Countering Mis- and Disinformation

A Narrative Review of Reactive Measures

Ola Svenonius, Elsa Isaksson, Johannes Lindgren

FOI-R--5263--SE

August 2022

FOI

Ola Svenonius, Elsa Isaksson, Johannes Lindgren

Countering Mis- and Disinformation

A Narrative Review of Reactive Measures

Titel	Countering Mis- and Disinformation – A Narrative Review of Reactive Measures
Title	Att möta desinformation – En litteraturöversikt över reaktiva motåtgärder
Report no	FOI-R--5263--SE
Month	Augusti
Year	2022
Pages	61
ISSN	1650-1942
Client	Myndigheten för psykologiskt försvar
Forskningsområde	Krisberedskap och civilt försvar
FoT-område	Inget FoT-område
Project no	B1041
Approved by	Malek Finn Khan
Ansvarig avdelning	Försvarsanalys

Cover: Oatawa, iStockphoto LP

Summary

This report consists of a review of recent research into reactive measures to counter mis- and disinformation, mainly, but not exclusively, on social media platforms. The target audience for this report is researchers, communicators and others who are engaged in countering misinformation, for example on social media.

53 articles have been selected for closer scrutiny based on method, relevance, date of publication, and publication language. The text reviews articles by focussing on tone, correction format, source rating, refutations by experts versus peers, and other direct countermeasures. Results show that the scientific evidence for the effectiveness of correcting false information is often inconsistent. Key conclusions that can be derived from current research are:

- Researchers recommend that communicators do not repeat misinformation unless necessary.
- Corrections on social media should emphasise the content over tone, as the effects of a correction have been found to be consistent regardless of the tone used.
- A direct rebuttal is more effective than a delayed one. If the correction appears days or even weeks later, there is a risk that the subject has accepted the claim as true.
- Correction credibility adheres to a hierarchy where self-corrections rank highest, followed by expert sources and, lastly, peers.
- Expert sources can be “borrowed” when users provide links to credible and trustworthy sources.
- The risk of a backfire effect occurring because of corrections to false and misleading information is limited.

The report further discusses the empirical validity of the results and identifies gaps for future research. These gaps include, first, the inconsistencies in research results. There are many instances where results contradict each other. Second, there is a lack of research on real-life social media behaviour. Third, there is a lack of research on how countermeasures work for practitioners. These aspects are key if the field is to develop into a more coherent literature.

Keywords: misinformation, disinformation, countermeasures, social media, continued influence effect, backfire effect.

Sammanfattning

Rapporten utgörs av en litteraturöversikt över nyligen publicerad forskning om åtgärder för att bemöta vilseledande information, huvudsakligen på sociala medier. Den riktar sig bland annat till forskare, kommunikatörer och andra som arbetar med att hantera vilseledande information, till exempel på sociala medier.

Ett urval om 53 artiklar ur vetenskapliga tidskrifter har gjorts med utgångspunkt i publiceringsdatum, metod, relevans och språk. Rapporten granskar artiklarnas resultat exempelvis med hänseende till hur rättelser eller korrigeringar bör gå till, vilken tonalitet de bör ha, vem som ska utföra rättelser, märkning av misstänkt vilseledande innehåll, med mera. De vetenskapliga beläggen är ofta svaga för hur effektiva olika metoder för rättelser på sociala medier är. Vissa generella slutsatser kan emellertid fras från nyligen publicerad forskning:

- Forskare rekommenderar att vilseledande information inte upprepas i samband med en rättelse om det inte är nödvändigt.
- Rättelser på sociala medier bör fokusera mer på budskapet än sättet som det förmedlas. Forskning visar att rättelser är lika effektiva oavsett hur de uttrycks.
- En omedelbar rättelse är mer effektiv än en som sker efter en tid. Om rättelsen sker dagar eller till och med veckor i efterhand finns det en risk att mottagaren redan har accepterat budskapet.
- Rättelser lyder under en sorts hierarki där självrättelser rankas högst, därefter rättelser av exporter och, till sist, rättelser av andra användare.
- Trovärdighet från experter kan “lånas” av användare som vill hänvisa till källor som de kan lita på.
- Risken för en motsatt effekt av en rättelse, en så kallad “backfire”, anses vara begränsad.

I rapporten diskuteras vidare resultatens validitet och luckor i forskningen för framtida studier identifieras. De luckor som diskuteras är, för det första, bristen på samstämmighet i forskningsresultaten. Det finns många exempel på motstridiga resultat i översikten. Vidare saknas det studier om verkligt beteende på sociala medier. Till sist saknas studier om hur motåtgärder faktiskt fungerar för praktiker som verkar för att möta vilseledande information. Dessa luckor behöver adresseras för att fältet ska utvecklas till ett mer sammanhållet forskningsområde.

Nyckelord: falsk och vilseledande information, desinformation, bemötande, sociala medier, continued influence effect, backfire effect.

Foreword

This report is the result of a Swedish Defence Research Agency research project on information influence and countermeasures. It aims to give the reader an improved understanding of the main results in this research field and outline future research needs. The text is meant to be read selectively. The chapters are kept short in order to increase the ease of access for all readers.

All authors shared in the responsibility for producing the core text. More specifically, Elsa Isaksson is the main author of Chapters 3, 6 and 7. Johannes Lindgren is the main author of Chapters 4, 5, 6.2, and 8. The general design and editorship, as well as the writing of Chapters 1, 2, and 9, were the responsibility of Ola Svenonius.

This report was written within the project *Mythbusting i en ding ding värld: En studie om informationskrigets mikropraktiker* (Mythbusting in a ding ding world: A study on the micropractices of the information war; MSB 2018:45), which was financed by the Swedish Civil Contingencies Agency. As of January 2022, the Swedish Psychological Defence Agency assumed responsibility for the project and is now the main recipient of this report.

Sincerely,

Ola Svenonius, Project Manager

Kista, June 16, 2022

Contents

1 Introduction

The democratic system of government, based on Enlightenment ideals, builds on the belief in man's rationality, a careful trust in the institutions governing our societies, as well as in the scientific mode of knowledge production that provides a foundation for societal progress (Peter 2017; Sigerist 1938). In what has been labelled the "post-truth era," knowledge is perceived as increasingly subjective and rationality is challenged, thus threatening the foundations of democracy itself (Lewandowsky, Ecker, and Cook 2017; Duncombe 2019). In 2021, widespread anti-vaccination sentiments in the wake of the Covid-19 pandemic expanded this threat beyond the political system to being a matter of global public health. At the time of writing, Russia's war in Ukraine highlights the importance of social media in modern warfare. Truth is indeed very much a question of world politics.

In the post-truth era, disinformation, fake news, misinformation, information influence, and information warfare have received increasing political and academic attention, especially during the last 5 years. Communication scholars have long studied how and why people choose to believe things that are not true (Frenda, Nichols, and Loftus 2011; Veil, Buehner, and Palenchar 2011). This field of research was expanded in both scope and importance in 2016, as a result of the Russian information influence campaigns during the US general election and the UK Brexit election, as well as the Cambridge Analytica scandal[1] (Akoz and Arbatli 2016; De Pryck and Gemenne 2017; Dobreva, Grinnell, and Innes 2019; McCombie, Uhlmann, and Morrison 2020; Solon and Graham-Harrison 2018). Misinterpreted facts, half-truths and outright lies are today viewed as a potent threat to society and constitute a broad field of research.

This field of research has come a long way in disentangling many of the problems associated with disinformation and related topics. We now know much about how to analyse and understand various forms of falsehoods, misconceptions and fake news (Chan et al. 2017), and we have come a long way in detecting fake content and accounts on social media networks (Figueira, Guimaraes, and Torgo 2018; Cresci et al. 2018). As we describe below, the most pressing current issue is how to understand the mechanisms and effects of exposure to online disinformation, and how to counter its adverse effects. How should authorities and fact-checkers communicate with social media users in order to prevent the spreading of harmful information, for example regarding Covid-19 vaccines? What tactics have researchers identified as being most effective, or ineffective, in these circumstances? These are questions that we sought answers to in this research review.

The motivation to produce this review stems from the fact that this research addresses pressing societal problems. Sweden is presently building new institutions

[1] The Cambridge Analytica scandal refers to a major data breach, where personal information was secretly extracted from Facebook users by means of a personality test. See Afriat et al. (2021).

and capacities to identify and analyse disinformation activities and influence campaigns, and to coordinate and support public authorities in their strategic communication (Psykförsvarsutredningen 2020). It is therefore important that up-to-date knowledge about the state of the art in disinformation research is disseminated beyond the research community. This report reviews the current state of research about countering mis- and disinformation, with special focus on reactive measures.

1.1 Aim and research question

The aim of this minor review is to provide an updated overview of recent research on how to counter mis- and disinformation, focussing on works published from January 2019 to June 2021. The following questions guided this work:

- What is the current state of scientific knowledge concerning reactive measures to counter mis- and disinformation?
- What communicative countermeasures to online mis- and disinformation are identified as the most effective ?
- What are the main research gaps?

The target audiences for this report are communicators working with social media and researchers in the fields of, in particular, mis- and disinformation, and, more generally, hybrid threats or information warfare. Journalists, NGO analysts and activists who work with fact-checking may also find this review rewarding. The text was written in English to ensure accessibility beyond Sweden's borders. Below we describe the method used to collect as well as exclude the papers considered for this review.

1.2 Scope and method

This is a narrative, state-of-the-art review in Grant and Booth's terminology (Grant and Booth 2009, 95). It focuses on a subset of the publications on countering mis- and disinformation that was selected using five criteria, which we describe below. The sources included databases available to the Swedish Defence Research Agency (hereafter, FOI). Searches were conducted using the terms "counter* misinformation" and "counter* disinformation" in titles and abstracts. Data collection was carried out in two stages. The first, in November 2020, was carried out using FOI's sources and resulted in 272 articles. The second stage was carried out in early April 2021; this stage resulted in 488 additional hits. A complementary scan was conducted in June 2021, in order to control the reliability of the previous searches. Naturally, the two stages contained multiple overlapping results. After deleting duplicates, the following criteria were used to siphon out the articles of interest:

- **Time frame**: the articles should have been published in 2019–2021. The search was ended in June 2021.
- **Empirical data**: The articles should report studies using their own empirical data, i.e., theoretical works were excluded. Meta-analyses with empirical foci, however, were included.
- **Relevance**: The articles should focus explicitly on *reactive* measures against misconceptions, or how to counter mis- and disinformation. Articles focusing on, e.g., prevention and media literacy, were not regarded in this review.[2] The article's relevance needed to be reflected in its title and/or abstract.
- **Type**: The review only includes peer-reviewed journal articles.
- **Language**: The publication should be in English.

The resulting database contained 53 articles that were selected for closer inspection and inclusion in this review. A list of all included works is available in the appendix. Zotero was used to collect and sort the articles.

In the analysis below, we cite additional articles as well. These are not as such included in sample of articles. Instead, they serve as references and are sometimes necessary because the articles included in the review builds on and goes into dialogue with previous work.

The articles were categorised inductively, according to their overarching theme and type of countermeasure. We were particularly interested in different types of countermeasures (narratives, fact-checking, corrections) as well as the specific techniques used (tags, source rating, the role of experts, etc.). In addition, the backfire effect is an interesting topic, which was given its own section in the review, because of its important but contested status.

It should be noted that this is not necessarily a complete or representative sample of all published works available. During 2020, a large number of works were published on misinformation regarding the Covid-19 pandemic; not all of those found their way into our dataset. We decided not to explicitly adapt our search term to include Covid-19-related publications, since it would require a different approach altogether. This does not imply that we ignored publications systematically, merely that the search term may not have caught all works related to, for example, public health. To prevent this, the authors agreed on the above-mentioned requirements and engaged in continuous discussion regarding the relevance criteria, from which a final selection was approved in a joint workshop.

This is not a comprehensive review, but a limited snapshot of the state of the art in mis- and disinformation research. The time frame is narrow, between 2019 and 2021, which affects the conclusions that can be drawn from the material. However,

[2] That does not means that these topics are not important. They may in fact be more important than direct countermeasures, but this review focuses on reactive measures. We discuss this in Chapter 2, below.

this snapshot presents a good view of the research field, and it is our opinion that a broader search would not have yielded significantly different results. While the review produces a good representation of the state of knowledge, it is not to be interpreted as the full or final truth. The review is narrative, which means that we discuss the results but do not systematise them, as is otherwise common in reviews, such as those in clinical research.

1.3 Disposition

The text is meant to be read selectively. The chapters are kept short in order to increase the ease of access for all readers.

Below is a brief outline of the contents:

- Chapter 2 discusses the **main concepts in the field** that are necessary to understand the rest of the review. All readers are encouraged to read at least Section 2.1. Section 2.2 outlines a general description of the field of mis- and disinformation research, including publication outlets and metrics.
- Chapter 3 introduces the review by discussing **corrections to misinformation,** in general.
- Chapter 4 briefly discusses **narratives** as countermeasures.
- Chapter 5 discusses the merits and potential problems of **fact-checking**.
- Chapter 6 proceeds through several techniques of **reactive countering, such as visual warnings, tone, and timing**.
- Chapter 7 focuses on **sources of corrections**. Experts as fact-checkers, as well as corrections by social peers, are discussed.
- Chapter 8 analyses the status of the so-called "**backfire effect**" and the recent publications on this topic.
- Chapter 9 concludes the review by discussing the overall picture and identifying **research gaps** in the literature on reactive countermeasures against mis- and disinformation.

1.4 Key takeaways

Mis- and disinformation research is difficult to grasp due to the disparate nature of the "field," its current close ties with political development, and inter-connectedness with other research areas (propaganda, marketing, intelligence, security studies). Results are not always coherent, and many publications reach different conclusions. This may or may not be seen as a problem. It does, however, send an important signal to communication practitioners to always consult at least two or preferably more studies on a given topic in order to form an informed opinion about the issues at hand.

There is agreement that simple corrections are generally not sufficient to prevent false or misleading information from being disseminated. Several types of corrections can be combined with advantage. Engaging in countering dis- and misinformation, and being fact-based and repetitive can, according to Ecker et al. (2022), lay the groundwork for a proper response to false or misleading information. Practitioners can help their audience learn to distinguish between facts and opinions.

Below follows a summary of the main results of the review, selected by topics.

Corrections, in general

Individuals' belief in incorrect information, low confidence in corrections, credibility and repetition of errors are factors that have been identified as negative in that they reinforce the "continued influence effect" (Walter and Tukachinsky (2020).

Fact-based corrections, where the reason for the correction is reported, are generally considered to be preferable to value-based corrections (see, e.g., Paynter et al. [2019]). It is more difficult to influence political, or politicised, opinions. Passive observance of corrections can be effective, argues Vraga & Tully (2020).

Narrative approaches

Stories as a method of disseminating knowledge are considered a good method in the long run. Sangalan et al. (2019) tested four different types of emotional appeal in terms of information about smoking. The authors were able to show that emotional appeal created a greater effect than that of the opposite. In the case of vaccinations, Kuru et al. (2021) do not demonstrate any specific effects of a narrative correction (as opposed to, for example, statistical facts).

Fact-check

Fact-checks generally show good results, but the effectiveness decreases over time. The fact that information is fact-checked on a site other than where it was originally published can also complicate the matter. It is best if the author also publishes the review.

"Differentiated acceptance" for information that contradicts people's perceptions – it is possible to educate to a certain extent, but it is more difficult to influence more basic opinions.

A question of trust: examination is potentially problematic if the subject concerns facts where the knowledge is either underdevelopment or where it can be traced back to different interpretations of reality.

There are different results regarding the design of fact-checks. Hameleers et al. (2020) and Walter et al. (2020) reach different conclusions when comparing text-

based and visual elements as carriers of fact-checking. Martel, Moshlen, and Rand (2021) tested different lengths of corrections but found no significant difference.

Warnings

Compared to other formats, warnings are uncertain: Is a visual warning better than a brief description of a fact? The results are somewhat scattered in the articles examined. Visual warnings have been compared with short snippets of text, humorous images and factual reviews. Other means then prove to be more effective, e.g., corrections by other users (Garrett and Poulsen 2019; Colliander 2019).

Pennycook et al. (2020) warns of the "implied truth effect," that everything that is not flagged as false could be perceived as true. Warnings used strategically can reduce the willingness to share false or misleading content on social media (Ardevol-Abreu et al. 2020).

The use of tags and warnings as a means of countering mis- and disinformation can be effective in the short term but does not "stick" with the addressee for a long period of time. Flagging suspicious content can be used in combination with other measures, such as different types of corrections, or fact-checks.

Language, humour or appeal to logical thinking

Vraga et al. (2019) tested logical and humour-based corrections of misleading information in different contexts. Both worked relatively well in contexts that were not politicised, where humour worked worse. Logical corrections work better with people who had a low acceptance of the prevailing scientific consensus. Humour was effective for those who had high acceptance of scientific knowledge. Corrective comments had a greater impact than reinforcing comments.

Tone in the address

Kim & Masullo (2020) tested different types of language. Rough or disturbing language resulted in poorer trust in the information. Bode and Vraga, in several studies, have tried to test similar effects, but have not found any significant differences. On social media, many users withdraw from correcting others, but among those who did, the tone played a minor role, according to Bode and Vraga.

The expert role

Experts can be used to give credibility to both factual reviews, individual articles, or sources. This is not always practically possible; the users themselves can also perform certain auditing tasks. Kim et al. (2019) Meer and Jin (2020) tested the difference between government, news media and social media users. The first two gave a significantly better result. Bode, Vraga, and Tully (2021) show that organisations could create credibility by engaging in e.g., expert reviews. The

same author shows how social media users can "borrow" the credibility of experts by linking and referring to these reviews. Experts may thus become instrumental in a user's self-presentation.

The backfire effect

The backfire effect can be assumed to occur in different ways (Paynter et al. (2019): as a reaction to a perceived authority; if the subject is familiar with the subject and "knows better"; as an emotional reaction, e.g., fear; or, if a statement conflicts with a person's worldview. In the past, the backfire effect has been considered to be well-documented in research, but attempts to reproduce it have failed. See, e.g., Wood and Porter (2019) and Ecker et al. (2019). Many potential sources of error make it difficult to assess how solid the evidence is.

Other factors

Can it be stigmatising to spread false or misleading information? Yes, say Altay, Hacquin, and Mercier (2020), who point out that social control between users is an effective mechanism for reducing the willingness to spread incorrect information. Repetition of false information is usually considered a mistake. However, new research by Ecker, Lewandowsky, and Chadwich (2020) finds no evidence that this is the case. The authors theorise that it may have to do with the number of repetitions.

2 The field of mis- and disinformation research

Research on mis- and disinformation includes several disciplines. The sample in this review stems mainly from psychology, media and communication studies, journalism, and political science. However, health sciences, business management, and computer science are also fields where mis- and disinformation is an important topic. In this chapter, we provide a brief overview over the field of mis- and disinformation research.

2.1 Main concepts

In psychology and communication studies, *misinformation*, *misperceptions*, or *misconceptions*, have been the key terms used to designate the research object. Increasingly, however, other terms such as *fake news* and *disinformation* have found their way into the scientific terminology. This reflects the above-mentioned shift in world politics, which led awareness of misinformation to become a top priority for political leaders by the mid-2010s and especially later, during the Covid-19 pandemic. In the wake of the "weaponisation" of information technology (Tong et al. 2020), other categories of scholars who were mainly interested in modern warfare also started discussing misinformation.

It is here that the distinction between mis- and disinformation becomes important: *misinformation* is "any piece of information that is initially processed as valid but that is subsequently retracted or corrected" (Lewandowsky et al. 2012b, 124f). *Disinformation*, according to a now popular definition provided in a report commissioned by the European Commission, can be seen as being "all forms of false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit" (DG CNECT 2018, 3). In other words, disinformation is *intentional*, whereas misinformation is a broader term signifying verifiably false information. In the new European security landscape, being able to determine whether misinformation is intentional, and who the source is, has become a major undertaking, although often impossible (Ördén and Pamment 2021). Since 2016, disinformation has become a major policy issue, as is evident from Figure 1 on page 17 below. It shows the increase in publications on both topics, using the year 2009 as index.

Other key terms used in the field are "confirmation bias" "continued influence effect," and "backfire effect." The first term, confirmation bias, refers to the common tendency to believe information that confirms an already existing opinion (Lewandowsky et al. 2012b). The second term, exposure to false information, may, in some cases, produce a so-called continued influence effect, or belief perseverance effect, which means that the corrected misinformation continues to

have an effect even after the misinformation is definitively corrected, despite the fact that the correction has been understood and remembered (Lewandowsky et al. 2020, 22; Ecker, Lewandowsky, and Chadwick 2020, 37; Lewandowsky et al. 2012). Lewandowski et al. (2012, 114) describe the previous research in this area as showing that retractions rarely have the effect of eliminating reliance on the misinformation after being exposed to it, and that it is difficult to readjust the beliefs of people previously exposed to the misinformation.

The backfire effect, the third term, refers to a type of reaction to corrections. Some people, in theory, given the continued influence effect and confirmation bias, would react negatively to being corrected. In such cases, those people not only disregard the correction, but create a more steadfast belief in the original misinformation. The correction "backfires" and produces the adverse effect. Lewandowsky et al. (2012) discuss several instances of backfire, while Nyhan and Reifler (2010), in a very influential article, also provide evidence of the phenomenon. We further discuss, in Chapter 8, the more recent literature on this topic. Below, we describe the research field(s) that focus on mis- and disinformation.

2.2 Disciplines and outlets

There is no single research field that focuses on mis- and disinformation. Research on these topics is carried out in a variety of disciplines, fields, and journals. As with the topic of strategic communication, there are ongoing attempts, such as with publication of the *HKS Misinformation Review*, to create a more unified research field, but there is still a long way to go. In this review, we highlight some issues that have sparked interest across disciplines. The question of whether a "backfire effect" exists is one of them. Still, before mis- and disinformation research can be called a field in its own right, more cohesion and common frameworks are needed.

Until fairly recently, the academic interest in mis- and disinformation outside of psychology and media & communication was minimal (Nyhan and Reifler 2010). That changed with the political developments described above. After 2016, the number of publications on mis- and disinformation has increased by 300–450 per cent, and a large number of publications today constitute a disparate field including a range of disciplines. An emerging, related research field, *strategic communication*, which deals with the organisation of purposeful communication to advance a specific mission, has also increased steadily during the same period (Hallahan et al. 2007, 9; Werder et al. 2018).[3]

[3] In the ProQuest database, in peer-reviewed sources with topic in abstract, it has increased from ca. 100 publications, in 2010, to 360, in 2020 .

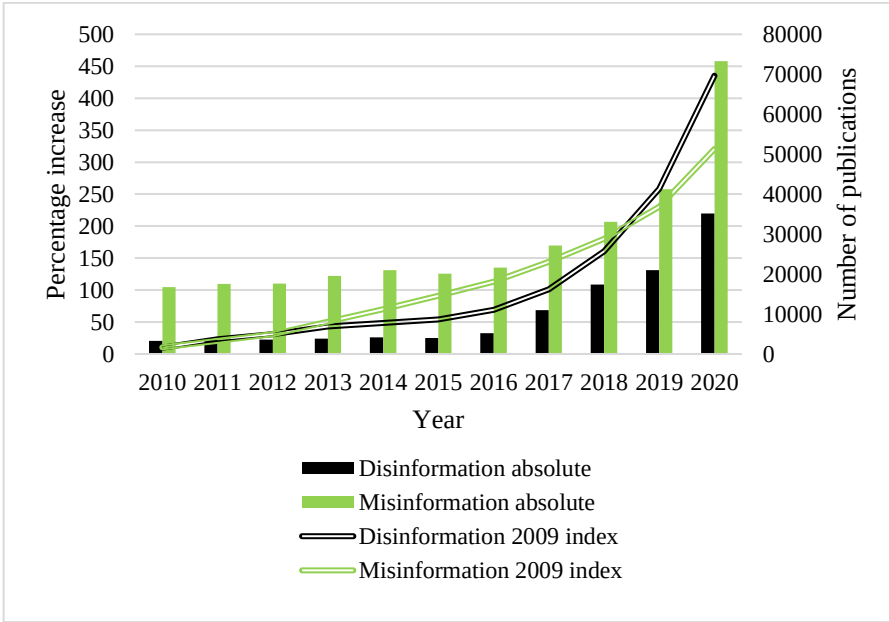

Figure 1. Development of the number of publications
Source: Lines: relative (index=2009); bars: absolute number. Searches on ProQuest, June 1, 2021, in peer-reviewed sources with topic (mis-/disinfo) in abstract. It should be noted that misinformation counts for, on average during the period, 4.8 times as many publications as disinformation, in absolute numbers.

Looking at the journals that published the works on mis- and disinformation included in this review, it is clear that a few disciplines dominate the research field. Table 1 shows the disciplinary home for the journals represented in the selection. It shows that media and communications studies, and related fields such as journalism, comprise the majority of the publications. Psychology is also well represented, especially considering that some of these publications also fall in the "interdisciplinary" category. Political science and sociology journals constitute the fourth-largest category in the sample.

Table 1. Articles in selection

Number of articles	Primary area	Example journal
18	Media and communication sciences	Communication Research
12	Psychology	British Journal of Psychology
6	Interdisciplinary	PLoS One
5	Political science and sociology	Political Behavior
4	Journalism	Journalism Practice
3	Health sciences	American Journal of Public Health
2	Information sciences	Internet Policy Review
2	Business management	Management Science
1	Computer science	Computers in Human Behavior
Total: 53		

Mis- and disinformation research is difficult to grasp due to the disparate nature of the "field," its current close ties with political development, and interconnectedness with other research areas (propaganda, marketing, intelligence, security studies). As we show below, the present topic is not easy to study. Results are not coherent, and many publications reach very different conclusions. This may or may not be seen as a problem. It does, however, send an important signal to communication practitioners to always consult at least two or preferably more studies on a given topic in order to form an informed opinion about the issues at hand.

3 Correcting false information

Using corrections is defined by Vraga and Bode (2020, 278), influential authors on this topic, as "the presentation of information designed to rebut an inaccurate claim or a misperception." This is broader than fact-checking and can be carried out through a range of techniques, such as inserting a general warning about the content of a message, or a misinformation meter. Throughout the field of research devoted to mis- and disinformation, the scientific evidence for the effectiveness of correcting false information is inconsistent. While some prior research found that providing correct information as a response to misinformation is very effective (Hameleers 2020; Vraga et al. 2020), other studies indicate that corrections do not eliminate the belief in dis- and misinformation (Nyhan and Reifler 2010; Rich and Zaragoza 2020; Walter and Tukachinsky 2020), and that a correction may even increase the belief in misinformation (Lewandowsky et al. 2012; Swire-Thompson, DeGutis, and Lazer 2020). Furthermore, a present limitation of these studies is that they do not easily permit generalisation, due to variations in study design; researchers sometimes use high-quality data sets but are more often limited to small samples of paid respondents. This complicates the relative weighing of existing evidence.

Although corrections generally reduce belief in misinformation, there is also evidence of the continued influence effect. In a meta-analysis, where the result of 32 studies on corrections to misinformation were aggregated, Walter and Tukachinsky (2020) found that misinformation continues to influence individuals' beliefs to a certain degree even after exposure to corrections. Individuals' faith in misinformation, together with low trust in the correction, credibility, and repetition of the misinformation, were factors that caused the greatest continued influence. Interestingly, while the result shows that corrections are not effective in entirely eliminating the effect of misinformation, studies in which corrections were delivered by the source of the misinformation, and consistent with the individual's worldview, did lower the likelihood of continued influence. This indicates that effects of misinformation, although not entirely eliminated, can still be reduced.

This conclusion is consistent with another meta-analysis (outside of the selected 53 articles included in the review) by Walter and Murphy (2018). Here, results from 65 studies were analysed and compiled. The meta-analysis was conducted on studies regarding attempts to correct misinformation and the authors conclude that a corrective message can influence the belief in misinformation. Across all studies, the effect of corrections was either moderate, positive, or significant. However, outcomes vary between different topics: politics or marketing are more difficult to influence than for example health issues. The same results can be found in other studies as well. For example, van der Linden, Leiserowitz, and Maibach (2019) demonstrate that corrections regarding the scientific consensus about climate change decreased belief in misinformation. Moreover, the study showed that

corrections also led to behavioural changes, such that participants demonstrated increased support for political action in the area. Similarly, Bode, Vraga and Tully (2021) found that corrective information about the (in)effectiveness of hot baths to prevent contraction of Covid-19 reduced belief in the misinformation. It also affected participants' behaviour related to the issue, also in the midterm time perspective (>1 week).

However, the results are inconsistent and other research, such as Rich and Zaragoza (2020), who did not see a durable effect of corrections. Their study investigated the interplay between the efficiency of corrections and the passage of time. The authors show that although participants' belief in misinformation was reduced immediately after the correction, this was not durable, as the belief in the misinformation increased over time. Thus, the study indicates that correcting misinformation is not durable.

Furthermore, research has found that there are various forms of correcting misinformation (Kim and Chen Masullo 2020) and a number of factors that could affect a correction to become more or less effective. Credibility, for example, has proven to influence how individuals perceive corrections (Kim and Chen Masullo 2020). While some studies focus on the credibility of the content (Tully, Bode, and Vraga 2020), others focus on the source credibility. In several studies, the use of scientific evidence and expert sources has shown to be an effective form of correction (Vraga and Bode 2017; Paynter et al. 2019; Ecker and Antonio 2020; Meer and Jin 2020). Citing credible, verified information that includes links to expert sources can also be effective when other internet users offer corrections (Vraga and Bode 2020). For example, van der Meer and Jin (2020) show that exposure to a correction can decrease believing in misinformation. Also, their results suggest that when government agencies and news media are sources of the correction, they are more effective than other social media users. The study indicates that corrective information that uses factual elaboration can affect an individual's behaviour by showing that participants' intentions to take protective actions regarding public health were affected. Similarly, by presenting participants with information that a specific autism treatment is ineffective, why that is, and why people would want to spread disinformation about it, Paynter et al. (2019) show that information regarding the importance of different types of evidence can be effective in combating mis- and disinformation. In their study, participants exposed to the corrective information about the autism treatment were prevented from initially believing the misinformation (Paynter et al. 2019). Although several studies indicate that correction does have great impact on an individual's belief in misinformation, the result of van der Meer and Jin (2020) indicates that these efforts appear especially successful in health misinformation rather than on more polarised issues.

4 Narratives as countermeasures

The power of using narratives as a strategy of persuasion and as a strategy to change attitudes and behaviours is something that has been shown to be effective (Wang and Huang 2020). This could partly be explained by the fact that narratives are not perceived as a persuasion, but rather something that would entertain the consumer and thus may limit the desire and ability to scrutinise the message (Dal Cin, Zanna, and Fong 2004). Mills and Robson (2020), for example, argue in their paper that the use of storytelling should be the way to respond to misinformation due to its emotional engagement. This could stand in contrast to non-narrative informational messages, for example public communication announcements, in which the inherit attempt of persuasion is something we as consumers tend to be aware of to a higher degree (Wang and Huang 2020). Narrative approaches have been tested in prior research in the context of various health communication settings. Wang and Huang (2020), for example, shows that narratives have the potential of triggering fewer defensive reactions than informational messages. The use of narratives as a strategy is also something currently being discussed in the context of brand management and strategic communication (Eberle and Daniel 2019; Mills and Robson 2020; Mohamad 2020; Winkler and Etter 2018).

Sangalan et al. (2019) tested narrative-based corrections containing four discrete emotions (sadness, fear, anger and happiness) in the context of organic tobacco. The findings suggest that the narrative corrections were effective in reducing all misinformation outcomes relative to a no-correction control condition. Also, narrative corrections with emotional endings compared to corrections using no emotions were even more effective in adjusting misinformed attitudes (Sangalang, Ophir, and Cappella 2019). These results are also in line with another recent empirical paper where the authors tested both narrative corrections and text rebuttals with the purpose of countering misleading pro-tobacco YouTube videos (Ophir et al. 2020). In this study, both narrative corrections and text rebuttals prove to be effective in reducing misinformed beliefs and attitudes. However, contrary to the authors' expectations, the narrative corrections did not prove to be more effective than the textual rebuttal, except for those participants who strongly identified with the character used in the correction (2020, 4973, 4982f).

As with other types of countermeasures discussed in this review, results are not unequivocal. Wang and Huang (2020) tested whether narratives would be a helpful way to counter misinformation related to the use of e-cigarettes. The results did not support the notion that narratives would be effective in countering misinformation; they did not reduce participants' counterarguments. Similar results were presented in a study testing narrative vs. non-narrative types of corrections of misinformation, with the conclusion that it did not matter which of the two was used as long as the correction was easy to comprehend and contained useful, relevant and credible information (Ecker, Butler, and Hamby 2020). In a recent

study, no significant effects of using narratives as part of a correction containing pro-science messages regarding vaccination were reported (Kuru et al. 2021). This stood in contrast to the use of statistical messages informing about the vaccination, which showed far better effectiveness (Kuru et al. 2021, 13). Narrative countering strategies, especially if they are carried out in a long-term perspective, thus find mixed support in the works included in this review, but not without exceptions.

5 The role of fact-checking

Fact-checking is "the practice of systematically publishing assessments of the validity of claims made by public officials and institutions with an explicit attempt to identify whether a claim is factual" (Walter et al. 2020, 351). Fact-checking is becoming an increasingly important tool in the media landscape of today's democratic societies (Ecker et al. 2019). More and more actors, such as social media companies, journalists and actors within civil society, now use fact-checking in a variety of contexts (Brandtzaeg, Følstad, and Chaparro Domínguez 2018; Ardèvol-Abreu et al. 2020). One example is Reuters' fact-checking as a part of their reporting on the war in Ukraine.[4] However, the effectiveness of fact-checking as a method of combatting mis- and disinformation is not always clear. Some researchers question its impact in relation to other methods (Lazer et al. 2018; Shao et al. 2018; Ecker et al. 2019). Limitations to the effectiveness of fact-checking could be cognitive dissonance that may occur when individuals are presented with factual information that does not match easily with their previous conceptions (Lazer et al. 2018). Another limitation could also be trust – in some countries fact-checks are not regarded as neutral or non-partisan (Ardèvol-Abreu et al. 2020), which affects the level of confidence and thus the effect of fact-checks (Ardèvol-Abreu et al. 2020; Brandtzaeg, Følstad, and Chaparro Domínguez 2018). In addition, some commentators have called for the need to fact-check fact-checkers themselves in order to assure legitimacy in these organisations (Brandtzaeg, Følstad, and Chaparro Domínguez 2018, 1114).

Despite some caveats regarding the reach of and confidence to fact-checks, the majority of the recent empirical research studied within the frame of this literature review indicates that fact-checking seems to be a relatively effective way to counter mis- and disinformation (Hameleers et al. 2020; Walter et al. 2020; Ecker et al. 2019; Nyhan et al. 2020). As an example, Walter et al. (2020) reviewed 30 studies to test the effect of fact-checking in correcting political misinformation. In the review, the authors conclude that fact-checking can positively affect beliefs regardless of context, pre-existing beliefs, or political ideology, even after exposure to a single fact-checking message. The authors also conclude that fact-checking is effective regardless of whether the refutation concerns an entire statement or just part of a statement (Walter et al. 2020). Similar results were presented in a study conducted by Nyhan et al. (2020), who tested exposure to fact-checks in the context of claims made by former US president Donald Trump. The study's results show that people expressed more factual beliefs after being exposed to a fact-check, even among Donald Trump's supporters, but that this did not affect their affinity for Trump as a presidential candidate (Nyhan et al. 2020). Furthermore, in studies by Ecker et al. (2019) and Hameleers et al. (2020), the

[4] See Reuters Fact Check Headlines: https://www.reuters.com/news/archive/factCheckNew (last accessed March 28, 2022).

results indicate that using fact-checkers to counter false information seems to be effective. Also, in a recent study by Walter and Salovich (2021), exposure to counter-attitudinal fact-checking was effective in rooting out previous misinformation. However, since the information that was debunked contained both opinion-based claims and factually-based claims, the participants had a hard time knowing whether political statements contained fact-based claims, which resulted in the fact-check's lack of effectiveness when participants were not aware of which statements could be factually verified (Walter and Salovich 2021). In another study testing the reactions of users when exposed to Facebook's labels for fact-checking, when analysing the quality and content of the fact-checkers, the labels were not perceived as determinant (Ardèvol-Abreu et al. 2020). In addition, several participants expressed distrust in the fact-checkers used by Facebook as well as in fact-checking as a process in itself, framing it as "the ministry of truth" (Ardèvol-Abreu et al. 2020, 9).

There are multiple ways to fact-check a specific statement and previous research states that the modality of the fact-check, for example using such visuals as "truth meters," could have an effect on the results (Hameleers et al. 2020, 297; Amazeen, Vargo, and Hopp 2019, 28). Hameleers et al. (2020) tested whether this was the case using both visual and textual fact-checkers. The result shows that the modality did not matter for its effectiveness. Also, the review by Walter et al. (2020) investigated whether visual elements would be beneficial in the fact-checks and came to the conclusion that including graphical elements was on the contrary less effective in affecting beliefs than non-visual fact-checkers (Walter et al. 2020).

In sum, fact-checking seem to be a fairly good way to tackle false and misleading information. However, it is a concern that some people view fact-checking and fact-checkers as biased, which indicates that the fact-checking organisation itself may prove to be of significant importance if fact-checking is to be an effective tool in countering mis- and disinformation. Future reviews could do well to focus specifically on empirical papers investigating fact-checking actors to see whether some actors reach a higher degree of effectiveness compared to others.

6 Correction techniques

6.1 Tags and warnings

A prominent way of correcting false information is to add tags or warnings regarding potentially false or misleading information. These types of measures take different forms, such as tags or banners indicating a specific content as "disputed" or "false," or browser plugins that warn users when they enter a page where the information provided on the page is suspected to be false (Clayton et al. 2020; Garrett and Poulsen 2019).

Multiple actors, such as Facebook and Google, have now taken steps to use warnings, whereas Facebook started to add "disputed" tags to stories in their newsfeed, in December 2016 (Clayton et al. 2020, 1074; Garrett and Poulsen 2019, 240). Previous research has suggested that adding these types of warnings or tags could be an effective strategy (Clayton et al. 2020, 1073). For example, Bolsen and Druckman (2015) find that warnings could be more effective than corrections when countering motivated reasoning about scientific claims (Bolsen and Druckman 2015). Furthermore, Ecker et al. (2010) showed that an issue-specific warning showed a large reduction in the continued influence effect of a misinformation claim (Ecker, Lewandowsky, and Tang 2010).

A risk in using warnings as countermeasures is the so-called "implied truth effect," which means that headlines that do *not* get tagged are considered valid and thus more accurate than if the tags on the other headlines did not exist. This was the case in a study carried out by Pennycook et al. (2020), where the presence of warnings caused untagged headlines to be perceived as more accurate. However, when the authors tried to attach verifications to true headlines, the implied truth effect was slightly reversed, meaning that people who were exposed to *both* warnings and verifications became less likely to consider sharing headlines that had no tag at all (Pennycook et al. 2020, 4956).

Regarding recent literature on the effectiveness of warnings, the results seem to vary. In a review of 26 articles, which among other things tested factors that contributed to individuals' belief in fake news, the authors conclude that empirical research is inconclusive with regard to warnings, different types of labelling and flagging (Bryanov and Vziatysheva 2021). For practical use, the authors argue that if these types of measures are used, they should be used cautiously. One of the studies examined in their review tested the effectiveness of flagging false political posts on social media. This was done in three different ways: using fact-checker flags, peer-generated flags, and self-identified humour flags (Garrett and Poulsen 2019). The results indicate that only the self-identified humour flags generated an effect with regard to the participants' engagement with the false post and their intentions of sharing it (Garrett and Poulsen 2019). Furthermore, in an article

investigating how social media users apply false content warnings on social media, warnings and labels issued by Facebook did not have a big impact on participants' perceptions about the quality and credibility of potentially false content (Ardévol-Abreu, Delponti, and Rodriguez-Wanguemert 2020). However, the authors point out that they did not measure the behavioural impact of the warnings, which leaves out the possibility that the warnings could have had some influence: a majority of the participants indicated that they would not share the false social media posts to their social media contacts (Ardévol-Abreu, Delponti, and Rodriguez-Wanguemert 2020).

Some studies, however, show an effect in using tags or warnings to counter false content online. Clayton et al. (2020) investigated whether different types of tags and general warnings would be an effective strategy to lower the credibility of a false headline. Results indicate that both the tags "Disputed" and "Rated false" modestly reduced belief in the false information, but that the "rated false" tags were more effective (Clayton et al. 2020). The effects of general warnings were small in comparison to the tags tested and could potentially induce a spillover-effect, since they also reduced the perceived accuracy of true headlines. However, when comparing tags to other forms of countermeasures, such as short-format corrections and critical comments from users, the apparent effectiveness of tags almost disappears (Colliander 2019; Ecker et al. 2019). For example, in a study comparing the effectiveness of the use of a "disputed" tag with critical comments by other users, the results indicate that critical comments were more effective in stopping the spread of false information on social media (Colliander 2019). In another study, researchers compared the effectiveness of using a "false" tag with short-format refutations (140-characters). The study showed that the impact after one day was similar. After a week, however, the short-format refutation yielded a stronger effect in reducing belief in false claims than the tag did (Ecker et al. 2019).

In sum, previous research indicates that the use of tags and warnings as a means of countering mis- and disinformation can be effective in the short term but does not "stick" with the addressee for a long period of time. Flagging suspicious content can be used in combination with other measures, such as different types of corrections, or fact-checks.

6.2 Source rating

Prior research on countering dis- and misinformation has shown that when presented with information that challenges someone's view on a matter, the source of the information is important in the validation of that information (Kim, Moravec, and Dennis 2019; Ecker et al. 2019). Kim, Moravec, and Dennis (2019) argue that fact-checking is most effective when presented to the user simultaneously with the article. However, most fact-checking takes place on different platforms than those where the original source was first published, sometimes even days after the misinformation or disinformation was first

published. Although automated solutions, such as fake fact-checking sites that verify news articles, can be faster than manual solutions, fact-checking takes place after much of the consumption of the news story and thereby the assumed damage has already been done (Kim, Moravec, and Dennis 2019). An alternative solution, or a complement, to fact-checking, they argue, is source-rating applied to articles on social media.

A difference between source-ratings and fact-checking is that source-rating is attached to every article when it is first published, rather than having verification attached to articles days after they have been published, sometimes even on a different platform than the one where the article was original published. The result of Kim, Moravec, and Dennis's (2019) study indicates that not only do the ratings seem to have the desirable effect of alarming users against the negatively rated sources, but they also have the second-order effect of stimulating users to think more critically about the truthfulness of the articles they see, without ratings present. In two studies, they tested the effects of three different types of rating mechanisms: expert ratings, user article ratings and user source ratings. The result showed that source-rating from experts was more effective than the others. However, the authors argue that since there are more ordinary users than experts available to rate articles, developing source ratings from other peers may be easier than finding appropriate experts (Kim, Moravec, and Dennis 2019).

6.3 Indirect effects of corrections

Corrections on social media can also be aimed at a wider audience, besides being directed towards the user's receiving and potentially sharing the misinformation. That is, this involves users who do not directly engage in the interaction but still witness the correction due to their presence on social media. Researchers had previously shown that passive corrections, referred to as "observational correction," could be an effective method from witch misperceptions can be reduced (Vraga and Bode 2017). In Bode, Vraga, and Tully's (2020) study, participants who saw someone else get corrected on social media experienced reduced belief in the misinformation. The observational correction is thus aimed at both the user who shares the misinformation as well as the wider audience who encounters a specific account. The approach emphasises the potential social cost of sharing misinformation and being publicly corrected, which may make users more unlikely to engage in such spreading behaviours (Altay, Hacquin, and Mercier 2020). Since the observational correction occurs simultaneously with users being exposed to the misinformation post, Vraga, Tully and Bode (2021) suggest that the likelihood that the misinformation reinforces misperceptions may be lower than for corrections to mis- and disinformation appearing days or weeks after. The result of their study indicates that observational corrections that point out scientific consensus can lead to more accurate beliefs on specific issues, as well as changes in behaviour.

6.4 Repeating misinformation

Another factor of corrections that has been studied is whether misinformation repetition in a correction message could affect the effectiveness. A study by Ecker, Lewandowsky, and Chadwich (2020) indicated that corrections that repeated misinformation did not lead to stronger misperceptions. Similarly, the result of Ecker, Butler, and Hamby (2020) indicates that exposure to corrections that repeat non-novel misinformation will not be counterproductive. Moreover, as for Ecker, Lewandowsky, and Chadwich (2020), the study shows that repeating misinformation while correcting it may even have some positive effects: participants in one of the experiments decreased their belief in the misinformation after the repeated correction. However, the authors point out that these benefits were seen after only one repetition of misinformation and may not occur when there is additional repetition of the misinformation. Therefore, they do not recommend that communicators repeat misinformation unless necessary.

6.5 Content and tone

In addition to the role that scientific evidence and expert sources play, other characteristics of corrections have also been studied. One is whether a correction consisting of logic and facts or a correction using humour would have greater outcomes in reducing individuals' belief in misperceptions and misinformation. Vraga, Tully, and Bode (2021) tested two different types of corrections, logic-based and humour-based, in a study using misinformation from three different topics, climate change, gun control and HPV vaccination. Both types of corrections affected issue attitude and credibility perception when applied to misinformation on the topic of HPV vaccination but not the other two topics. Moreover, the logic-based corrections were more effective among participants that had low agreement with the scientific consensus on the issue, while the humour-based approach was more effective for participants with high agreement with the scientific consensus.

In order to investigate how the efficacy of the corrections is dependent on the correction style, Martel, Moshlen, and Rand (2021) manipulated the correction message to either be more or less detailed. The study suggests that these manipulations have minimal, if any, effects on social media users' likelihood of replying to or accepting a correction.

The effect of the tone of peer corrections, that is, whether users oppose one another politely or rudely, was studied by Kim and Chen Masullo (2020). They looked at online comments to a news story and how the tone affected the credibility of the content. Are uncivil comments less credible than polite ones? Might they even reinforce the misinformation? The researchers found that participants perceived polite comments to be more credible than the uncivil ones. Also, participants perceived the comments that corrected the misinformation as more credible than

the ones reinforcing it. However, rude reinforcement of a news story was not necessarily more effective than a polite one. This leads the authors to conclude that people seem to consider these two factors separately when evaluating the credibility of the corrections. Bode, Vraga, and Tully (2020), too, suggest that neither neutral, civil, uncivil, nor affirmative corrections affect the effectiveness of corrections. In the study, 610 participants were first shown a meme that contained misinformation. Then the respondents were shown a correction; although the respondents were shown corrections of different types of tone and content. The corrections varied with respect to civility, affirmation, or topic neutrality. The facts of the correction were the same regardless of tone. Although there was no effect of the tone of the correction, all corrections reduced belief in the misinformation. Similarly, Tully, Bode, and Vraga (2020), in an experiment examining the willingness of other users to reply and engage in corrections, manipulated the tone of the corrections used. The different tones used in the corrections were either neutral, affirmative, or uncivil, but the facts were all the same. Moreover, the researchers also examined whether the tone of a correction affects how other users reply to the misinformation. Among those who engaged in corrections, the tone of the corrections had little effects on how they responded. For example, the correction that used an uncivil tone did not lead participants to respond uncivilly.

The studies reviewed here thus show that corrections on social media should emphasise the content over tone, as the effects of a correction were consistent regardless of the tone used.

6.6 Timing

The timing of a countermeasure, that is, whether it occurs before, during, or after, having received the misinformation, likely affects its outcome. According to Rich and Zaragoza (2020), the design of the experiments in much previous research on mis- and disinformation is such that the observations take place immediately after participants are exposed to the misinformation (Rich and Zaragoza 2020). This means that measurements of the impact of corrections, rebuttals, and fact-checks may be biased by the close proximity in time. In real life, the corrections and fact-checks often occur days, weeks, or sometimes months after the misinformation was first posted. Rich and Zaragoza (2020) addressed this possibility in a recent study.

Other researchers, such as Brashier et al. (2021), indicate that providing fact-checks after headlines is more effective than presenting them during or before exposure to misinformation online.[5] In the study by Brashier et al., participants were presented with “true” and “false” headlines that appeared before, during, and

[5] Ecker et al. (2020), whose results we discussed above in the context of repetitions of misinformation while debunking it, follow the same reasoning.

after the participants read an article. They were then asked to reclassify the articles one week later. The respondents who received “true” or “false” tags that appeared immediately after exposure to (mis)information were most successful in correctly identifying false information one week later. The respondents who were shown the corrections before or parallel to the misinformation exposure were significantly worse at re-identifying the false information one week after. Similarly, a two-wave online experiment from Dai, Yu, and Shen (2021) that tested how the effect of corrections depended on timing, showed that corrections provided after the misinformation, compared to before, were more effective.

As with several other topics covered in this review, the research is not entirely conclusive. Despite Vraga and Bode (2020) arguement that corrections should be made as soon as possible before the misperceptions are entrenched, and the fact that fact-checks and corrections seem to be more effective if presented after the misinformation (as discussed above), one study shows a different result. In examining how the efficacy of a correction of misinformation interacted with the passage of time, Rich and Zaragoza (2020) found no evidence that the timing of the correction appearing after the misinformation actually impacted the efficacy of the correction. This was tested by assessing the effects of a two-day delay, compared to a correction applied minutes after participants were exposed to the misinformation.

In sum, a direct rebuttal is more effective than a delayed one. If the correction appears days or even weeks later, there is a risk, if the subject has accepted the claim as true, that the continued influence effect occurs. Rich and Zaragosa’s study is interesting, because they show that it does not matter when the correction is made; in their study, the effect does not hold over time. Only days after receiving the correction, in time, that is, and with it in fresh memory, subjects had still increased their belief in the misinformation with time. Therefore, especially with news relating to polarised issues, the effect of debunking, or corrections, should not be overstated.

7 Who should counter mis- and disinformation?

Expert sources, including health agencies and governmental organisations, news media, and other internet users have the potential to correct misperceptions among the public. However, although previous work indicates that the source of the correction matters when correcting misinformation, studies show that this is not the case for every strategy in combating it. As for fact-checks, for example, the study of Wintersieck, Fridkin, and Kenney (2021) indicates that the content of the information presented is more significant for individuals' perceptions than its source. Below, we first discuss expert sources, then the use of social peers.

7.1 Expert sources

Several studies show that corrections of mis- and disinformation are more effective if they come from an expert source (Vraga and Bode 2017, 20; Ecker and Antonio 2020; Meer and Jin 2020). By comparing whether an expert's, as opposed to a normal user's, sharing of WHO graphics designed to address Covid-19 misinformation could reduce misperceptions, Vraga and Bode (2021) show that exposure to the WHO graphic reduced immediate misperceptions about the science of a false preventative for the illness. Similarly, the result of Bowles et al.'s (2020) study in Zimbabwe indicates that social media messaging from trusted sources may play a large role not only for individuals' knowledge but, ultimately, their behaviour. Furthermore, corrective information from an expert source, as opposed to non-expert, about health misinformation may increase the likelihood that individuals find correct and accurate information instead of relying on misinformation (Vraga et al., 2020). Vraga and Bode's (2017) study shows that a single message by a reputable scientific expert, correcting the information about the causes of the Zika virus, reduced misperceptions about the virus on social media.

Apart from expert agencies and organisations, other authoritative sources have been shown to have a substantial impact on reducing misperceptions when they have engaged in corrections. Among other things, van der Meer and Jin (2020) investigated the impact of different sources (government health agency, news media, or social peer) of corrective information on health misinformation and found that governmental agencies, as well as news media, compared to social peers on social media, were more likely to be successful in correcting and debunking misperceptions among the public. Moreover, the result shows that individuals tend to experience more anxiety in response to a public health crisis when corrective information comes from government agencies or news media, compared to when it comes from social peers. Fear and anxiety tend to increase preventive actions taken by individuals (Meer and Jin 2020).

Recent studies on using expert sources to correct misinformation online have also identified opportunities for organisations to improve their credibility. Vraga and Bode (2017) indicated that expert organisations would not lose credibility when correcting misinformation, and in a more recent article, the authors develop their account. They show that an organisation's credibility may even increase after it engages in corrections (Bode, Vraga, and Tully 2021). By examining whether expert organisations can correct misinformation on social media, Bode, Vraga and Tully (2021) find that misperceptions among the public about genetically manipulated food were reduced and that the credibility improved after expert organisations highlighted the scientific consensus on the matter. The result of this study indicates that providing corrective information about a scientific misperception may be a good strategy for expert organisations, as it not only provides citizens with correct information but also improves the organisation's credibility among the public. According to the study's results, expert organisations should consider providing the public with corrective information as well as emphasise the scientific and expert agreement on the matter.

Source credibility has been shown to be effective and crucial for correcting information (Walter and Tukachinsky, 2020). Although perceived expertise has proven to be a successful ingredient in successful corrections, Ecker and Antonio (2020) previously demonstrated that trust may matter even more than perceived expertise. They tested whether perceived trustworthiness of the source of a retraction determines its effectiveness and what role the perceived expertise of the source plays. The findings indicate that perceived trustworthiness of the retraction source matters. Moreover, in one of two experiments, they found that retractions from expert sources were ineffective if the source trustworthiness was low, which indicates that trustworthiness is a crucial factor in source credibility. These findings further indicate that corrections from expert sources are not *per se* effective since rebuttal messages from non-expert sources may have greater impact if the level of trustworthiness is higher.

Misleading corrections that provide *more* misinformation have been documented (Vraga and Bode 2020). These may be non-intentional results of a lack of knowledge of how to accurately correct misinformation. Moreover, it not only matters about the source credibility of a correction, but also that the credibility of the source that shared the misinformation in the first place can affect how effective a correction is. Some research indicates that if the source of the misinformation is perceived as more credible than the source that aims to correct it, continued influence can occur despite subsequent retractions and corrections (Connor Desai, Pilditch, and Madsen 2020). Walter and Tukachinky (2020) suggest that if the credibility of the source of the misinformation trumps the source of the correction, the credibility of the retraction source may have minimal impact on the size of the continued influence effect. Walter and Tukachinky's (2020) results also show that corrections coming from the same source as the misinformation will be more credible than a correction coming from a different source.

In short, the output from recent studies suggests that correction credibility adheres to a hierarchy where self-corrections rank highest, followed by expert sources and, lastly, peers; but also that this hierarchy can be offset if the trustworthiness of a misinformation source is deemed to be higher. In such a case, the backfire effect may occur, as discussed separately below. The hierarchy may be fragile, however, as peer corrections may also trump expert sources. This is the topic of the following section.

7.2 Corrections by peers

Mis- and disinformation on social media platforms can have a widespread effect when users engage with it by reposting, commenting, and replying. But social media users can also respond with corrective information. Whether non-expert sources such as other social media users who engage in combating mis- and disinformation online are effective or not has been investigated in several studies. As stated above, experts seem to be more effective in correcting misinformation than other users, but it also seems that engaging in making such corrections on social media can improve their organisational credibility (Vraga and Bode 2020). For example, in addition to showing that a single correction from the Center for Disease Control and Prevention (CDC) was effective in reducing misinformation about the Zika virus, the result from Vraga and Bode's (2017) study indicates that if there is only one user correction it does not produce the same response. In the study, a single user was not able to reduce misperceptions on their own, nor could a user who added their rebuttal to an existing CDC correction further contribute to its effectiveness beyond the CDC.

For peers, citing highly credible information with links to expert sources can be an effective response to misinformation. Vraga and Bode (2017) show that when other users, rather than just correcting the misinformation, also provided a reliable source for it, they were effective in reducing misperceptions about the Zika virus, both on Facebook and Twitter. Expert sources can be "borrowed" when users provide links to credible and trustworthy sources (Bode and Vraga 2021). This research, indicating that social media users can play an important role by responding to online misinformation with a link to accurate information from expert sources (Bode and Vraga 2018), is promising, considering the number of social media users compared to the number of professional fact-checkers and expert organisations online (Tully, Bode, and Vraga 2020). Moreover, the results of Vraga et al. (2020) suggest that a user who debunks a myth pre-emptively by using facts might be less effective than when sharing a correction made by an expert source.

A person's real-life network is an important factor that can affect their ability to resist mis- and disinformation. In a recent article, Ecker and Antonio (2020) discuss the possibility that trust might have an even greater impact on corrections than expertise does. Close peers on social media, such as friends and family, might

therefore be well suited to engage in correcting and debunking misinformation online (Bode and Vraga 2021). Colliander's (2019) study of fake news concludes that users exposed to comments by critical peers responded with lower trust in the misinformation and were more likely to produce critical comments themselves. They were also less likely to share the misinformation than other users were.

Social corrections, where other internet users rather than experts, news media, or algorithms correct the online mis- and disinformation, can function as a parallel support system for users at risk of "falling down the rabbit hole." By comparing a correction produced by a platform algorithm with a social correction from another internet user, Vraga and Bode (2018) found that social corrections were as effective in reducing the acceptance of misinformation as the algorithm.

While individuals, in a study by Tandoc et al. (2020), expressed concerns about engaging in making corrections, other studies suggest that many individuals do hold beliefs that are the public's responsibility to respond to and correct misinformation about online (Bode and Vraga 2021). By examining how exposure to misinformation, and the associated correction, on Twitter affected the likelihood that users would respond to the misinformation, Tully, Bode, and Vraga's (2020) study indicates that individuals are unlikely, overall, to respond to misinformation tweets but more likely to respond with correct information after seeing other corrections. Little research has been done on what drives these individual users to conduct social correction (Sun et al. 2021). Tandoc et al. (2020) indicates that believing that the misinformation will be harmful to oneself or others, as well as being emotionally attached to the topic, can be driving forces behind the activity. This result is confirmed in Sun et al.'s (2020) study, in which they found that people take the influence of misinformation on others into consideration when engaging in making social corrections. Similarly, by examining how exposure to active corrections of Covid-19 misinformation by other social media users induced participants' threat appraisals of the influence of the misinformation, the results of Sun et al.'s (2021) study suggest that the anticipation of guilt strengthened users' intentions to correct misinformation related to the illness. According to the authors, being aware of the potential influence that the misinformation can have on other people engendered guilt, which strengthened participants' intentions to engage in social correction of the misinformation. These results offer some guidance on how to engage social media users in social corrections, that, despite mixed results, can be a successful mechanism in correcting mis- and disinformation online.

8 The backfire effect

Efforts to counter mis- and disinformation can, theoretically, have unintended and counterproductive effects (Carey et al. 2020). First, corrections can spur directionally motivated reasoning (the fact that people are more accepting of false information that is in line with their pre-existing beliefs) among those with a predisposition to endorse conspiracy theories (Carey et al. 2020, 20; Walter et al. 2020, 353), or among people who believe in the specific misperceptions that are being debunked. For these people, corrections may fail to reduce misperceptions (Carey et al. 2020, 21). Second, in countering a specific phenomenon with a correction is that it can reduce belief in other facts that are true (Carey et al. 2020, 6). This appeared to be the case in Carey et al.'s (2020) experiment to test the effectiveness of corrections against false information about the Zika virus and Yellow fever in Brazil. The results of the study indicate that not only did the corrections fail to reduce belief in the false information, they also reduced belief in true facts about the virus (Carey et al. 2020, 1, 8–9). A third potential consequence of trying to correct false and misleading information is the phenomena identified in previous research and called the *backfire effect*. The backfire effect suggests that not only can corrections fail, resulting in continued influence of misinformation, but they can also "backfire," leading to a strengthening of an individual's belief in the very same misperception it intended to correct (Swire-Thompson, DeGutis, and Lazer 2020, 287; Lewandowsky et al. 2012; Nyhan and Reifler 2010).

8.1 Different types of backfire

According to Paynter et al. (2019), a backfire effect can arise, theoretically, in one or more of the following potential cases: first, a correction can be rejected and potentially backfire because of a psychological reactance due to the authoritative nature of the correction. Second, it can also be rejected because of familiarity with the claim, which means that when correcting a false statement, a repetition of the misconception could sometimes be included, thus leading to a strengthening of the misconception (Paynter et al. 2019, 2). This form of backfire effect is usually referred to as the familiarity backfire effect (Swire-Thompson, DeGutis, and Lazer 2020). This phenomena is also confirmed by previous research, which states that repeating a statement could eventually increase its acceptance as truth, since repetition could lead to a higher acceptance of the credibility of the statement and thus create a social consensus of its truthiness (Peter and Koch 2016, 6; Lewandowsky et al. 2012, 113). Furthermore, corrections that involve emotional statements could potentially backfire due to the fear they evoke and, finally, a backfire effect could occur if the correction involves an attack on a person's core beliefs, which leads to a desire to protect one's worldview (Paynter et al. 2019, 2; Swire-Thompson, DeGutis, and Lazer 2020, 287). According to Swire-Thompson

et al., who reviewed the current literature on backfire effects, two of these above-mentioned reasons for the occurrence of backfire, the worldview backfire effect and the familiarity backfire effect, have been popularised in the literature (Swire-Thompson, DeGutis, and Lazer 2020, 286–87).

The worldview backfire effect occurs when a person's belief system is threatened, which motivates a defensive reaction to protect one's own worldview and thus strengthens the original belief in the misinformation (Swire-Thompson, DeGutis, and Lazer 2020, 287). The worldview backfire effect stems, according to Swire-Thompson, DeGutis, and Lazer (2020) from Nyhan and Reifler's highly influential article (2010), which not only proclaimed the failure of corrections to correct misperceptions among a certain targeted ideological group, but also suggests that the corrections actually increase the misperceptions among the target group and thus result in a backfire effect. The familiarity backfire effect originated, according to Swire-Thompson et al., from a highly cited unpublished manuscript, where participants viewed a flyer containing both myths and facts about a flu vaccine. In the study, the participants reported less favourable attitudes toward the vaccination than those who did not view the flyer (Swire-Thompson, DeGutis, and Lazer 2020, 288). Both these types of backfire effects have since been tested in different settings and contexts, in the literature, and in some cases a backfire effect appears to exist (Nyhan and Reifler 2010; Hart and Nisbet 2012; Nyhan, Reifler, and Ubel 2013; Nyhan et al. 2014; Nyhan and Reifler 2015; Zhou 2016; Ecker and Ang 2019; Pluviano, Watt, and Sala 2017; Pluviano et al. 2019). Due in part to difficulties in replicating these studies, Swire-Thompson et al. (2020) conclude that the backfire is not a robust empirical phenomena.

8.2 Results

The results of our review are consistent with the findings of Swire-Thompson, DeGutis, and Lazer (2020), in that most of the literature studied within the frame of this article suggest that corrections do not trigger a backfire effect. For example, no backfire effect could be found in a study made by Wood and Porter (2019), which, with inspiration from the study by Nyhan and Reifler (2010), tested the backfire effect along 52 polarised issues, with more than 10,000 participants (Wood and Porter 2019). This null effect was the outcome despite testing the phenomena in the context of polarised issues where the backfire effect is said to occur and across five different experiments (Wood and Porter 2019, 135). Furthermore, no backfire effects were found in a 2019 study by Ecker et al. (2019), where the authors tested the familiarity backfire effect through using simple retractions that repeat a false claim, while tagging it false (Ecker et al. 2019). Similarly, no backfire effect was found in several studies using both visual and textual fact-checkers or debunking strategies (Hameleers 2020, 297; Paynter et al. 2019; Vraga and Bode 2021, 402). Zooming into fact-checking, specifically, a literature review of 30 studies using fact-checking practices also provides no

evidence of factual backfire. On the contrary, fact-checking seemed to positively affect beliefs irrespective of political ideology, context or pre-existing positions (Walter et al. 2020, 367). Nor does using different tones when correcting seem to lead to a backfire effect (Tully, Bode, and Vraga 2020, 9). Finally, the backfire effect was also absent in a study that tested the attachment of warnings to false information; and not only so, the study also showed that the warnings were *more* effective for false headlines that were in line with the political ideology of the participants (Pennycook et al. 2020, 4955), which stands against the fact of motivated reasoning in their review.

Some cases of a backfire effect have been identified, however, although the results are not clear. For example, a backfire effect may have occurred in a study conducted by Vraga et al. (2019), where the authors tested logic- and humour-based corrections in the context of three specific topics. Here, the logic-based correction boosted the credibility perceptions of a misinformation tweet in the context of HPV vaccination among the people who already believed the scientific consensus within the field (Vraga, Kim, and Cook 2019, 407). This, the authors argue, may have been due to an occurrence of a backfire effect, but might also have been the result of a greater uncertainty about the HPV vaccination, in general, and perhaps not due to the correction itself. Also, in a study by Ecker, Butler, and Hamly(2020), some evidence for a familiarity backfire effect was identified in their first study, but failed to be repeated in their second and third study.

8.3 Conclusions on the backfire effect

It is therefore difficult to rule out that the backfire effect does not exist at all. Some of the studies reviewed, such as Ecker et al. (2019), reflect on this. The authors argue that it is difficult to rule out the backfire effect entirely, since some of the claims they used in the study could have been familiar to the participants. If they would have used entirely novel claims, it might have led to a different result, including an identified backfire effect (Ecker et al. 2019, 13). In addition, Lewandowsky et al. (2012), before Nyhan and Reiflers article was released, in 2010, mention several studies that identify a backfire effect. The implications of these findings are not clear-cut, since the authors used a slightly different definition of “backfire” than in, e.g., Nyhan and Refiler (2010). Also, other previously conducted research, such as Vraga and Bode (2017), identified a potential case of backfire, where a second correction that took place after an expert correction reinforced original misperceptions among participants who originally had a low level of misperceptions (Vraga and Bode 2017, 16). There are also some further cases of identifying the backfire effect, which Thompson et al. (2020) mention in their review.

Taken together, it seems that the risk of a backfire effect is limited. That said, there are other types of risks in countering mis- and disinformation, which some of the studies reviewed here identify. As an example, Nyhan et al. (2020) tested the

effects of exposure to fact-checks of claims made by Donald Trump, both among his supporters and non-supporters. The results indicate that the fact-checks seem to update participants' factual beliefs, even among the Trump supporters, who at the same time nevertheless viewed the articles as less accurate and fair when a fact-check was included (Nyhan et al., 948, 957). This, the authors argue, may indicate that updates in factual beliefs and motivated reasoning can coexist and thus need not be a mutually exclusive phenomena. Other studies, such as Lewandowsky et al. (2012), also point to the risk of the continued influence effect, among other things. To that end, there seem to be multiple risks in countering dis- and misinformation, risks that future literature reviews could do well to look into. However, regarding the backfire effect, specifically, the present review shows that the risk of a backfire effect's occurring because of countermeasures against or corrections to false and misleading information seems limited.

9 What we have learned and may still learn

This review seeks to discern the state of scientific knowledge about countering mis- and disinformation. It proceeds by using recently published works as a gateway to the field of research in focus. This chapter summarises and discusses the results and outlines paths for future research.

9.1 The state of knowledge regarding efficiency of reactive countermeasures

The selected works discussed above show a large variety in approaches and focal points. Their key commonality is that they seek to understand under what circumstances people tend to be susceptible to changing their beliefs, opinions, and perceptions about specific factual matters. The general result, if one chooses to interpret the outcomes in a positive manner, is that corrections generally have the ability to influence peoples' ways of thinking. Fact-checking, visual cues to potentially misleading informational content, and corrections by social peers, to name a few examples, are effective in the short term, at least on certain types of issues.

However, as the review shows, there are great differences between studies regarding their conclusions about how effective various countermeasures can be, and for how long. As Walter et al. (2020, 366f) discuss, most measures are only able to affect recipients in matters in which they are not particularly knowledgeable. The effect generally decreases as soon as the correction concerns issues that are either politically normative or issues where the subject possesses a higher degree of knowledge. In situations where previously highly scientific issues become politicised, such as in the case of countermeasures against the Covid-19 pandemic, opinions are less easily swayed. As Walter et al. (2020, 367) note, fact-checking and other measures are perhaps most needed in situations such as election campaigns because these are the times when people tend to encounter a higher degree of misleading information. These are also at the same time a context in which many issues tend to become politicised. Presenting a timely response when and where it is most needed is perhaps also one of the most substantial problems for the art of countering mis- and disinformation. The nature of the response, whether through visual cues or extensive factual descriptions, for example, may matter less.

The selected works also discuss some of the potential problems in trying to address misinformation. Clayton et al. (2020) discovered that using general warnings, as a way of combating dis- and misinformation, reduced the perceived accuracy of not

only false headlines, but of true headlines. This points to a dangerous aspect of using general, sweeping warnings as a countermeasure against mis- and disinformation (Clayton et.al, 2020: 1091). Further research would be well suited to investigate whether this "spill-over effect" might occur under different circumstances.

The potential for backfire, which used to be viewed as a verified fact (see Lewandowsky et al. 2012), would pose a substantial headache to any communicator trying to produce a narrative to counter mis- or disinformation. The current state of scientific knowledge – and this is one of the more tangible results of the review – does not lend much support to the existence of backfire effects. On the contrary, several large-scale experiments have shown no such effect at all, even given favourable conditions (Swire-Thompson, DeGutis, and Lazer 2020; Wood and Porter 2019). The question of the effect of repetition has undergone a similar transformation: whereas Lewandovsky et al. (2012) claim, in their seminal review, that repetition increases the chance that recipients will believe the misinformation rather than the correction, the tendency today is the contrary: repeating misinformation may in certain circumstances produce even *positive* effects (see Section 4.4, above). Some problems in correcting mis- and disinformation that were previously thought of as more or less proven have been shown to be highly contextual and difficult to validate.

One reason for the differences in results regarding effects that have previously been taken-for-granted and that has not yet received attention is the evolving nature of the misinformation landscape. As Section 2.2 shows, the number of publications on this topic has virtually exploded since roughly 2015, which may be a trend related to the fact that disinformation has become a buzzword. However, the fact that it has done so can be viewed as a consequence of the information-influence campaigns by, *inter alia*, Russian state or state-sponsored actors, during the Crimea annexation; the 2016 US general elections; and the Brexit election of 2016. Since then, several elections throughout Europe have been targeted (Fjällhed, Pamment, and Bay 2021; Jeangène Vilmer et al. 2018; Nothhaft et al. 2019). During this period, Western populations have both increased their general social media usage and likely become more aware of the dubious nature of online information-sharing, fake news, and media manipulation. The lack of panel data that spans long periods of time, as well as the evolving nature of social media platforms and information consumption generally in society, makes such knowledge difficult or even impossible to generate. Even though it does seem unlikely that the time factor would play such a large role, this has not been tested and cannot be ruled out *prima facie*.

9.2 Three research gaps

The currently most pressing issues are how to understand the (1) mechanisms and (2) effects of exposure to online disinformation, and (3) how to counter adverse

effects. In order to achieve this, we need to look beyond existing research. When reviewing research on reactive countermeasures to mis- and disinformation, three main discrepancies in the existing research stand out: first, the uncertainties in research results, as there are many instances where results contradict each other; second, the lack of research on real-life social media behaviour; and third, the lack of research on how countermeasures work for practitioners. These gaps concern not so much topical aspects as methodological ones. They are key if the field is to develop into a more coherent literature.

First, as identified in this review, there are many cases where research results are contradictory. This may be inevitable due to the volatile nature of the research subject and the fact that respondents live in a rapidly changing world. However, more coherent results are needed to produce results that are generalisable. One aspect of this issue is connected to the fact that most extant research is carried out in the US. This is perhaps understandable due to the events surrounding the 2016 US general election, as well as the polarised US political landscape, between Republicans and Democrats. However, there is a need to carry out studies in other cultural contexts in order to analyse whether the results reported in this review also hold outside the USA or whether they are defined by that specific context. At least in the Swedish context, few studies in this direction have been carried out, with the notable exception of Nygren et al. (2021), who applied a professional fact-checking tool on the curricular digital activities of pupils in four different European countries with good results.[6] In this article, Nygren et al. show that there are significant differences between the different countries included in the study, which again accentuates the need to verify results in different contexts. This is an important task that would contribute considerably to a field in dire need of replication due to the many different outcomes, as reported above.

The second gap in the research is methodological in nature. Generally, the field needs stronger evidence-based results from large representative-sample studies. Today, most of the quantitative work cited above is experimental, based on comparably small samples using either paid participants (Amazon Turk, Yougov, etc.), or students, as test subjects. This is acceptable in a model-testing and theory development phase, but as the field draws closer to producing more solid hypotheses, a stronger evidence base is needed.

A complicating feature of the research discussed here is the structural conditions under which it operates. This field of research concerns human behaviour mediated through both text, graphical elements, social media platforms' design features, and researchers' own data-collection instruments. The difference between research results from tentative behavioural experiments and actual behaviour on social media platforms is likely substantial. Participating in a survey experiment is not the same as encountering mis- or disinformation in real life. Adding to the

[6] This article was published outside of the time range for the collection of data for this review.

complexity is the multifaceted use of semantic symbols, neologisms, subliminal implications, and "extramoral" relationships to the truth and lies that flourish on social media today. Interpreting text-based information can indeed be challenging. These are factors that should be addressed by researchers, but ultimately are under the control of the social media platforms and their parent corporations. As long as researchers do not have access to comprehensive data about online behaviour, and as long as large social media platforms restrict and do not disclose research data, this problem will remain (Kaye 2021; Wall Street Journal 2021).

The third gap in the research reviewed here intersects with the first two. It concerns the lack of research on how communicative countermeasures to mis- and disinformation work in practice, from the perspective of *practitioners*. A vast number of corporations, agencies, fact-checkers, NGOs, international organisations, etc., employ communicators who work continuously on social media platforms. This group of communication professionals in many cases employs the countermeasures discussed above and is reasonably well suited as respondents and research partners if one wants to study what works, and for whom. For example, Vraga and Bode (2017) found evidence, contrary to other studies, such as the ones by Wood and Porter (2019), that multiple corrections might induce backfire effects. Future research should study whether this really is the case, and could use communicators as entrance points to the empirical data collection. This type of research could of course also be used to replicate results, as discussed above, in this section. Taking practitioners' experiences seriously and involving them as partners in research could open up new avenues in the study of how to counter mis- and disinformation.

9.3 Concluding remarks

This review sought to scan existing, recently published works to find out what research can tell us about effective countering of mis- and disinformation. The findings are somewhat diversified, and some researchers have shown that corrections are only occasionally effective, while others suggest that they may even be counterproductive (Nyhan and Reifler 2010; Rich and Zaragoza 2020). However, meta-analyses such as the ones by Walter and Murphy (2018) and Walter and Tukachinsky (2020) provide evidence that exposure to corrections often encourages individuals in changing their misinformed beliefs, especially regarding health issues and when the correction is provided by an expert source, due to source credibility. In a few years, it would be interesting to perform an updated review to see whether these results still hold. In general, although corrections do not always lead to decreased belief in misinformation, and may theoretically sometimes even be counterproductive, correcting misinformation is still better than not doing anything at all.

10 References

Afriat, Hagar, Shira Dvir-Gvirsman, Keren Tsuriel, and Lidor Ivan. 2021. "'This Is Capitalism. It Is Not Illegal': Users' Attitudes toward Institutional Privacy Following the Cambridge Analytica Scandal." *The Information Society* 37 (2): 115–27. https://doi.org/10.1080/01972243.2020.1870596.

Akoz, Kemal Kivanç, and Cemal Eren Arbatli. 2016. "Information Manipulation in Election Campaigns." *Economics & Politics* 28 (2): 181. https://doi.org/10.1111/ecpo.12076.

Altay, Sacha, Anne-Sophie Hacquin, and Hugo Mercier. 2020. "Why Do so Few People Share Fake News? It Hurts Their Reputation." *New Media & Society*, November, 1461444820969893. https://doi.org/10.1177/1461444820969893.

Amazeen, Michelle A., Chris J. Vargo, and Toby Hopp. 2019. "Reinforcing Attitudes in a Gatewatching News Era: Individual-Level Antecedents to Sharing Fact-Checks on Social Media." *Communication Monographs* 86 (1): 112–32. https://doi.org/10.1080/03637751.2018.1521984.

Ardevol-Abreu, Alberto, Patricia Delponti, and Carmen Rodriguez-Wanguemert. 2020. "Intentional or Inadvertent Fake News Sharing? Fact-Checking Warnings and Users' Interaction with Social Media Content." *Profesional De La Informacion* 29 (5): e290507. https://doi.org/10.3145/epi.2020.sep.07.

Bode, Leticia, and Emily K. Vraga. 2018. "See Something, Say Something: Correction of Global Health Misinformation on Social Media." *Health Communication* 33 (9): 1131–40. https://doi.org/10.1080/10410236.2017.1331312.

———. 2021. "The Swiss Cheese Model for Mitigating Online Misinformation." *Bulletin of the Atomic Scientists* 77 (3): 129–33. https://doi.org/10.1080/00963402.2021.1912170.

Bode, Leticia, Emily K. Vraga, and Melissa Tully. 2020. "Do the Right Thing: Tone May Not Affect Correction of Misinformation on Social Media." *Harvard Kennedy School Misinformation Review*, June. https://doi.org/10.37016/mr-2020-026.

———. 2021. "Correcting Misperceptions About Genetically Modified Food on Social Media: Examining the Impact of Experts, Social Media Heuristics, and the Gateway Belief Model." *Science Communication* 43 (2): 225–51. https://doi.org/10.1177/1075547020981375.

Bolsen, Toby, and James N. Druckman. 2015. “Counteracting the Politicization of Science.” *Journal of Communication* 65 (5): 745–69. https://doi.org/10.1111/jcom.12171.

Bowles, Jeremy, Horacio Larreguy, and Shelley Liu. 2020. “Countering Misinformation via WhatsApp: Preliminary Evidence from the COVID-19 Pandemic in Zimbabwe.” *PLoS One* 15 (10): e0240005. http://dx.doi.org.ep.fjernadgang.kb.dk/10.1371/journal.pone.0240005.

Brandtzaeg, Petter Bae, Asbjørn Følstad, and María Ángeles Chaparro Domínguez. 2018. “How Journalists and Social Media Users Perceive Online Fact-Checking and Verification Services.” *Journalism Practice* 12 (9): 1109–29. https://doi.org/10.1080/17512786.2017.1363657.

Brashier, Nadia M., Gordon Pennycook, Adam J. Berinsky, and David G. Rand. 2021. “Timing Matters When Correcting Fake News.” *Proceedings of the National Academy of Sciences* 118 (5): e2020043118. https://doi.org/10.1073/pnas.2020043118.

Bryanov, Kirill, and Victoria Vziatysheva. 2021. “Determinants of Individuals’ Belief in Fake News: A Scoping Review Determinants of Belief in Fake News.” *PLoS One* 16 (6): e0253717. http://doi.org/10.1371/journal.pone.0253717.

Carey, John M., Victoria Chi, D. J. Flynn, Brendan Nyhan, and Thomas Zeitzoff. 2020. “The Effects of Corrective Information about Disease Epidemics and Outbreaks: Evidence from Zika and Yellow Fever in Brazil.” *Science Advances* 6 (5): eaaw7449. https://doi.org/10.1126/sciadv.aaw7449.

Chan, Man-pui Sally, Christopher R. Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. “Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation.” *Psychological Science* 28 (11): 1531–46. https://doi.org/10.1177/0956797617714579.

Clayton, Katherine, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, et al. 2020. “Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media.” *Political Behavior* 42 (4): 1073–95. https://doi.org/10.1007/s11109-019-09533-0.

Colliander, Jonas. 2019. “‘This Is Fake News’_ Investigating the Role of Conformity to Other Users’ Views When Commenting on and Spreading Disinformation in Social Media.” *Computers in Human Behavior* 97: 202–15. https://doi.org/10.1016/j.chb.2019.03.032.

Connor Desai, Saoirse A., Toby D. Pilditch, and Jens K. Madsen. 2020. “The Rational Continued Influence of Misinformation.” *Cognition* 205 (December): 104453. https://doi.org/10.1016/j.cognition.2020.104453.

Cresci, Stefano, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2018. “Social Fingerprinting: Detection of Spambot Groups Through DNA-Inspired Behavioral Modeling.” *IEEE Transactions on Dependable and Secure Computing* 15 (4): 561–76. https://doi.org/10.1109/TDSC.2017.2681672.

Dai, Yue, Wenting Yu, and Fei Shen. 2021. “The Effects of Message Order and Debiasing Information in Misinformation Correction.” *International Journal of Communication* 15: 1039–59. https://doi.org/1932–8036/20210005.

Dal Cin, Sonya, Mark P. Zanna, and Geoffrey T. Fong. 2004. “Narrative Persuasion and Overcoming Resistance.” In *Resistance and Persuasion*, 175–91. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

De Pryck, Kari, and François Gemenne. 2017. “The Denier-in-Chief: Climate Change, Science and the Election of Donald J. Trump.” *Law and Critique* 28 (2): 119–26. https://doi.org/10.1007/s10978-017-9207-6.

DG CNECT. 2018. “A Multi-Dimensional Approach to Disinformation.” Report of the independent High level Group on fake news and online disinformation. Brussels: European Commission. https://data.europa.eu/doi/10.2759/739290.

Dobreva, Diyana, Daniel Grinnell, and Martin Innes. 2019. “Prophets and Loss: How ‘Soft Facts’ on Social Media Influenced the Brexit Campaign and Social Reactions to the Murder of Jo Cox MP.” *Policy and Internet*, May. https://doi.org/10.1002/poi3.203.

Duncombe, Constance. 2019. “Popular Culture, Post-Truth and Emotional Framings of World Politics.” *Australian Journal of Political Science* 54 (4): 543–55. https://doi.org/10.1080/10361146.2019.1663405.

Eberle, Jakub, and Jan Daniel. 2019. “‘Putin, You Suck’: Affective Sticking Points in the Czech Narrative on ‘Russian Hybrid Warfare.’” *Political Psychology* 40 (6): 1267–81. https://doi.org/10.1111/pops.12609.

Ecker, Ullrich K. H., and Li Chang Ang. 2019. “Political Attitudes and the Processing of Misinformation Corrections.” *Political Psychology* 40 (2): 241–60. https://doi.org/10.1111/pops.12494.

Ecker, Ullrich K. H., and Luke Antonio. 2020. “Can You Believe It? An Investigation into the Impact of Retraction Source Credibility on the Continued Influence Effect.” PsyArXiv. https://doi.org/10.31234/osf.io/qt4w8.

Ecker, Ullrich K. H., Lucy H. Butler, and Anne Hamby. 2020. “You Don’t Have to Tell a Story! A Registered Report Testing the Effectiveness of Narrative versus Non-Narrative Misinformation Corrections.” *Cognitive Research: Principles and Implications* 5 (1): 64. https://doi.org/10.1186/s41235-020-00266-x.

Ecker, Ullrich K. H., Stephan Lewandowsky, and Matthew Chadwick. 2020. “Can Corrections Spread Misinformation to New Audiences? Testing for the Elusive Familiarity Backfire Effect.” *Cognitive Research: Principles and Implications* 5 (1): 41. https://doi.org/10.1186/s41235-020-00241-6.

Ecker, Ullrich K. H., Stephan Lewandowsky, and David T. W. Tang. 2010. “Explicit Warnings Reduce but Do Not Eliminate the Continued Influence of Misinformation.” *Memory & Cognition* 38 (8): 1087–1100. https://doi.org/10.3758/MC.38.8.1087.

Ecker, Ullrich K. H., Ziggy O’Reilly, Jesse S. Reid, and Ee Pin Chang. 2019. “The Effectiveness of Short-Format Refutational Fact-Checks.” *British Journal of Psychology* 111 (1): 36–54. https://doi.org/10.1111/bjop.12383.

Figueira, Alvaro, Nuno Guimaraes, and Luis Torgo. 2018. “Current State of the Art to Detect Fake News in Social Media: Global Trendings and Next Challenges:” In *Proceedings of the 14th International Conference on Web Information Systems and Technologies*, 332–39. Seville, Spain: SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0007188503320339.

Fjällhed, Alicia, James Pamment, and Sebastian Bay. 2021. “A Swedish Perspective on Foreign Election Interference.” In *Defending Democracies*, edited by Duncan Hollis and J.D. Ohlin, p. 139-161. United Kingdom: Oxford University Press.

Frenda, Steven J., Rebecca M. Nichols, and Elizabeth F. Loftus. 2011. “Current Issues and Advances in Misinformation Research.” *Current Directions in Psychological Science* 20 (1): 20–23. https://doi.org/10.1177/0963721410396620.

Garrett, R Kelly, and Shannon Poulsen. 2019. “Flagging Facebook Falsehoods: Self-Identified Humor Warnings Outperform Fact Checker and Peer Warnings.” *Journal of Computer-Mediated Communication* 24 (5): 240–58. https://doi.org/10.1093/jcmc/zmz012.

Grant, Maria J., and Andrew Booth. 2009. “A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies: A Typology of Reviews, *Maria J. Grant & Andrew Booth*.” *Health Information & Libraries Journal* 26 (2): 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x.

Hallahan, Kirk, Derina Holtzhausen, Betteke van Ruler, Dejan Verčič, and Krishnamurthy Sriramesh. 2007. “Defining Strategic Communication.” *International Journal of Strategic Communication* 1 (1): 3–35. https://doi.org/10.1080/15531180701285244.

Hameleers, Michael. 2020. “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* 0 (0): 1–17. https://doi.org/10.1080/1369118X.2020.1764603.

Hameleers, Michael, Thomas E. Powell, Toni G. L. A. Van Der Meer, and Lieke Bos. 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.

Hart, P. Sol, and Erik C. Nisbet. 2012. “Boomerang Effects in Science Communication: How Motivated Reasoning and Identity Cues Amplify Opinion Polarization About Climate Mitigation Policies.” *Communication Research* 39 (6): 701–23. https://doi.org/10.1177/0093650211416646.

Jang, Jeong-woo, Eun-Ju Lee, and Soo Yun Shin. 2019. “What Debunking of Misinformation Does and Doesn’t.” *Cyberpsychology, Behavior, and Social Networking* 22 (6): 423–27. https://doi.org/10.1089/cyber.2018.0608.

Jeangène Vilmer, Jean-Baptiste, Alexandre Escorcia, Marine Guillaume, and Janaina Herrera. 2018. “Les manipulations de l’information: Un défi pour nos démocraties.” Paris: CAPS (ministère de l’Europe et des Affaires étrangères) & IRSEM (ministère des Armées).

Kaye, Kate. 2021. “Princeton Researchers Ditch Facebook Political Ad Project.” *Digiday* (blog). August 12, 2021. https://digiday.com/marketing/princeton-researchers-ditch-facebook-political-ad-project-after-the-platform-used-a-debunked-ftc-privacy-defense/.

Kim, Antino, Patricia L. Moravec, and Alan R. Dennis. 2019. “Combating Fake News on Social Media with Source Ratings: The Effects of User and Expert Reputation Ratings.” *Journal of Management Information Systems* 36 (3): 931–68. https://doi.org/10.1080/07421222.2019.1628921.

Kim, Ji Won, and Gina Chen Masullo. 2020. “Exploring the Influence of Comment Tone and Content in Response to Misinformation in Social Media News.” *Journalism Practice*, March, 1–15. https://doi.org/10.1080/17512786.2020.1739550.

Kuru, Ozan, Dominik Stecula, Hang Lu, Yotam Ophir, Sally Chan Man-pui, Ken Winneg, Kathleen Hall Jamieson, and Dolores Albarracín. 2021. “The Effects of Scientific Messages and Narratives about Vaccination.” *PLoS One* 16 (3): e0248328. http://doi.org/10.1371/journal.pone.0248328.

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, et al. 2018. “The Science of Fake News.” *Science* 359 (6380): 1094–96. https://doi.org/10.1126/science.aao2998.

Lewandowsky, Stephan, Ullrich K. H. Ecker, and John Cook. 2017. “Beyond Misinformation: Understanding and Coping with the ‘Post-Truth’ Era.” *Journal of Applied Research in Memory and Cognition* 6 (4): 353–69. https://doi.org/10.1016/j.jarmac.2017.07.008.

Lewandowsky, Stephan, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. “Misinformation and Its Correction: Continued Influence and Successful Debiasing.” *Psychological Science in the Public Interest* 13 (3): 106–31.

Martel, Cameron, Mohsen Mosleh, and David G. Rand. 2021. “You’re Definitely Wrong, Maybe: Correction Style Has Minimal Effect on Corrections of Misinformation Online.” *Media and Communication* 9 (1): 120–33. http://doi.org/10.17645/mac.v9i1.3519.

McCombie, Stephen, Allon J. Uhlmann, and Sarah Morrison. 2020. “The US 2016 Presidential Election & Russia’s Troll Farms.” *Intelligence and National Security* 35 (1): 95–114. https://doi.org/10.1080/02684527.2019.1673940.

Meer, Toni G. L. A. van der, and Yan Jin. 2020. “Seeking Formula for Misinformation Treatment in Public Health Crises: The Effects of Corrective Information Type and Source.” *Health Communication* 35 (5): 560–75. http://doi.org/10.1080/10410236.2019.1573295.

Mills, Adam J., and Karen Robson. 2020. “Brand Management in the Era of Fake News: Narrative Response as a Strategy to Insulate Brand Value.” *The Journal of Product and Brand Management* 29 (2): 159–67. http://doi.org/10.1108/JPBM-12-2018-2150.

Mohamad, Siti Mazidah. 2020. “Creative Production of ‘COVID-19 Social Distancing’ Narratives on Social Media.” *Tijdschrift Voor Economische En Sociale Geografie* 111 (3): 347–59. https://doi.org/10.1111/tesg.12430.

Nothhaft, Howard, James Pamment, Henrik Agardh-Twetman, and Alicia Fjällhed. 2019. “.” In *Countering Online Propaganda and Violent Extremism*, edited by Corneliu Bjola and James Pamment. United Kingdom: Routledge.

Nygren, Thomas, Mona Guath, Werner Axelsson, Carl-Anton, and Divina Frau-Meigs. 2021. “Combatting Visual Fake News with a Professional Fact-Checking Tool in Education in France, Romania, Spain and Sweden.” *Information* 12 (5): 201. http://doi.org/10.3390/info12050201.

Nyhan, Brendan, Ethan Porter, Jason Reifler, and Thomas J. Wood. 2020. “Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability.” *Political Behavior* 42 (3): 939–60. https://doi.org/10.1007/s11109-019-09528-x.

Nyhan, Brendan, and Jason Reifler. 2010. “When Corrections Fail: The Persistence of Political Misperceptions.” *Political Behavior* 32 (2): 303–30. https://doi.org/10.1007/s11109-010-9112-2.

———. 2015. “Does Correcting Myths about the Flu Vaccine Work? An Experimental Evaluation of the Effects of Corrective Information.” *Vaccine* 33 (3): 459–64. http://doi.org./10.1016/j.vaccine.2014.11.017.

Nyhan, Brendan, Jason Reifler, Sean Richey, and Gary L. Freed. 2014. “Effective Messages in Vaccine Promotion: A Randomized Trial.” *Pediatrics* 133 (4): e835–42. https://doi.org/10.1542/peds.2013-2365.

Nyhan, Brendan, Jason Reifler, and Peter A. Ubel. 2013. “The Hazards of Correcting Myths About Health Care Reform.” *Medical Care* 51 (2): 127–32.

Ophir, Yotam, Dan Romer, Patrick E. Jamieson, and Kathleen Hall Jamieson. 2020. “Counteracting Misleading Protobacco YouTube Videos: The Effects of Text-Based and Narrative Correction Interventions and the Role of Identification.” *International Journal of Communication* 14: 4973–88.

Paynter, Jessica, Sarah Luskin-Saxby, Deb Keen, Kathryn Fordyce, Grace Frost, Christine Imms, Scott Miller, David Trembath, Madonna Tucker, and Ullrich Ecker. 2019. “Evaluation of a Template for Countering Misinformation—Real-World Autism Treatment Myth Debunking.” *PLOS ONE* 14 (1): e0210746. https://doi.org/10.1371/journal.pone.0210746.

Pennycook, Gordon, Adam Bear, Evan T Collins, and David G Rand. 2020. “The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings.” *Management Science* 66 (11): 15.

Peter, Christina, and Thomas Koch. 2016. “When Debunking Scientific Myths Fails (and When It Does Not): The Backfire Effect in the Context of Journalistic Coverage and Immediate Judgments as Prevention Strategy.” *Science Communication* 38 (1): 3–25. https://doi.org/10.1177/1075547015613523.

Peter, Fabienne. 2017. “Political Legitimacy.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2017/entries/legitimacy/.

Pluviano, Sara, Caroline Watt, Giovanni Ragazzini, and Sergio Della Sala. 2019. “Parents’ Beliefs in Misinformation about Vaccines Are Strengthened by pro-Vaccine Campaigns.” *Cognitive Processing* 20 (3): 325–31. https://doi.org/10.1007/s10339-019-00919-w.

Pluviano, Sara, Caroline Watt, and Sergio Della Sala. 2017. “Misinformation Lingers in Memory: Failure of Three pro-Vaccination Strategies.” *PLOS ONE* 12 (7): e0181640. https://doi.org/10.1371/journal.pone.0181640.

Psykförsvarsutredningen. 2020. “En Ny Myndighet För Att Stärka Det Psykologiska Försvaret.” SOU 2020:29. Stockholm: Regeringskansliet.

Rich, Patrick R., and Maria S. Zaragoza. 2020. “Correcting Misinformation in News Stories: An Investigation of Correction Timing and Correction Durability.” *Journal of Applied Research in Memory and Cognition* 9 (3): 310–22. https://doi.org/10.1016/j.jarmac.2020.04.001.

Sangalang, Angeline, Yotam Ophir, and Joseph N Cappella. 2019. “The Potential for Narrative Correctives to Combat Misinformation†.” *Journal of Communication* 69 (3): 298–319. https://doi.org/10.1093/joc/jqz014.

Shao, Chengcheng, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. 2018. “Anatomy of an Online Misinformation Network.” *PLoS ONE* 13 (4): 1–23. https://doi.org/10.1371/journal.pone.0196087.

Sigerist, Henry E. 1938. “Science and Democracy.” *Science & Society* 2 (3): 291–99.

Solon, Olivia, and Emma Graham-Harrison. 2018. “The Six Weeks That Brought Cambridge Analytica Down.” The Guardian. May 3, 2018. http://www.theguardian.com/uk-news/2018/may/03/cambridge-analytica-closing-what-happened-trump-brexit.

Sun, Yanqing, Stella C. Chia, Fangcao Lu, and Jeffry Oktavianus. 2020. “The Battle Is On: Factors That Motivate People to Combat Anti-Vaccine Misinformation.” *Health Communication* 0 (0): 1–10. https://doi.org/10.1080/10410236.2020.1838108.

Sun, Yanqing, Jeffry Oktavianus, Sai Wang, and Fangcao Lu. 2021. “The Role of Influence of Presumed Influence and Anticipated Guilt in Evoking Social Correction of COVID-19 Misinformation.” *Health Communication* 0 (0): 1–10. https://doi.org/10.1080/10410236.2021.1888452.

Swire-Thompson, Briony, Joseph DeGutis, and David Lazer. 2020. “Searching for the Backfire Effect: Measurement and Design Considerations.” *Journal of Applied Research in Memory and Cognition* 9 (3): 286–99. http://doi.org/10.1016/j.jarmac.2020.06.006.

Tandoc, Edson C, Darren Lim, and Rich Ling. 2020. “Diffusion of Disinformation: How Social Media Users Respond to Fake News and Why.” *Journalism*. https://doi.org/10.1177/1464884919868325.

Tong, Chau, Hyungjin Gill, Jianing Li, and Sebastián Valenzuela. 2020. “‘Fake News Is Anything They Say!’ — Conceptualization and Weaponization of Fake News among the American Public.” *Mass Communication & Society* 23 (5): 755–78. https://doi.org/10.1080/15205436.2020.1789661.

Tully, Melissa, Leticia Bode, and Emily K. Vraga. 2020. “Mobilizing Users: Does Exposure to Misinformation and Its Correction Affect Users’ Responses to a Health Misinformation Post?” *Social Media + Society* 6 (4): 205630512097837. https://doi.org/10.1177/2056305120978377.

van der Linden, Sander, Anthony Leiserowitz, and Edward Maibach. 2019. “The Gateway Belief Model: A Large-Scale Replication.” *Journal of Environmental Psychology* 62 (April): 49–58. https://doi.org/10.1016/j.jenvp.2019.01.009.

Veil, Shari R., Tara Buehner, and Michael J. Palenchar. 2011. “A Work-In-Process Literature Review: Incorporating Social Media in Risk and Crisis Communication: Social Media and Crisis Communication.” *Journal of Contingencies and Crisis Management* 19 (2): 110–22. https://doi.org/10.1111/j.1468-5973.2011.00639.x.

Vraga, Emily K., and Leticia Bode. 2017. “Using Expert Sources to Correct Health Misinformation in Social Media:” *Science Communication*, September. https://doi.org/10.1177/1075547017731776.

———. 2020. “Correction as a Solution for Health Misinformation on Social Media.” *American Journal of Public Health* 110: S278–80. http://doi.org/10.2105/AJPH.2020.305916.

———. 2021. “Addressing COVID-19 Misinformation on Social Media Preemptively and Responsively.” *Emerging Infectious Diseases* 27 (2). http://doi.org/10.3201/eid2702.203139.

Vraga, Emily K., Sojung Claire Kim, and John Cook. 2019. “Testing Logic-Based and Humor-Based Corrections for Science, Health, and Political Misinformation on Social Media.” *Journal of Broadcasting & Electronic Media* 63 (3): 393–414. https://doi.org/10.1080/08838151.2019.1653102.

Vraga, Emily K., Sojung Claire Kim, John Cook, and Leticia Bode. 2020. "Testing the Effectiveness of Correction Placement and Type on Instagram." *International Journal of Press-Politics* 25 (4): 632–52. https://doi.org/10.1177/1940161220919082.

Vraga, Emily K., Kathryn H. Jacobsen. 2020. "Strategies for Effective Health Communication during the Coronavirus Pandemic and Future Emerging Infectious Disease Events." *World Medical & Health Policy* 12 (3): 233–41. http://doi.org/10.1002/wmh3.359.

Vraga, Emily K., Melissa Tully, and Leticia Bode. 2021. "Assessing the Relative Merits of News Literacy and Corrections in Responding to Misinformation on Twitter." *New Media & Society*, 1461444821998691. https://doi.org/10.1177/1461444821998691.

Wall Street Journal. 2021. "The Facebook Files." *Wall Street Journal*, October 1, 2021, sec. Tech. https://www.wsj.com/articles/the-facebook-files-11631713039.

Walter, Nathan, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. 2020. "Fact-Checking: A Meta-Analysis of What Works and for Whom." *Political Communication* 37 (3): 350–75. https://doi.org/10.1080/10584609.2019.1668894.

Walter, Nathan, and Sheila T. Murphy. 2018. "How to Unring the Bell: A Meta-Analytic Approach to Correction of Misinformation." *Communication Monographs* 85 (3): 423–41. https://doi.org/10.1080/03637751.2018.1467564.

Walter, Nathan, and Nikita A. Salovich. 2021. "Unchecked vs. Uncheckable: How Opinion-Based Claims Can Impede Corrections of Misinformation." *Mass Communication and Society* 24 (4): 500–526. https://doi.org/10.1080/15205436.2020.1864406.

Walter, Nathan, and Riva Tukachinsky. 2020. "A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It? - Nathan Walter, Riva Tukachinsky, 2020." *Communication Research* 47 (2): 155–77.

Wang, Weirui, and Yan Huang. 2020. "Countering the 'Harmless E-Cigarette' Myth: The Interplay of Message Format, Message Sidedness, and Prior Experience With E-Cigarette Use in Misinformation Correction." *Science Communication*. https://doi.org/10.1177/1075547020974384.

Werder, Kelly Page, Howard Nothhaft, Dejan Verčič, and Ansgar Zerfass. 2018. "Strategic Communication as an Emerging Interdisciplinary Paradigm." *International Journal of Strategic Communication* 12 (4): 333–51. https://doi.org/10.1080/1553118X.2018.1494181.

Winkler, Peter, and Michael Etter. 2018. “Strategic Communication and Emergence: A Dual Narrative Framework.” *International Journal of Strategic Communication* 12 (4): 382–98. https://doi.org/10.1080/1553118X.2018.1452241.

Wintersieck, Amanda, Kim Fridkin, and Patrick Kenney. 2021. “The Message Matters: The Influence of Fact-Checking on Evaluations of Political Messages.” *Journal of Political Marketing* 20 (2): 93–120. https://doi.org/10.1080/15377857.2018.1457591.

Wood, Thomas, and Ethan Porter. 2019. “The Elusive Backfire Effect: Mass Attitudes’ Steadfast Factual Adherence.” *Political Behavior* 41 (1): 135–63. https://doi.org/10.1007/s11109-018-9443-y.

Zhou, Jack. 2016. “Boomerangs versus Javelins: How Polarization Constrains Communication on Climate Change.” *Environmental Politics* 25 (5): 788–811. https://doi.org/10.1080/09644016.2016.1166602.

Ördén, Hedvig, and James Pamment. 2021. *What Is So Foreign About Foreign Influence Operations?* Washington D.C.: Carnegie Endowment for International Peace. https://carnegieendowment.org/2021/01/26/what-is-so-foreign-about-foreign-influence-operations-pub-83706.

Appendix

List of articles included in the review

Authors	Title	Journal	Year
Amazeen, Michelle A.; Vargo, Chris J.; Hopp, Toby	Reinforcing attitudes in a gatewatching news era: Individual-level antecedents to sharing fact-checks on social media	*Communication Monographs*	2019
Ardèvol-Abreu, Alberto; Delponti, Patricia; Rodríguez-Wangüemert, Carmen	Intentional or inadvertent fake news sharing? Fact-checking warnings and users' interaction with social media content	*El Profesional de la Información*	2020
Bode, Leticia; Vraga, Emily K.; Tully, Melissa	Correcting Misperceptions About Genetically Modified Food on Social Media: Examining the Impact of Experts, Social Media Heuristics, and the Gateway Belief Model	*Science Communication*	2021
Bowles, Jeremy; Larreguy, Horacio; Liu, Shelley	Countering misinformation via WhatsApp: Preliminary evidence from the COVID-19 pandemic in Zimbabwe	*PLoS One*	2020
Brashier, Nadia M.; Pennycook, Gordon; Berinsky, Adam J.; Rand, David G.	Timing matters when correcting fake news	*Proceedings of the National Academy of Sciences*	2021
Bryanov, Kirill; Vziatysheva, Victoria	Determinants of individuals' belief in fake news: A scoping review determinants of belief in fake news	*PLoS One*	2021

Butterfuss, Reese; Kendeou, Panayiota	Reducing interference from misconceptions: The role of inhibition in knowledge revision	*Journal of Educational Psychology*	2019
Carey, John M.; Chi, Victoria; Flynn, D. J.; Nyhan, Brendan; Zeitzoff, Thomas	The effects of corrective information about disease epidemics and outbreaks: Evidence from Zika and yellow fever in Brazil	*Science Advances*	2020
Clayton, Katherine, et al.	Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media	*Political Behavior*	2020
Colliander, Jonas	"This is fake news"_ Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media	*Computers in Human Behavior*	2019
Connor Desai, Saoirse A.; Pilditch, Toby D.; Madsen, Jens K.	The rational continued influence of misinformation	*Cognition*	2020
Dai, Yue; Yu, Wenting; Shen, Fei	The Effects of Message Order and Debiasing Information in Misinformation Correction	*International Journal of Communication*	2021
Ecker, Ullrich K. H.; Ang, Li Chang	Political Attitudes and the Processing of Misinformation Corrections	*Political Psychology*	2019
Ecker, Ullrich K. H.; Antonio, Luke M.	Can you believe it? An investigation into the impact of retraction source credibility on the continued influence effect	*Memory & Cognition*	2021
Ecker, Ullrich K. H.; Butler, Lucy H.; Hamby, Anne	You don't have to tell a story! A registered report testing the effectiveness of narrative versus non-narrative misinformation corrections	*Cognitive Research: Principles and Implications*	2020

Ecker, Ullrich K. H.; Lewandowsky, Stephan; Chadwick, Matthew	Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire effect	*Cognitive Research: Principles and Implications*	2020
Ecker, Ullrich K. H.; O'Reilly, Ziggy; Reid, Jesse S.; Chang, Ee Pin	The effectiveness of short-format refutational fact-checks	*British Journal of Psychology*	2019
Garrett, R Kelly; Poulsen, Shannon	Flagging Facebook Falsehoods: Self-Identified Humor Warnings Outperform Fact Checker and Peer Warnings	*Journal of Computer-Mediated Communication*	2019
Hameleers, Michael; Powell, Thomas E.; Van D er Meer, Toni G. L. A.; Bos, Lieke	A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media	*Political communication*	2020
Jang, Jeong-woo; Lee, Eun-Ju; Shin, Soo Yun	What Debunking of Misinformation Does and Doesn't	*Cyberpsychology, Behavior, and Social Networking*	2019
Kim, Antino; Moravec, Patricia L.; Dennis, Alan R.	Combating Fake News on Social Media with Source Ratings: The Effects of User and Expert Reputation Ratings	*Journal of Management Information Systems*	2019
Kim, Ji Won; Chen Masullo, Gina	Exploring the Influence of Comment Tone and Content in Response to Misinformation in Social Media News	*Journalism practice*	2020
Kuru, Ozan; Stecula, Dominik; Lu, Hang; Ophir, Yotam; Man-pui, Sally Chan; Winneg, Ken; Jamieson, Kathleen Hall; Albarracín, Dolores	The effects of scientific messages and narratives about vaccination	*PLoS One*	2021
Luengo, Maria; Garcia-Marin, David	The performance of truth: politicians, fact-checking journalism, and the struggle to tackle COVID-19 misinformation	*American Journal of Cultural Sociology*	2020

Mills, Adam J.; Robson, Karen	Brand management in the era of fake news: narrative response as a strategy to insulate brand value	*The Journal of Product and Brand Management*	2020
Nyhan, Brendan; Porter, Ethan; Reifler, Jason; Wood, Thomas J.	Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability	*Political Behavior*	2020
Ó Fathaigh, Ronan; Helberger, Natali; Appelman, Naomi	The perils of legally defining disinformation	*Internet Policy Review*	2021
Ophir, Yotam; Romer, Dan; Jamieson, Patrick E.; Jamieson, Kathleen Hall	Counteracting Misleading Protobacco YouTube Videos: The Effects of Text-Based and Narrative Correction Interventions and the Role of Identification	*International Journal of Communication*	2020
Paynter, Jessica; Luskin-Saxby, Sarah; Keen, Deb; Fordyce, Kathryn; Frost, Grace; Imms, Christine; Miller, Scott; Trembath, David; Tucker, Madonna; Ecker, Ullrich	Evaluation of a template for countering misinformation—Real-world Autism treatment myth debunking	*PLOS ONE*	2019
Pennycook, Gordon; Bear, Adam; Collins, Evan T.; Rand, David G.	The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings	*Management Science*	2020
Pluviano, Sara; Watt, Caroline; Ragazzini, Giovanni; Della Sala, Sergio	Parents' beliefs in misinformation about vaccines are strengthened by pro-vaccine campaigns	*Cognitive Processing*	2019

Rich, Patrick R.; Zaragoza, Maria S.	Correcting Misinformation in News Stories: An Investigation of Correction Timing and Correction Durability	*Journal of Applied Research in Memory and Cognition*	2020
Sangalang, Angeline; Ophir, Yotam; Cappella, Joseph N	The Potential for Narrative Correctives to Combat Misinformation†	*Journal of Communication*	2019
Sun, Yanqing; Chia, Stella C.; Lu, Fangcao; Oktavianus, Jeffry	The Battle is On: Factors that Motivate People to Combat Anti-Vaccine Misinformation	*Health Communication*	2020
Sun, Yanqing; Oktavianus, Jeffry; Wang, Sai; Lu, Fangcao	The Role of Influence of Presumed Influence and Anticipated Guilt in Evoking Social Correction of COVID-19 Misinformation	*Health Communication*	2021
Swire-Thompson, Briony; DeGutis, Joseph; Lazer, David	Searching for the Backfire Effect: Measurement and Design Considerations.	*Journal of applied research in memory and cognition*	2020
Tandoc, Edson C; Lim, Darren; Ling, Rich	Diffusion of disinformation: How social media users respond to fake news and why	*Journalism*	2020
Tully, Melissa; Bode, Leticia; Vraga, Emily K.	Mobilizing Users: Does Exposure to Misinformation and Its Correction Affect Users' Responses to a Health Misinformation Post?	*Social Media + Society*	2020
van der Linden, Sander; Leiserowitz, Anthony; Maibach	The Gateway Belief Model: A Large-Scale Replication	*Journal of Environmental Psychology*	2019
van der Meer, Toni G. L. A.; Jin, Yan	Seeking Formula for Misinformation Treatment in Public Health Crises: The Effects of Corrective Information Type and Source	*Health Communication*	2020
Vraga, Emily K.; Bode, Leticia	Correction as a Solution for Health Misinformation on Social Media	*American Journal of Public Health*	2020

Vraga, Emily K.; Bode, Leticia	Addressing COVID-19 Misinformation on Social Media Preemptively and Responsively	*Emerging Infectious Diseases*	2021
Vraga, Emily K.; Bode, Leticia; Tully, Melissa	Creating News Literacy Messages to Enhance Expert Corrections of Misinformation on Twitter	*Communication Research*	2020
Vraga, Emily K.; Jacobsen, Kathryn H.	Strategies for Effective Health Communication during the Coronavirus Pandemic and Future Emerging Infectious Disease Events	*World Medical & Health Policy*	2020
Vraga, Emily K.; Kim, Sojung Claire; Cook, John	Testing Logic-based and Humor-based Corrections for Science, Health, and Political Misinformation on Social Media	*Journal of Broadcasting & Electronic Media*	2019
Vraga, Emily K.; Kim, Sojung Claire; Cook, John; Bode, Leticia	Testing the Effectiveness of Correction Placement and Type on Instagram	*International Journal of Press-Politics*	2020
Vraga, Emily K.; Tully, Melissa; Bode, Leticia	Assessing the relative merits of news literacy and corrections in responding to misinformation on Twitter	*New Media & Society*	2021
Walter, Nathan; Cohen, Jonathan; Holbert, R. Lance; Morag, Yasmin	Fact-Checking: A Meta-Analysis of What Works and for Whom	*Political Communication*	2020
Walter, Nathan; Salovich, Nikita A.	Unchecked vs. Uncheckable: How Opinion-Based Claims Can Impede Corrections of Misinformation	*Mass Communication and Society*	2021
Walter, Nathan; Tukachinsky, Riva	A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It? - Nathan Walter, Riva Tukachinsky, 2020	*Communication Research*	2020

Wang, Weirui; Huang, Yan	Countering the “Harmless E-Cigarette” Myth: The Interplay of Message Format, Message Sidedness, and Prior Experience With E-Cigarette Use in Misinformation Correction	*Science communication*	2020
Wintersieck, Amanda; Fridkin, Kim; Kenney, Patrick	The Message Matters: The Influence of Fact-Checking on Evaluations of Political Messages	*Journal of Political Marketing*	2021
Wood, Thomas; Porter, Ethan	The Elusive Backfire Effect: Mass Attitudes’ Steadfast Factual Adherence	*Political Behavior*	2019

This report consists of a review of recent research into reactive measures to counter mis- and disinformation, mainly, but not exclusively, on social media platforms. 53 articles have been selected for closer scrutiny based on method, relevance, date of publication, and publication language. The text reviews articles by focussing on tone, correction format, source rating, refutations by experts versus peers, and other aspects of direct countermeasures. The report further discusses the empirical validity of the results and identifies gaps for future research.

The target audience for this report is researchers, communicators and others who are engaged in countering misinformation, for example on social media.

FOI

ISSN 1650-1942

www.foi.se