Is the alarm on deception ringing too loudly? The effects of different forms of misinformation warnings on risk perceptions of misinformation exposure

Hameleers, M.

DOI
10.1177/02673231241271015

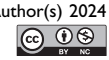
Publication date
2024

Document Version
Final published version

Published in
European Journal of Communication

Full Length Article

European Journal of Communication
2024, Vol. 39(4) 360–374



DOI: 10.1177/02673231241271015
journals.sagepub.com/home/ejc

Is the alarm on deception ringing too loudly? The effects of different forms of misinformation warnings on risk perceptions of misinformation exposure

Michael Hameleers
Political Communication at the Amsterdam School of Communication Research (ASCoR), University of Amsterdam, The Netherlands

Abstract

Misinformation is widely regarded as an undermining force to European democracies. Yet, to date, empirical research shows that the amount of misinformation people encounter is rather low, and not in proportion to the strong alarming messages spread throughout society. In this light, current interventions that pre-bunk misinformation by using warning messages may disproportionally prime suspicion and result in inflated estimates of misinformation. To assess whether messages that pre-bunk misinformation result in disproportionate risk perceptions related to inaccurate or false information, and to explore the effectiveness of alternative interventions, this article relied on an online between-subjects experiment in the Netherlands ($N = 437$). Our main findings indicate that exposure to a media literacy intervention does not result in higher first- or third-person risk perceptions related to misinformation exposure. However, a warning message that emphasizes the identification of reliable news while contextualizing the threats of misinformation significantly lowers perceived misinformation salience. As an important implication of our findings, we suggest that pre-bunking interventions should relativize the threats of misinformation by facilitating the recognition of honest and reliable information as an alternative path to help people identify reliable information.

Keywords

Misinformation, disinformation, pre-bunking, media literacy, inoculation, corrective information

Corresponding author:
Michael Hameleers, Political Communication at the Amsterdam School of Communication Research (ASCoR), University of Amsterdam, The Netherlands 1000 GG
Email: m.hameleers@uva.nl

Misinformation, which can be understood as an umbrella term for false or inaccurate information that is either spread unintentionally or intentionally (e.g. Wardle and Derakhshan, 2017), has been regarded as an undermining influence across democracies (e.g. Bennett and Livingston, 2018). Although pre-bunking approaches to misinformation, such as inoculation interventions, have been received with great optimism (Roozenbeek et al., 2022), empirical evidence on the effectiveness of pre-bunking is mixed at best (e.g. Hameleers, 2022; Modirrousta-Galian and Higham, 2023; Pennycook et al., 2023). In light of survey research indicating that people across the globe distrust the media most of the time, while being uncertain about how to detect deception (Newman et al., 2023), pre-bunking interventions may unintentionally exacerbate uncertainty related to truth discernment in an era of factual relativism (Van Aelst et al., 2017).

Given that factually accurate or honest information is much more prevalent than misinformation in people's online newsfeeds (e.g. Acerbi et al., 2022), lowering the trustworthiness of all information beyond misinformation potentially indicates that the remedy of pre-bunking is worse than the informational disease it aims to cure. Hence, when people perceive that the information they encounter is flooded with misinformation, they may avoid all information, or become disproportionately cynical toward trustworthy sources of information (Van der Meer et al., 2023). Although the trustworthiness of information may comprise of more than (lacking) facticity, warning people specifically about the threats of alleged omnipresent 'fake news' or information that is inaccurate and false may not provide the accurate skills needed for audiences to discern between information that is more or less trustworthy.

Against this background, this article relies on a survey experiment to answer to what extent pre-bunking messages offering suggestions on how to resist misinformation (dis) proportionally affect people's risk perceptions about encountering misinformation. In the experimental design, we distinguish between various types of warnings to assess the extent to which different pre-bunking interventions (relativized, established, and deception-focused) affect people's estimates of misinformation's risk. The aim is to assess whether a more relativized media literacy message that offers suggestions on how to find trustworthy information may be able to circumvent the disproportionate activation of the perceived prevalence of misinformation.

One explanation for the potential unintended effect of pre-bunking interventions on disproportionate risk perceptions related to misinformation is that current interventions often formulate a generic warning that cultivates beliefs related to suspicious content without pointing people to specific false statements (e.g. Pennycook et al., 2023). An example is the 'How to spot fake news' infographic used by Factcheck.org that offers suggestion on how to spot deception, without offering cues on how reliable news may be found. Such pre-bunking messages function as a trigger event for suspicion, which makes people more aware of the potential risk of misinformation in their environment (Van der Meer et al., 2023).

How to move forward? In this article, we suggest that pre-bunking messages can have unintended side-effects due to the manner in which the warning message is formulated. Specifically, the emphasis on a pre-defined set of indicators (i.e. clickbait headlines, emotionality) may insufficiently relate to the actual nature of misinformation, and primes

suspicion related to all information, including true content. To understand how disproportionate risk perceptions may be remedied, we contrast traditional approaches in pre-bunking interventions to a media literacy message that focuses on the recognition of truthful information whilst relativizing the threats of misinformation.

The findings of the experiment indicate that the different interventions used did not result in overly high-risk perceptions of encountering misinformation. However, the relativized version of the pre-bunking message that placed the warning messages in a wider context of how to detect and access trustworthy information resulted in a significantly lower risk perception when it comes to misinformation exposure. As a major contribution to the misinformation and pre-bunking literature, this study shows how side-effects of pre-warning messages may be mitigated by relativizing and contextualizing the misinformation warning offered in such approaches.

Theory

The perceived prevalence of misinformation

In this article, we define misinformation as an umbrella-term for information that is inaccurate, false, or not based on relevant expert knowledge (Vraga and Bode, 2020). It can refer to both intentionally false information (disinformation, see e.g. Chadwick and Stanyer, 2022) or unintentionally false information (misinformation, see e.g. Wardle and Derakhshan, 2017). Although the distinction between unintentionally false information and goal-directed disinformation is important to make in light of the disruptive consequences of disinformation (e.g. Bennett and Livingston, 2018), associating covert intentions to false content is a complex empirical endeavor. Hence, whether a given piece of false information is disseminated with the intention to deceive depends on the context of communication and the actor-perspective (Hameleers, 2023). On a content level, the same information can thus be misinformation and disinformation depending on who sends it with what intentions.

Considering that the intentions behind the dissemination of false information are only known when the actor-perspective and context of spreading false information are uncovered, we focus on misinformation as an overarching term throughout this article. However, we do consider that misinformation can become disinformation in certain contexts of deception (Hameleers, 2023), and that different motives may drive the communication and creation of statements that lack in facticity. We therefore seek to understand the extent to which highlighting the risks and alleged fixed content indicators central to misinformation (i.e. an emotional tone or the circumvention of experts) as both intentionally and unintentionally false information affects the perception of being at risk.

Although we consider that untrustworthy information can comprise of more than information that lacks facticity intentionally (disinformation) or unintentionally (misinformation), we specifically focus on how media literacy messages aimed to warn people about mis- and disinformation may contribute to risk perceptions of encountering misinformation. Hence, in light of findings suggesting that the perceived threat of misinformation in particular is higher than the base rate of such information (e.g. Acerbi et al., 2022), we postulate that media literacy interventions that aim to provide audiences with

skills to discern trustworthy from untrustworthy information should not trigger too strong levels of suspicion related to fake news. In line with this, we contrast traditional interventions focused on fake news or misinformation with a media literacy intervention designed to provide audiences with the skills to identify trustworthy and untrustworthy information, which may also relate to propaganda, strongly biased reporting, or poor-quality journalism.

Because we are mainly interested in the effects of pre-bunking interventions on (dis) proportional risk perceptions related to misinformation, we regard perceived misinformation exposure as the central dependent variable of the study (also see e.g. Wagner and Boczkowski, 2019). Thus, rather than measuring the actual levels of misinformation people encounter, we ask participants to estimate the amount of false information they receive in their daily media environment. Based on the findings of various content analyses on the actual prevalence of misinformation in the Global North (e.g. Acerbi et al., 2022; Guess et al., 2019; Yang et al., 2023), it can be inferred that misinformation makes up anywhere between .7% (i.e. desktop media consumption in the United States) and 25% (i.e. visuals on Facebook, see Yang et al., 2023) of people's information diets. Although the findings of existing content analyses may exclude certain information platforms that are more prone to misinformation, we believe it is safe to assume that the large majority of the information people encounter online does *not* consist of misinformation (e.g. Acerbi et al., 2022; Guess et al., 2019).

Yet, we do take into account that the actual amount of misinformation people are exposed to may be higher than the extremely low estimates of existing approaches for different reasons. First of all, people agreeing to donate their data or tracking media exposure may be less prone to mis- or disinformed media diets. Second, mapping approaches often focus on hyper-partisan platforms associated with fake news, partisan content, or other problematic forms of false information instead of identifying mis- and disinformation on the content level (e.g. Acerbi et al., 2022). Third, access to data on social media or regular online channels may miss out on mis- and disinformation shared privately or via messaging apps. As such spaces may be vulnerable to mis- and disinformation, the low estimates may underestimate some of the risks related to mis- and disinformation on ungoverned platforms or interpersonal communication.

Finally, it has to be acknowledged that people may associate misinformation with forms of problematic information that move beyond information that lacks facticity intentionally or unintentionally. Hence, qualitative evidence suggests that hyper-partisan content, biased reporting, bad journalism, or highly emotional content are considered as misinformation by audiences (Kyriakidou et al., 2023). As such, people may arrive at high estimates of misinformation as they also consider other distortions as part of this concept.

In line with these prevalent and broad risk perceptions, survey research suggests that about 50% of the population distrusts information most of the time whilst being uncertain about how to discern true from false information (Newman et al., 2023). Distrust may be fueled by elite discourse in which accusations of disinformation and fake news abound (e.g. Van Duyn and Collier, 2019) as well as the mediatized emphasis on the 'uncontrolled' flows of misinformation reaching people on a daily basis. We also consider warning messages that aim to instill resilience to misinformation are part of this

discursive emphasis on the threats of false information (see Van der Meer et al., 2023). But how may interventions that prime suspicion by pre-bunking misinformation affect the estimate of misinformation prevalence?

The merits and pitfalls of pre-bunking misinformation

Generally, extant literature has looked at two different strategies of responding to misinformation through journalistic or educational interventions: De-bunking messages that respond to misinformation directly after its spread (e.g. Thorson, 2016; Wood and Porter, 2019) and pre-bunking interventions offering guidelines and suggestions on how to refute misinformation before it is encountered (e.g. Roozenbeek and Van der Linden, 2019; Tully et al., 2020). In this article, we focus on pre-bunking interventions for two reasons. First, extant literature more or less reached consensus that de-bunking messages such as fact-check interventions are effective in correcting factual misperceptions (e.g. Walter et al., 2020). Yet, they may not have a reach beyond the scope of the fact-checked statements, which may make them less suitable as large-scale interventions to instill general resilience to misinformation.

Second, although the larger reach of pre-bunking messages may instill resilience on a wider scope than specific fact-checks, these interventions may also negatively impact trust in real information (Van der Meer et al., 2023). Especially given the ongoing debate in extant literature on the merits and pitfalls of pre-bunking (e.g. Modirrousta-Galian and Higham, 2023; Pennycook et al., 2023), we believe that it is crucial to further assess the (un)intended effects of pre-bunking messages on risk perceptions related to misinformation. Considering that the risk of pre-bunking interventions is more or less a blind spot in current policies and recommendations communicated to stakeholders, we think it is crucial to explore the extent to which pre-bunking messages may contribute to levels of risk perceptions that are disproportional and undesirable from a democratic viewpoint. Hence, when people come to distrust true and false information to similar extents, and when interventions contribute to uncertainty related to the application of the suggestions to real versus false information, pre-bunking's negative consequences may outweigh its impact on lowering the credibility of misinformation.

Pre-bunking interventions may either be based on the principles of stimulating media literacy through offering constructive suggestions on how to critically navigate the information landscape or comprise of an inoculation message that aims to instill resilience through a small dose of misinformation and subsequent suggestions on how to refute misinformation (Roozenbeek et al., 2022). The central assumption of such interventions is that there is an analogy between false information and viruses: Exposing people to a small dose of this virus may act as a vaccine that protects people against real doses of the threat. Although there are differences between media literacy interventions and gamified inoculation messages, the suggestions they forward for the detection of misinformation often overlap. Hence, they typically point recipients to the emotional nature of misinformation, the role of clickbait headlines and negativity, or stylistic features that may set misinformation apart from other information (e.g. Tully et al., 2020). Crucially, we can discern between various real-life applications of pre-bunking that vary related to the strength of the warning message they forward. Most interventions,

such as the Factcheck.org 'how to spot fake news' campaign highlight the idea of deception, without contextualizing the threat by also suggesting how reliable information may be accessed. However, other initiatives such as the Dutch website and media literacy training program 'is this real?' more clearly forward suggestion on how true and false information may differ, suggesting people to be critical towards information whilst also offering tips on how to find trustworthy news.

Most studies that explore the effectiveness of pre-bunking or media literacy interventions rely on between-subjects experiments (e.g. Modirrousta-Galian and Higham, 2023). Roozenbeek and Van der Linden (2019) randomly exposed students between 16 and 19 to either the experimental condition (gamified inoculation) or a control condition (an unrelated stimulus) and then asked both groups to rate a misinformation story. Modirrousta-Galian and Higham (2023) re-analyzed the experimental evidence from various gamified inoculation approaches and found that the selection of the true and false statements to rate, as well as the extent to which the rating of accurate information is taken into account, can have a large influence on the conclusions drawn. Hence, across the board, they found very limited or even no effects that exposure to a pre-bunking message enhanced truth discernment.

Tully et al. (2020) relied on an experiment conducted with Amazon Mechanical Turk participants, and exposed half of the participants to a control condition unrelated to misinformation and half of participants to an 'how to spot fake news' intervention. After this, they were exposed to misinformation on the flu vaccine. The dependent variable of interest was the credibility rating of this article, and the difference in credibility between the control and treatment condition was used to assess the effectiveness of the intervention.

Although these experimental studies indicate that pre-bunking messages are effective in helping people to detect false information (e.g. Roozenbeek and Van der Linden, 2019), recent evidence suggests that they do not consistently contribute to truth discernment—which is the intended outcome of interventions (e.g. Pennycook et al., 2023). Here, we understand truth discernment as the extent to which people deem false information as less credible whilst maintaining trust in true information as a consequence of being exposed to a pre-bunking intervention. Experimental studies by Modirrousta-Galian and Higham (2023) and Pennycook et al. (2023) failed to replicate the effects of inoculation messages on truth discernment. At the same time, Hameleers (2022) found that media literacy messages offering suggestions on how to detect false information lowered the credibility of misinformation and true information to similar extents—which indicates that truth discernment did not improve as a consequence of exposure to a pre-bunking message. How can we explain these conflicting findings?

In this article, we postulate that the (disproportionate) activation of unguided suspicion is responsible for the potentially unintended effects of pre-bunking messages. The problem here is that the discernment between trustworthy and untrustworthy information is a difficult task for audiences, given the complexity of the information environment and the various forms of biased and problematic information they are confronted with. Yet, current interventions that focus on 'fake news' or disinformation mostly affect subjective interpretations of encountering misinformation, whilst not offering guidance on how to identify trustworthy news.

The effects of the prime of deception offered by current interventions can be understood in the context of the truth-default-theory. In line with the truth-default-theory (Levine, 2014) people have a general tendency to accept new information as honest and true. The theory also holds that people can deviate from the default state of accepting the honesty of information. Such deviations are a likely consequence of so-called 'trigger events' (Clare and Levine, 2019). Crucially, a pre-bunking intervention that tells people to be aware of misinformation can be regarded as a trigger event (Van der Meer et al., 2023). Hence, by telling people that misinformation is a salient threat, or by exposing people to a small dose of misinformation, people may deviate from the truth-default and perceive to be at risk of misinformation exposure. Against this backdrop, and in line with empirical evidence showing that warning people about misinformation fuels distrust in true or factually accurate information (e.g. Van der Meer et al., 2023), we postulate the following hypothesis:

H1: Pre-bunking misinformation results in higher levels of perceived misinformation than the absence of such an intervention.

In line with the truth-default-theory, the strength of the trigger event may play a role in the activation of risk perceptions related to misinformation. Considering that existing pre-bunking messages mostly emphasize a warning for misinformation instead of enhancing the acceptance of real information (Acerbi et al., 2022), they may offer a relatively strong trigger for suspicion. By offering people insights into the content features that allegedly define misinformation, such as emotionality and conflict (e.g. Carrasco-Farré, 2022), they may contribute to a strong risk perception about the abundance of misinformation. After all, as emotional language and negativity are likely to be encountered on a daily basis, people exposed to such an intervention may perceive that misinformation is all around. Yet, in line with the aim to improve awareness rather than cynicism related to (mis)information, recent media literacy interventions in practice have used a more subtle approach that informs people on the potential indicators of misinformation that could warrant further investigation and verification.

Although such formulations may refrain from equating misinformation with a set of pre-defined content characteristics, they still emphasize deception instead of trust. Against this backdrop, in our experiment, we contrast the two aforementioned approaches with an even more relativized pre-bunking message that focuses on finding trustworthy information instead of misinformation (also see the suggestions formulated by Acerbi et al. (2022) in this regard). This alternative 'relativized' condition stresses that misinformation occurs in the context of predominately accurate and honest information. The suggestions forwarded in this alternative pre-bunking interventions emphasize the checks that can be done to find accurate information, instead of detecting deceptive content. For example, the claim that 'fake news is often targeting emotions' was rephrased as 'be critical toward the emotional tone of information' in the relativized media literacy condition focusing on finding trustworthy information. Considering that the level of suspicion triggered by the intervention varies across the three versions of pre-bunking interventions tested in this study, we also expect that they fuel different levels of risk perceptions related to misinformation. Specifically:

H2: Pre-bunking interventions that focus on how fake news can be detected result in higher levels of perceived misinformation than a pre-bunking intervention relativizing and contextualizing the threat of misinformation.

Based on the overview of extant research on the effectiveness of pre-bunking interventions, it can be concluded that although exposure to such interventions may lower the credibility of misinformation (e.g. Roozenbeek and Van der Linden, 2019; Tully et al., 2020), they can also impact the credibility of true information (e.g. Hameleers & van der Meer, 2023). Against this background, this paper aims to explore how varying the strength of the warning message and the threat of deception in interventions may result in higher or lower levels of perceived misinformation.

Method

Design

To test our hypotheses, we rely on a between-subjects experimental design in which participants were randomly assigned to: (1) a control condition in which a nonrelated list of suggestions on how to use search engines was presented; (2) an established media literacy message following the argumentation and structure of the most prevalent examples of pre-bunking applied in Europe and beyond; (3) a deception-focused pre-bunking message with a stronger warning about misinformation and an explicit reference to the defining content characteristics of misinformation; (4) a relativized media literacy intervention that relativized the threats of misinformation by focusing the suggestions on finding trustworthy information instead of ringing the alarm on misinformation. The group sizes across conditions were equal, and post hoc randomization checks confirmed that the conditions did not differ significantly in the composition of key factors, including age, gender, education, and ideology.

In Supplemental Appendix A, the scripts used in the interventions and control condition are included. Here, we also included screenshots of the lay-out of the interventions that closely matched the application of media literacy interventions used in educational materials and public service announcements. Hence, to enhance external validity, we based the design of the tested intervention on existing pre-bunking approaches developed in the context of the study.

All three pre-bunking messages essentially presented the same suggestions on how media users can navigate their information ecology and check for the presence of trustworthy or deceptive information. All pre-bunking messages were based on existing content analyses on the characteristics associated with misinformation (e.g. Damstra et al., 2021) and the actual tips offered by media literacy organizations in the Dutch setting (e.g. the media literacy foundation, 2023). We extended the list of indicators and suggestions with recent evidence pointing to the problematic role that expert knowledge and empirical evidence can play as legitimacy enhancing tools for misinformation (e.g. Hameleers and Yekta, 2023). We also referred to the role that images can play as 'proof' for deceptive claims—even if the images themselves are authentic (see e.g. Brennen et al., 2021). The resulting pre-bunking messages included tips referring to

the emotional tone of the message, the context of social media, the intentions of communicators, the role of visual decontextualization and manipulation, the importance of reading the news article, and paying attention to the source information of the text.

Concretely, the *established media literacy intervention* that was based on existing applications in Europe and beyond (i.e. 'is this real' or the 'how to spot fake news' tips of Factcheck.org) offered suggestions on useful checks to assess whether information could be fake or deceptive. Here, no clear suggestions on how to detect trustworthy information were offered. For the second treatment condition presenting a *deception focused pre-bunking message*, the warning message was framed in a stronger manner: Instead of pointing to potential indicators that warrant further investigation, the message associated the same content features with deception and fake news. Finally, the third *relativized media literacy* condition reversed the emphasis from deception to trustworthy news. Although it relied on the same arguments and suggestions to check information, the warning message framed them as suggestions to check for reliable and trustworthy information instead of looking out for deception.

To avoid priming beliefs about misinformation and its risks, the control condition referred to a completely different issue. Suggestions for using search queries in Google and other search engines. This message was formulated in a similar way as the treatment conditions in the sense that it offered a list of suggestions for online news use. However, the control condition did not refer to the honesty, accuracy, deception, or bias of news online. It rather argued how people could obtain more relevant findings by using different terms and formulations of their questions in search engines. Inspecting open-ended answers of participants exposed to the control condition, we can confirm that none of the participants exposed to this condition thought about fake news, misinformation, or the accuracy of news. In Table B1 of Supplemental Appendix B, a table with an overview of all the conditions is included.

Sample

Participants were recruited by Kantar, an international research agency with access to large and nationally representative panels of participants across countries. We conducted our experimental study in the Netherlands—a country that is relatively resilient to mis- and disinformation due to high levels of media trust and low levels of polarization (Humprecht et al., 2020). Considering that citizens in this country are often reminded of the harms of misinformation, and given that different interventions are presented to citizens with the intention to enhance resilience and media literacy, we perceive it as a theoretically relevant context for understanding whether warning about misinformation to different degrees may fuel risk perceptions. In other words, the potentially high discrepancy between risk perceptions and actual resilience to misinformation may make the Netherlands a relevant case for understanding the effects of different interventions.

Based on an a priori power analysis, we estimated to need about 100 completes per cell to obtain sufficient power (.80). As we are mainly interested in the direct effects of the interventions, we did not require a larger sample that could also result in small and less important differences to be significant. In total, we achieved 437 valid completes in the experiment (the completion rate was 94.2%). The average age of participants

was 44.56 years ($SD = 14.49$). 45.3% identified as male (.2% other and 54.5% female). A total of 16.0% was lower educated, whereas 38.9% obtained a higher level of education (45.1% moderate). These distributions reflect the Dutch population closely. Finally, the sample composition reflects the vote share of left-wing and right-wing parties in the current political landscape, with 45.1% identifying as (mostly) right-wing and 39.1% as (mostly) left-wing. All in all, we obtained a sample that reflects the Dutch population regarding age, gender, education, and ideology as close as possible.

Measures

We assessed the (dis)proportionality of risk perceptions related to perceived misinformation exposure by asking participants to indicate their perceived level of exposure to false information. On a scale from 0 to 100, participants indicated their estimate of how much of all the information they encountered contained inaccurate or false information (we did not distinguish between misinformation and disinformation as these concepts are difficult to distinguish for media users, also see Hameleers et al. 2023). Importantly, these perceptions were not restricted to certain domains, such as political communication or social media platforms. Hence, given that perceptions of misinformation may be highly contingent upon individual-level differences in media use and associations with the term, the question wording was kept as open as possible. Yet, we aimed to restrict the estimation to information that would classify as misinformation based on scholarly definitions (i.e. information that lacks facticity, or content i.e. based on false, manipulated, or inaccurate information).

As risk perceptions related to misinformation may be subject to third-person biases, meaning that people perceive that mostly others are vulnerable to the risks of misinformation (e.g. Jang and Kim, 2018), we additionally included a question asking people to estimate the amount of misinformation encountered by 'most people in the Netherlands'.

Results

Estimated risk perceptions in response to misinformation interventions

To test our hypotheses, we rely on one-way analysis of variances (ANOVAs) in which the conditions variable was included as categorical independent variable, and the estimated proportion of misinformation for participants themselves and others in the Netherlands included as dependent variables tapping risk perceptions. We report corrected pairwise mean score comparisons based on t-tests for our more specific expectations. Generally, we see significant differences in personal risk assessments between the conditions ($F(3431) = 2.55$, $p = .033$, partial $\eta^2 = .017$). Contrary to H1, however, all pre-bunking interventions resulted in lower risk perceptions than the absence of a pre-bunking message. Based on the Bonferroni pairwise mean score comparisons, this difference was only significant between the control condition without pre-bunking ($M = 46.50$, $SD = 20.72$) and the relativized media literacy message ($M = 38.74$, $SD = 21.01$, $p = .008$). The same patterns were found for the perceived risk of misinformation exposure

for most other people in the Netherlands (third-person perceptions), although the model was not significant ($F(3431) = 1.76$, $p = .155$, partial $\eta^2 = .012$). Thus, instead of priming suspicion and resulting in disproportionate risk perceptions, we see that the relativized variant of the pre-bunking message relativized risk perceptions related to misinformation. In the control condition, the baseline perception of misinformation was close to 50% ($M = 46.50$, $SD = 20.72$). This decreased substantially for the relativized media literacy interventions (see Table 1).

Our findings do, however, lend some support for H2: the stronger the warning and emphasized risk of misinformation in the intervention, the higher the perceived risk perceptions about the abundance of misinformation. Hence, whereas the relativized media literacy intervention resulted in significantly lower estimated misinformation prevalence compared to the control condition, this was not the case for the established media literacy intervention or the deception-focused pre-bunking message. Specifically, the stronger warning message yielded higher levels of perceived misinformation exposure ($M = 43.43$, $SD = 22.04$) than the relativized media literacy message ($M = 38.74$, $SD = 21.01$, $p = .05$). For third-person risk perceptions, H2 was only confirmed when comparing the established media literacy message currently used in the Netherlands ($M = 49.29$, $SD = 20.25$) to the relativized media literacy message ($M = 43.01$, $SD = 20.58$). In line with these findings, the weaker the prime of suspicion in the pre-bunking message, the lower the estimated third-person risk perception.

Exploratory findings explicating uncertainty and risk

To contextualize these findings, we included an open-ended question asking participants to list their thoughts related to the intervention they were exposed to. One important finding is that many participants emphasized the high degree of uncertainty in determining whether something is true or false: 'It is difficult to find out when a message is real or fake. It is unpleasant having to study every message and judge its truthfulness'. People further emphasized their worries and fears related to fake news: 'It is very concerning that there is so much fake news around us'. Another theme that stood out was the association of misinformation with social media, which was most pronounced in the condition voicing the strongest warning about misinformation: 'It makes me really sad that the rise

Table 1. Overview of mean estimates misinformation prevalence across conditions.

Experimental condition	*M*	*M* 95% CI [LL, UL]	*SD*
Control group	46.50_a	[42.63, 50.36]	20.72
Established ML	43.43_a	[39.22, 47.63]	22.04
Deception-focused pre-bunking	42.33_a	[38.42, 46.24[	20.90
Relativized ML	38.74_b	[34.61, 42.86]	21.01
$F(3, 431) = 2.55$ $p = .033$			

Note. *M* and *SD* represent mean and standard deviation, respectively. *LL* and *UL* indicate the lower and upper limits of the 95% confidence intervals (CI) for the mean scores. Letters with different subscripts indicate significant differences in pairwise mean score comparisons (Bonferroni corrected).

of social media comes with a flood of misinformation'. In the relativized condition, in contrast, many participants emphasize the need to be critical and verify information: 'You have to check whether you can trust information, for example, by checking the source'.

The thematic analysis of the open-ended answers offers support for a deception-default and high degrees of concerns related to misinformation, especially when the pre-bunking messages emphasized a strong warning: People find it difficult to discern true from false information, and believe that misinformation is flooding social media. Although the relativized warning triggered more critical beliefs toward information in general, the strong warning message was related to a more cynical outlook on social media, and triggered the idea of overwhelming amounts of misinformation online.

Discussion

In light of research suggesting that warning people about misinformation can negatively affect trust in real information (Van der Meer et al., 2023), this study aimed to explore to what extent different forms of pre-bunking would result in (dis)proportionate risk estimates related to misinformation. Can warning about misinformation's omnipresence and content characteristics cause disproportionately high risk perceptions related to misinformation exposure?

Our findings first of all indicate that, irrespective of seeing an intervention, risk perceptions about misinformation are relatively high. Overall, people perceive that about 50% of all the information they encounter can be classified as misinformation, which mirrors the findings of comparative survey research indicating that about half of the population is (very) concerned about encountering misinformation online (Newman et al., 2023). Although current estimates of the exact proportion of misinformation in people's newsfeeds may be incomplete or subject to various biases, misinformation may represent a vast minority of all information encountered (e.g. Acerbi et al., 2022; Yang et al., 2023). In light of this, we suggest that the average proportion of estimated misinformation exposure is not congruent with the actual threat of false information online.

Contrary to our expectations, our findings do not lend support for the expectation that pre-bunking always strengthens risk perceptions. If anything, the relativized and established media literacy interventions that did not offer a strong warning about misinformation lowered perceived risk perceptions. One explanation for these findings is a potential ceiling effect, and a shift from a truth-default to a deception-default in the current media ecology where the alarm on misinformation is constantly ringing (Van der Meer et al., 2023). Even without the presence of an explicit trigger event (Clare and Levine, 2019) offered by a pre-bunking intervention, people are highly aware of the potential of being deceived. Arguably, the constant warnings about 'floods' and 'uncontrolled' or 'overwhelming' streams of misinformation, for example, voiced in the context of the Israel/Palestine war, may have made the trigger of deception inevitable and omnipresent.

Our findings do suggest that risk perceptions may be lowered by exposing people to a pre-bunking message that relativizes and contextualizes the threats of misinformation by emphasizing how to navigate the media ecology for truthful and honest information. Exposure to a relativized media literacy message, at least compared to the absence of an intervention, substantially lowered risk perceptions below the 50% mark. By

contextualizing the threats of misinformation in the context of mostly accurate and honest information—and by offering suggestions on how truthful content may be recognized—the trigger of deception may be substantially lowered in line with the actually lower prevalence of mis- and disinformation in the Western European context of this study.

This can be translated to practical implications and recommendations. Given the fact that misinformation is not as easy to distinguish from true content as often assumed (e.g. Hameleers and Yekta, 2023), and the disproportionally high-risk perceptions among the public, we suggest pre-bunking interventions to avoid emphasizing the harms of misinformation based on certain fixed content characteristics, such as an emotional tone or the lack of empirical evidence mentioned. We rather suggest that interventions highlight that disinformation exists in an ecosystem of mostly accurate and honest information. Yet, in this context, media users should be critical toward information as some information may not be trustworthy. Here, it is important to explicate when information can (most likely) be trusted, instead of distrusted. Given that pre-bunking messages are not able to directly inform a verdict on (un)truthfulness, their suggestions should be formulated as tips for finding additional information and sources to double check information when in doubt. Thus, pre-bunking messages should (1) provide citizens with the agency to verify information themselves, (2) offer suggestions on how the information landscape can be navigated for trustworthy information, and (3) consolidate trust in real information by not raising high levels of suspicion.

Despite offering some concrete pointers for pre-bunking interventions, this study comes with a number of limitations. First, we assessed perceived risk perceptions of encountering misinformation instead of people's actual ability to discern between true and false information. Although we postulate that risk perceptions may be an important aspect of the epistemological crisis surrounding misinformation, future research may also explore the effects of different pre-bunking messages on truth discernment related to various forms of misinformation and reliable information. Second, we conducted the study in just one country that is relatively resilient to misinformation. Future research may explore the (dis)proportionality of risk perceptions in contexts that are more likely to be threatened by actually undermining disinformation campaigns (Humprecht et al., 2020). Finally, future research may investigate the effects of pre-bunking in more ecologically valid settings, for example, by creating a selective exposure environment.

Despite these limitations, this study indicates that although risk perceptions related to encountering misinformation are high across the board, they are not reinforced by exposing people to pre-bunking interventions. If anything, a more relativized approach to media literacy interventions that focuses on the detection of reliable and trustworthy news may contribute to more moderate risk perceptions related to misinformation.

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Michael Hameleers https://orcid.org/0000-0002-8038-5005

Supplemental material

Supplemental material for this article is available online.

References

Acerbi A, Altay S and 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.

Bennett LW and Livingston S (2018) The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication* 33(2): 122–139.

Brennen JS, Simon FM and Nielsen RK (2021) Beyond (mis) representation: Visuals in COVID-19 misinformation. *The International Journal of Press/Politics* 26(1): 277–299.

Carrasco-Farré C (2022) The fingerprints of misinformation: How deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities and Social Sciences Communications* 9(1): 1–18.

Chadwick A and 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.

Clare DD and Levine TR (2019) Documenting the truth-default: The low frequency of spontaneous unprompted veracity assessments in deception detection. *Human Communication Research* 45(3): 286–308.

Damstra A, Boomgaarden HG, Broda E, et al. (2021) What does fake look like? A review of the literature on intentional deception in the news and on social media. *Journalism Studies* 22(14): 1947–1963.

Guess A, Nagler J and Tucker J (2019) Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances* 5(1): eaau4586.

Hameleers M (2022) Separating truth from lies: Comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the US and Netherlands. *Information, Communication & Society* 25(1): 110–126.

Hameleers M (2023) 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.

Hameleers M, Humprecht E, Möller J, et al. (2023) Degrees of deception: The effects of different types of COVID-19 misinformation and the effectiveness of corrective information in crisis times. *Information, Communication and Society* 26(9): 1699–1715.

Hameleers M and van der Meer T (2023) Striking the balance between fake and real: under what conditions can media literacy messages that warn about misinformation maintain trust in accurate information? *Behaviour and Information Technology*: 1–13.

Hameleers M and Yekta N (2023) Entering an information era of parallel truths? A qualitative analysis of legitimizing and de-legitimizing truth claims in established versus alternative media outlets. *Communication Research*: 1–19. DOI: 10.1177/00936502231189685.

Humprecht E, Esser F and 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.

Jang SM and Kim JK (2018) Third person effects of fake news: Fake news regulation and media literacy interventions. *Computers in Human Behavior* 80: 295–302.

Kyriakidou M, Morani M, Cushion S, et al. (2023) Audience understandings of disinformation: Navigating news media through a prism of pragmatic skepticism. *Journalism* 24(11): 2379–2396.

Levine TR (2014) Truth-default theory (TDT) a theory of human deception and deception detection. *Journal of Language and Social Psychology* 33(4): 378–392.

Modirrousta-Galian A and Higham PA (2023) Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General* 152(9): 2411–2437.

Newman N, Fletcher R, Eddy K, et al. (2023) *Digital News Report 2023*. Reuters Institute. Retrieved from https://policycommons.net/artifacts/4164711/digital_news_report_2023/4973510/ on 11 Oct 2023. CID: 20.500.12592/3sq026.

Pennycook G, Bhargava P, Cole R, et al. (2023) Misinformation inoculations must be boosted by accuracy prompts to improve judgments of truth. Preprint available on https://osf.io/preprints/psyarxiv/5a9xq.

Roozenbeek J and Van der Linden S (2019) Fake news game confers psychological resistance against online misinformation. *Palgrave Communications* 5(1): 1–10.

Roozenbeek J, Van Der Linden S, Goldberg B, et al. (2022) Psychological inoculation improves resilience against misinformation on social media. *Science Advances* 8(34): eabo6254.

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

Tully M, Vraga EK and Bode L (2020) Designing and testing news literacy messages for social media. *Mass Communication and Society* 23(1): 22–46.

Van Aelst P, Strömbäck J, Aalberg T, et al. (2017) Political communication in a high-choice media environment: A challenge for democracy? *Annals of the International Communication Association* 41(1): 3–27.

Van der Meer TG, Hameleers M and 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* 24: 1–21.

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

Vraga EK and Bode L (2020) Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication* 37(1): 136–144.

Wagner MC and Boczkowski PJ (2019) The reception of fake news: The interpretations and practices that shape the consumption of perceived misinformation. *Digital Journalism* 7(7): 870–885.

Walter N, Cohen J, Holbert RL, et al. (2020) Fact-checking: A meta-analysis of what works and for whom. *Political Communication* 37(3): 350–375.

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

Wood T and Porter E (2019) The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior* 41: 135–163.

Yang Y, Davis T and Hindman M (2023) Visual misinformation on Facebook. *Journal of Communication* 73: 316–328.