrected mis- or disinformation in both the US and the Netherlands, the difference is most pronounced in the US (ΔM = .87, SE = .25, p = .005). In the US, but not in the Netherlands, there is a significant and substantial mean score difference between factually accurate information and disinformation in the absence of a media literacy intervention (ΔM = .84, SE = .25, p = .005). Although the intervention makes factual information less credible (ΔM = .49), the intervention does not have a significant effect on the credibility of mis- or disinformation (ΔM = .02). The findings thus point to a harmful side-effect of the pre-bunking intervention in the US: The media literacy message lowers the credibility of factually accurate information, and does not have a significant effect on the credibility of mis- or disinformation. Comparing the two country cases, it can be concluded that the intervention had more negative side effects than positive intended effects in the US. We can also conclude that the averaged findings across both country settings are driven by the US: There are no significant differences if we look at the Netherlands separately.

As a main finding, it can be concluded that the presence of a pre-bunking intervention in the form of a warning label combined with suggestions on how mis- and disinformation should be detected diminishes the credibility gap between factually accurate information and mis- or disinformation in the US. Hence, the intervention significantly lowers the credibility of factually accurate information, but not mis- or disinformation (see Table 1).

5. Discussion

What to make of these findings? Although warning people about the threats of mis- and disinformation, for example by exposing them to media literacy messages, can reduce the perceived credibility of mis- and disinformation, it can also harm the credibility of factually accurate information. Even more so, taking into account the negative side-effect of a media literacy message on the perceived credibility of factual information, the positive impact of the media literacy message can diminish. This was especially shown in the US, where people were better able to discern truthful information from mis- or disinformation without an intervention than with the presence of a media literacy intervention. The warning seemed to confuse news users, and made the task of truth and deception discernment more difficult.

This illustrates that media literacy messages that discuss the threats of mis- and disinformation may work most efficiently when they are shown to people that expose themselves to mis- or disinformation. When people are warned about mis- or disinformation but exposed to factual news, the intervention may harm trust in factually accurate information. In practice, targeting pre-bunking measures to mis- or disinformation exposure is virtually impossible. Hence, looking at existing applications of inoculation strategies, media literacy messages, and other pre-bunking information (e.g., Roozenbeek & van der Linden, 2019), these interventions are presented as general warnings that do not take into account people's preferences and selection patterns for mis- or disinformation compared to reliable information.

Although the effectiveness of news media literacy interventions and pre-bunking techniques requires more empirical research, our findings suggest that it is important to look at the discernment between factually accurate information and mis- or disinformation to measure the effectiveness of interventions (see also e.g., Modirrousta-Galian & Higham, 2022). Hence, under some conditions, exposure to a media literacy intervention can lower the credibility of information irrespective of its veracity. Overall, the effectiveness of interventions should be considered in the wider context of the prevalence of factually accurate, reliable information vis-à-vis mis- or disinformation (e.g., Acerbi et al., 2022; Allen et al., 2020). More specifically, mis- and disinformation is estimated to make up about 1% of news users' media diets (e.g., Acerbi et al., 2022). Although this estimate excludes blogs, chat groups, and communication in closed communities, it can be argued that the large majority of information that people are exposed to is reliable and factually accurate. Against this background, the trend suggested by this study is that the presence of media literacy interventions can also result in lower levels of credibility when presented before factually accurate information may have consequences for how pre-bunking techniques could be presented and placed in people's newsfeeds. Although this study has only offered limited and preliminary evidence for this, future research needs to further explore under which conditions media literacy interventions may have unintended consequences that could harm their effectiveness.

This negative side-effect of talking about the threats of false information corresponds to Acerbi et al.'s (2022) recommendation to focus interventions on enhancing the credibility of reliable and factually accurate information instead of combating mis- and disinformation. Hence, based on the substantially lower prevalence of deceptive compared to factually accurate information in people's newsfeeds, it may be equally effective to enhance the trustworthiness of factually accurate information with 1% as to lower the acceptance of mis- and disinformation to zero. However, to date, most interventions and media literacy programs target the threats and perceived effectiveness of false and deceptive information (e.g., Hwang et al., 2021; Jones-Jang et al., 2021). Therefore, it may be important for interventions to relativize the threats associated with mis- and disinformation and to additionally focus on the consolidation of trust in authentic and factually accurate information and news sources.

In line with this suggestion, a practical implication of this study is to—after conducting more research on the unintended consequences of pre-bunking messages—reconsider media literacy interventions, as well as journalists' and media practitioners' attention to mis- or disinformation as a threat. More specifically, it may be worthwhile to emphasize that misinformation—although a problematic and amplified issue—is far less prevalent than factually accurate information. As part of media literacy programs, it is relevant to offer suggestions and practical recommendations on how news users can find reliable and trustworthy sources. Concretely, a list of suggestions on how the reliability and accuracy of trustworthy information and news sources may be assessed can be included, supplemented with source recommendations for verified news on different issues. Considering that all sources can be wrong sometimes, news users should be recommended to not blindly accept sources that are likely to contain factually accurate information. Yet, media literacy interventions can offer concrete guidance on how critical news users can discern truthful information from intentionally deceptive sources.

Although these media literacy interventions should motivate citizens to be critical toward information—including the information from reliable sources that are recommended—it is important to not instill cynicism. One suggestion in that regard would be to inform people about the reasons why information may not be accurate (i.e., due to a lack of expert knowledge, changing analyses, or updated evidence), instead of triggering suspicion related to perceived intended manipulation and conspiracies. Thus, although moderate misinformation perceptions may be conducive to media trust and news selection, mis- and disinformation perceptions that revolve around the perceived dishonesty of the press may be more detrimental to media trust and democracy at large (Hameleers et al., 2020). In any case, it is important that media literacy interventions focus on truth discernment instead of mis- and disinformation detection.

Theoretically, our findings offer relevant input for the truth versus deception default theory (Levine, 2014). Although it has traditionally been argued that people are biased toward the truth, and that people accept the honesty of information unless deception is triggered (Levine, 2014), discussions about mis- and disinformation may be an important trigger event. Hence, media literacy messages, the weaponization of fake news, and public and politicized debates on how to "fight" mis- and disinformation may contribute to a gradual shift toward a deception bias (also see Bond et al., 2005; Burgoon, 2015). As suggested by the findings of our study, people may become overly sensitive toward deception as frames and discourses around mis- and disinformation are omnipresent, potentially resulting in an overestimation of the threat. Although people should not, at least from a normative perspective, blindly follow a truth bias, the complexity of currently digitized information ecologies dictates that people should be wary about the likelihood to encounter deception or truthfulness in different communication contexts. Specifically, when people consume news on social media and when relevant expert knowledge is absent, they should arguably be more critical and motivated to verify information than when they use established news sources that quote different relevant expert sources.

Despite offering new insights into the dynamics of accuracy judgments driven by deception and truth biases in the context of mis- and disinformation, this study comes with a number of substantial limitations. First and foremost, we base ourselves on just one specific media literacy intervention and the subsequent effect of the credibility of just one mis- or disinformation and factual news article on one issue. It could be the case that this specific media literacy intervention caused cynicism and distrust due to its framing, and it could be argued that the difference between the deceptive and the factually framed article was too small to detect meaningful differences in perceived credibility. We thus need more research using different types of interventions applied to different forms of mis- and disinformation to triangulate and validate the trends suggested here. Although we based ourselves on real mis- or disinformation articles and media literacy interventions, we consider this as one case study that needs to be supplemented with more robust evidence based on different interventions and a more diverse selection of mis- and disinformation narratives. We consider the artificiality of the experimental setting as another shortcoming. Although the media literacy intervention was not directly placed in front of the articles that participants had to evaluate, there is arguably more distance and distraction between warnings and actual (mis)information in people's newsfeeds. This means that the real-life effects of interventions may be smaller in real life. However, discussions on mis- and disinformation and warnings about their threats reach people in a multitude of formats that are constantly repeated. Therefore, the findings of this study may also underestimate the effects of warning people about mis- and disinformation. This calls for more realistic study designs that take delayed exposure as well as repetition into account, for example, by relying on a multi-wave survey experiment.

Yet, this study's aim was to offer first tentative evidence on the possibility that well-intended responses to mis- and disinformation could, under some conditions, have negative unintended side-effects on the trustworthiness of reliable information. We believe that our findings at the very least call for some sensitivity to undesired effects of mis- or disinformation as a discursive issue, and more research that takes unintended consequences into account when mapping the effectiveness of interventions.

Conflict of Interests

The author declares no conflict of interests.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

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

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14), Article eaay3539. https://doi.org/10.1126/sciadv.aay3539

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

Bond, G. D., Malloy, D. M., Arias, E. A., Nunn, S. N., & Thompson, L. A. (2005). Lie-biased decision making in prison. *Communication Reports*, *18*(1/2), 9–19. https://doi.org/10.1080/08934210500084180

Burgoon, J. K. (2015). Rejoinder to Levine, Clare et al.'s comparison of the Park-Levine probability model versus interpersonal deception theory: Application to deception detection. *Human Communication Research*, *41*(3), 327–349. https://doi.org/10.1111/hcre.12065

Chadwick, A., & Stanyer, J. (2022). Deception as a bridging concept in the study of disinformation, misinformation, and misperceptions: Toward a holistic framework. *Communication Theory*, *32*(1), 1–24. https://doi.org/10.1093/ct/qtab019

Chan, M. P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, *28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Clare, D. D., & Levine, T. R. (2019). Documenting the truth-default: The low frequency of spontaneous unprompted veracity assessments in deception detection. *Human Communication Research*, *45*(3), 286–308. https://doi.org/10.1093/hcr/hqz001

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. https://doi.org/10.1177/1940161220944364

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. https://doi.org/10.1080/23808985.2019.1602782

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. https://doi.org/10.1093/joc/jqac031

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

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, *117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Hameleers, M. (2022). Disinformation as a context-bound phenomenon: Toward a conceptual clarification integrating actors, intentions and techniques of creation and dissemination. *Communication Theory*, *33*(1), 1–10. https://doi.org/10.1093/ct/qtac021

Hameleers, M., Powell, T. E., Van Der Meer, T. G., & Bos, L. (2020). A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, *37*(2), 281–301. https://doi.org/10.1080/10584609.2019.1674979

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

Hwang, Y., Ryu, J. Y., & Jeong, S. H. (2021). Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 188–193. https://doi.org/10.1089/cyber.2020.0174

Jones-Jang, S. M., Mortensen, T., & Liu, J. (2021). Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *American Behavioral Scientist*, *65*(2), 371–388. https://doi.org/10.1177/0002764219869406

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. https://doi.org/10.1177/0261927X14535916

Luo, M., Hancock, J. T., & Markowitz, D. M. (2022). Credibility perceptions and detection accuracy of fake news headlines on social media: Effects of truth-bias and endorsement cues. *Communication Research*, *49*(2), 171–195. https://doi.org/10.1177/0093650220921321

Marwick, A., & Lewis, R. (2017). *Media manipulation and disinformation online*. Data & Society. https://datasociety.net/output/media-manipulation-and-disinfo-online

Modirrousta-Galian, A., & Higham, P. A. (2022). *How effective are gamified fake news interventions? Reanalyzing existing research with signal detection theory*. PsyArXiv. https://psyarxiv.com/4bgkd

Moore, R. C., & Hancock, J. T. (2022). A digital media literacy intervention for older adults improves resilience to fake news. *Scientific Reports*, *12*(1), Article 6008. https://doi.org/10.1038/s41598-022-08437-0

Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson, C. T., & Nielsen, R. K. (2022). *Reuters Institute digital news report 2022*. Reuters Institute. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf

Roozenbeek, J., & van der Linden, S. (2019). The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research*, *22*(5), 570–580. https://doi.org/10.1080/13669877.2018.1443491

Shin, J., & Thorson, K. (2017). Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication*, *67*(2), 233–255. https://doi.org/10.1111/jcom.12284

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

Tully, M., Vraga, E. K., & Bode, L. (2020). Designing and testing news literacy messages for social media. *Mass Communication and Society*, *23*(1), 22–46. https://doi.org/10.1080/15205436.2019.1604970

Uscinski, J. E., & Butler, R. W. (2013). The epistemology of fact checking. *Critical Review*, *25*(2), 162–180. https://doi.org/10.1080/08913811.2013.843872

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. https://doi.org/10.1080/15205436.2018.1511807

Vraga, E. K., & Bode, L. (2020). Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication*, *37*(1), 136–144. https://doi.org/10.1080/10584609.2020.1716500

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

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

Wardle, C. (2017, February 16). *Fake news. It's complicated*. First Draft. https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79

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

Zimmermann, F., & Kohring, M. (2020). Mistrust, disinforming news, and vote choice: A panel survey on the origins and consequences of believing disinformation in the 2017 German parliamentary election. *Political Communication*, *37*(2), 215–237. https://doi.org/10.1080/10584609.2019.1686095

About the Author

Michael Hameleers (PhD, University of Amsterdam, 2017) is assistant professor in political communication at the Amsterdam School of Communication Research, Amsterdam, The Netherlands. His research interests include populism, disinformation, and corrective information. He has published extensively on the impact of populism, (visual) disinformation, fact-checking, media literacy interventions, and (media)trust in leading peer-reviewed journals.