[1]

The Landscape of User-centered Misinformation Interventions - A Systematic Literature Review

KATRIN HARTWIG, FREDERIC DOELL, and CHRISTIAN REUTER, Science and Technology for Peace and Security (PEASEC), Technische Universität Darmstadt, Germany

Misinformation is one of the key challenges facing society today. User-centered misinformation interventions as digital countermeasures that exert a direct influence on users represent a promising means to deal with the large amounts of information available. While an extensive body of research on this topic exists, researchers are confronted with a diverse research landscape spanning multiple disciplines. This review systematizes the landscape of user-centered misinformation interventions to facilitate knowledge transfer, identify trends, and enable informed decision-making. Over 6,000 scholarly publications were screened, and a systematic literature review ($N = 172$) was conducted. A taxonomy was derived regarding intervention design (e.g., labels, showing indicators of misinformation, corrections, removal, or visibility reduction of content), user interaction (active or passive), and timing (e.g., pre or post exposure to misinformation or on request of the user). We provide a structured overview of approaches across multiple disciplines and derive six overarching challenges for future research regarding transferability of approaches to (1) novel platforms and (2) emerging video- and image-based misinformation, the sensible combination of automated mechanisms with (3) human experts and (4) user-centered feedback to facilitate comprehensibility, (5) encouraging media literacy without misinformation exposure, and (6) adequately addressing particularly vulnerable users such as older people or adolescents.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; HCI theory, concepts and models;

Additional Key Words and Phrases: misinformation, disinformation, fake news, user intervention, countermeasure, media literacy

ACM Reference Format:
Katrin Hartwig, Frederic Doell, and Christian Reuter. 2024. The Landscape of User-centered Misinformation Interventions - A Systematic Literature Review. *ACM Comput. Surv.* 1, 1, Article 1 (January 2024), 55 pages. https://doi.org/10.1145/3674724

1 INTRODUCTION

The fast spread of misinformation is an enormous challenge both for society and individuals, with a great impact on democracy. Severe and fatal consequences can be observed about misinformation shared on social media related to COVID-19, with mistrust sowed in health measures required for combating a pandemic. With this in mind, Pennycook et al. [137] even go so far as calling it a 'matter of life and death'. In light of the grave consequences, the need for digital misinformation interventions to slow down its propagation is evident. Those technical approaches can be divided

[1] This work was supported by the German Federal Ministry for Education and Research (BMBF) in the project NEBULA (13N16361) as well as by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

Authors' address: Katrin Hartwig, hartwig@peasec.tu-darmstadt.de; Frederic Doell, frederic@doell-online.de; Christian Reuter, reuter@peasec.tu-darmstadt.de, Science and Technology for Peace and Security (PEASEC), Technische Universität Darmstadt, Pankratiusstraße 2, Darmstadt, Germany, 64285.

roughly into two main steps [143]: The (automatic) detection of misinformation and, second, the implementation of countermeasures as a concrete decision on what to do after successful detection. A great deal of research exists on detecting misinformation, which is often based on machine learning algorithms (e.g., [26, 79, 172]). Such algorithms are typically so-called 'black box' algorithms, which – while producing promising results with regard to detection accuracy – are not transparent in their reasoning. In order to make an algorithm's decisions transparent to users, interventions may profit from using 'white box' algorithms/explainable AI, which give greater insights into how the algorithm behaves and what variables influence the model [36]. The reliance on these automatic detection measures is increasing. For example, because of the COVID-related increase in traffic, Twitter (now X) has increased its use of machine learning and automation against misinformation [61]. Complementary research has been done on the implementation of concrete interventions after successful automatic detection (e.g., [18, 152, 157]). Those interventions are available in a wide range: Some aim at efficiently and automatically deleting content before exposure, while others try to educate users by showing corrections or flagging problematic content. While there is a large and heterogeneous field of interventions, they have a direct impact on end users of social media, as they focus on whether and how to communicate their output and findings, for instance, via information visualization. Although promising approaches have been established, the ongoing challenge of users being confronted and influenced by misinformation on diverse social media platforms such as TikTok, Twitter/X, Instagram, Facebook, and Co. suggests a need for further systematic design, implementation, and evaluation of effective digital interventions.

This review study aims to systematize knowledge on digital user-centered misinformation interventions. The term 'misinformation' is often used as an umbrella term for better readability, encompassing misleading information that has been created deliberately (frequently referred to as 'disinformation' or 'fake news') as well as unintentionally (frequently referred to as 'misinformation') [5, 36, 205]. For example, Li et al. [112] justify the use of misinformation as an umbrella term by stating that the majority of studies use this term to encompass different types of misleading information in general without denying the proper distinction between disinformation and misinformation. Indeed, misinformation and related phenomena such as rumors and conspiracy theories can all lead to severe consequences, even if those were not intended. Thus, in accordance with other research and systematic reviews [36, 112], in this paper, we will use 'misinformation' for better readability and to allow for a broader perspective on different kinds of misleading information while not denying the significant differences of phenomena. While the term 'misinformation intervention' has already been established by other researchers [15, 156, 157], we define *user-centered misinformation interventions* as digital countermeasures that go beyond a purely algorithmic back-end solution and exert a direct influence on the user in the form of information presentation or information withholding. Accordingly, we do not include approaches that deal exclusively with the automatic detection of misinformation without describing the subsequent communication to the user.

We provide a taxonomy that classifies and aggregates interventions regarding multiple relevant dimensions, such as time of intervention, addressed platform, and the thorough differentiation between intervention categories (e.g., correction, (binary) labeling, transparent indicators) to help identify promising research directions and encourage cross-disciplinary transferability. Researchers are faced with a very diverse research landscape on user-centered countermeasures, which is spread across multiple disciplines, such as computer science, human-computer interaction, information systems, psychology, communication sciences, journalism, and even medical research. Hence, we address the challenge of gaining an overview and help build on existing research while considering and learning from current insights of different relevant disciplines as well as research on different social media platforms. Thereby, we seek to facilitate informed decision-making of researchers and practitioners when analyzing, designing, and evaluating (novel) digital countermeasures to combat misinformation. We are especially interested in approaches communicating to users how an algorithm arrives at its results (e.g., white box algorithms) instead of giving a top-down answer (e.g., misleading, not misleading) without explanation. In our paper, we understand a ‘transparent’ intervention as an intervention that allows for informed decisions and the ability to comprehend why the content potentially contains misinformation, for example, via explanations of varying degrees, and can, thus, be considered as more user-centered than top-down interventions. There is evidence that transparently assisting users in their own assessment of misinformation is more promising than a top-down approach that provides social media posts solely with a label stating ‘This is/isn’t misinformation’ without cues to help comprehend the decision [104] or simply removes misinformation [10]. Research indicated that giving explanations or comprehensible cues can be significant to establish trust in the intervention [104], and counteract feelings of reactance or related backfire effects [128] that are controversially discussed in research [211].

While the topic of misinformation has been studied in systematic reviews, e.g., regarding specific contexts such as health [112] or political misinformation [94], existing literature reviews on interventions against misinformation and similar phenomena focus on more general overviews. For example, a related literature map by Almaliki [5] focuses on the research field of misinformation. It provides a general overview rather than analyzing the characteristics of concrete interventions and comparing the different approaches. They state that “less than 2% of the selected papers proposed digital intervention techniques”, while our focus lies on those studies that fall into the 2% as a promising subgroup of interventions with growing research interest. Furthermore, when literature reviews, or meta-analyses deal with concrete interventions, they often focus on the detection step and machine learning interventions (e.g., [29, 68, 124, 149, 212, 219, 220]) instead of user-centered interventions or focus on a specific subgroup like corrections [33, 147], warnings [123], accuracy prompts [138], or contexts like COVID-19 [91]. A first systematic overview of strategies against misinformation, including countermeasures with a direct influence on end users, was given by Chen et al. [36], who differentiate between five broad categories of solutions according to communication elements: message-based, source-based, network-based, policy-based, and education-based approaches. This contrasts our approach, which is not based on communication elements (e.g., message versus source) but considers interventions more in terms of their in-depth design. This design can be applied to the content itself, within a network-based approach, or to sources (e.g., by highlighting components in color as a passive intervention during exposure to misinformation). The authors give an overview of exemplary implementations within the four clusters. We build on that by providing an in-depth analysis of the design, interaction type, and timing of interventions as central aspects for user-centered implementations. In addition, we provide an overview of the methodological characteristics of intervention studies. Furthermore, Aghajari et al. [1] reviewed misinformation interventions with a focus on underlying driving factors of misinformation like social contexts and beliefs. Thus, in a constrained search process around the term ‘misinformation’, they categorize interventions according to content-based, source-based, individual user-based, and community-based strategies. Our study complements the analysis of strategies in terms of their driving factors (e.g., content-based strategies like corrections) with an HCI perspective detached from individualistic or community-based emphasis. Differentiating misinformation interventions on an individual trial system level, Roozenbeek et al. [151] review boosting interventions, nudging, debunking, and content labeling in comparison to interventions that rely on algorithms, business models, legislation, and politics. We complement the findings of related reviews by (a) shedding light on user-centered aspects of concrete misinformation interventions by performing an in-depth analysis of their design, implementation, and methodological evaluation for a broad perspective that offers a more comprehensive understanding of misinformation interventions. Thereby, we (b) specifically discuss and categorize characteristics impacting end users, such as the intervention design, user interaction, and timing of the intervention. We further complement the existing research landscape by (c) performing a review on publications of diverse disciplines and not limited to a specific period until 2024, searching three major databases. In doing so, we address calls for future work on a review including multiple disciplines and phenomena [1], and, when combined with findings of existing reviews, we provide a different perspective and more nuanced understanding of the research landscape. To our knowledge, a systematization of knowledge on specific user-centered misinformation interventions has yet not been conducted to this extent and with this perspective.

Our overarching goal is to deeply examine and classify misinformation intervention studies in terms of methodological characteristics in study design and evaluation, content characteristics of user interventions, and derived trends and challenges for future research. Addressing that goal, all our considerations lead us to the following research questions:

RQ1: *What are the typical methodological characteristics of existing studies on misinformation interventions?*
RQ2: *How do existing forms of user-centered misinformation interventions assist users in dealing with misinformation online?*
RQ3: *Which trends and chances for future research can be derived from the existing literature?*

The paper is structured as follows: First, we present our methodology of a systematic literature review and the procedure of deriving a taxonomy of user-centered misinformation interventions (see Section 2). Then, we present our results, including methodological aspects of analyzed publications such as addressed formats and platforms, applied methods of user studies, sample size, and participant details (see Section 3.1). Then we present our taxonomy (see Section 3.2), distinguishing nine intervention designs, active versus passive user interaction, and five points in time at which an intervention can be applied. We further discuss transparency as a specific measure to facilitate users in dealing autonomously with misinformation (see Section 3.3). Lastly, we present 'nudging' as a concept applied in many extracted publications (see Section 3.4). In Section 4, we answer our research questions regarding the design of user-centered misinformation interventions, methodological characteristics of existing studies, and the derived trends, open questions, and challenges for future research.

2 METHODOLOGY

In this section, we present our methodological approach of performing a systematic literature review, comprising of the identification and screening of relevant literature (Section 2.1) and the thorough analysis and structuring of publications and interventions included therein (Section 2.2).

2.1 Identification of Literature

To identify and categorize relevant literature on misinformation interventions, we performed a systematic literature review, following the PRISMA guidelines [130] (see Figure 1). Schryen et al. [164] state that literature reviews are important for "developing domain knowledge" and to identify knowledge-building activities, such as synthesizing, aggregating evidence, criticizing, theory building, identifying research gaps, and developing a research agenda. In accordance with these principles, we set up our literature search as follows: The initial search spans the ACM Digital Library, Web of Science, as well as the IEEE Xplore database. With this set of databases, we encompass a broad corpus of diverse literature as well as the ten conferences and journals listed by Google Scholar as the best regarding human-computer interaction: ACM Conference on Human Factors in Computing Systems (CHI), IEEE Transactions on Affective Computing, Proceedings of the ACM on Interactive Mobile, Wearable and Ubiquitous Technologies (IMWUT), Proceedings of the ACM on Human-Computer Interaction (PACM), International Journal of Human-Computer Studies

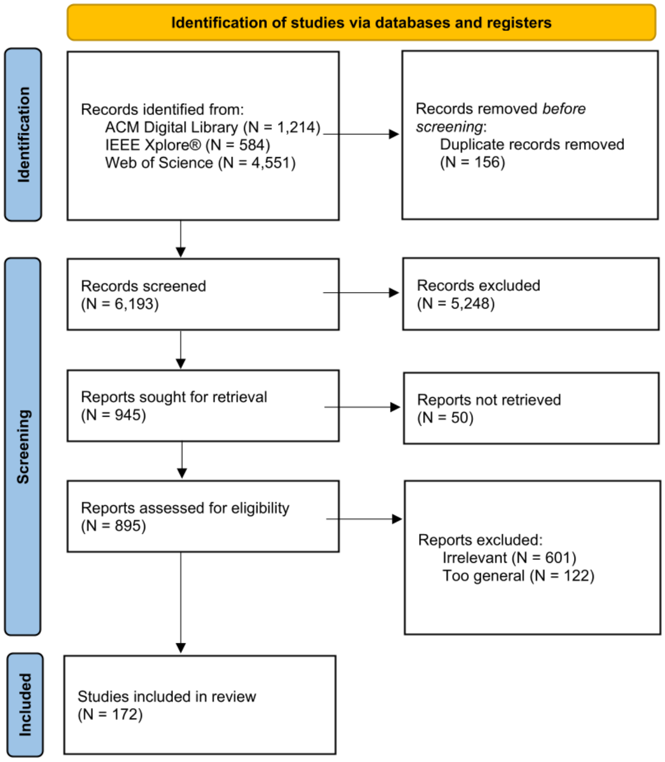

Identification of studies via databases and registers
Identification
Records identified from:
ACM Digital Library (N = 1,214)
IEEE Xplore® (N = 584)
Web of Science (N = 4,551)
Records removed *before screening*:
Duplicate records removed
(N = 156)
Screening
Records screened
(N = 6,193)
Records excluded
(N = 5,248)
Reports sought for retrieval
(N = 945)
Reports not retrieved
(N = 50)
Reports assessed for eligibility
(N = 895)
Reports excluded:
Irrelevant (N = 601)
Too general (N = 122)
Included
Studies included in review
(N = 172)

Fig. 1. PRISMA Diagram demonstrating our data flow within the systematic literature review (created with the template from [130])

(IJHCS), ACM/IEEE International Conference on Human-Robot Interaction (HRI), ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW), IEEE Transactions on Human-Machine Systems, Behaviour & Information Technology (BIT), and the ACM Symposium on User Interface Software and Technology (UIST). The search took place until June 2024, and only papers that were published until that date could be considered. All publication years up to this date have been included, however, there were no relevant publications for our final set concerning content-wise inclusion and exclusion criteria before 2011.

The search term consists of two parts: The first part are terms included in or related to our umbrella term of 'misleading information'. The second part includes synonyms for 'intervention' and related concepts addressing user-centered measures. Only papers containing at least one term of each part in their title or abstract were included in our search. We did not filter for a publication year. The complete search term is the following:

((rumour OR rumor* OR "misleading information" OR "fake news" OR "false news" OR misinformation OR disinformation OR "news credibility") AND (combat* OR correct* OR interven* OR countermeasur* OR counteract* OR treatment OR relief OR educat* OR warning OR nudg* OR user-centered OR "media literacy"))*

The term was modified to adhere to database requirements and to run comparable searches. Furthermore, because Web of Science returned many results, the term was adjusted to exclude obviously irrelevant disciplines (e.g., chemistry).

The broad interdisciplinary nature of Web of Science explains its large amount of 'false positives' during the initial search in comparison to the other two databases that already focus on disciplines relevant to digital misinformation interventions (e.g., computing and information technology). The search returned 1,214 results from the ACM Digital Library, 584 results from IEEE Xplore and 4,551 results from Web of Science, in total 6,349 results. After removing 156 duplicates, we screened 6,193 records, of which we excluded 5,248.

Records were removed for a multitude of reasons: First, some words of the search term had multiple meanings, which is why papers using a different interpretation were excluded. While we decided to take a broad perspective on diverse kinds of misleading information, including misinformation, disinformation, rumors, and related phenomena (e.g., conspiracy theories), as all types can have severe consequences, papers were excluded when the phenomenon investigated was not referring to our definition of the umbrella term 'misinformation' (see Section 1). This was particularly the case for misinformation referring to eyewitnesses remembering something inaccurately (e.g., due to suggestion) during a testimony in court. For phenomena included in our broad definition, there was a variety of terms included in our sample (see row 'Concept' in Table 2). Furthermore, some technical terms have different meanings in different fields, for example, network science uses the term rumor in the context of nodes spreading information (e.g., [27]). Second, when papers concentrated solely on the technical detection step with no involvement of the user at all, e.g., machine learning approaches focusing on increasing the detection rate, they were excluded. Furthermore, interventions that took a network-based approach, for example, by simulating which nodes to delete in order to reduce the spread of misinformation, were also excluded. In addition, we excluded psychological experiments without concrete reference to misinformation as well as surveys and questionnaires exploring background information (e.g., Which demographics are susceptible to misinformation?). Additionally, we decided to exclude reviews.

945 papers were sought for retrieval, of which 50 were removed because they could not be accessed, leaving us with a total of 895 publications that were assessed for eligibility. In the last step, a total of 723 papers were finally sorted out, 601 thereof because of the aforementioned criteria. Another 122 papers were excluded because they were too general, including papers that did not meet this review's focus because the intervention occurred long before the actual usage of social media, like educational school lectures, trainings, or serious games, but also implementations that focus exclusively on psychological phenomena (e.g., Do corrections of content generate reactance?). Particularly in the context of corrections or debunking of misinformation (e.g., in comment sections), there is a lot of in-depth research on factors impacting user reactions and interactions. Often, these studies focus on psychological or social phenomena that are particularly valuable to consider when designing interventions tailored to a specific persona. To receive a reasonable number of publications and thus allow for a thorough focus on research regarding the design and evaluation of digital interventions, we decided to exclude studies that rather address (psychological or social) impact factors without a particular focus on intervention design and evaluation. The final set of papers contained 172 items which were included in our analysis and were categorized according to our taxonomy.

2.2 Development of a Taxonomy

For the development of the taxonomy, we first collected different relevant dimensions to compare and differentiate studies on user interventions within the context of misinformation. We developed and applied those categories in an iterative process of brainstorming sessions with two researchers with expertise in computer science, psychology, and human-computer interaction based on already familiar studies within the field of interest (e.g., [18, 104]). The coding process was initiated by a training phase where a common understanding of each category was obtained. When disagreeing on a categorization during the coding phase, the study was discussed to achieve a consensus. This approach of consensus coding is commonly applied in other research [209]. First, we defined our target group: researchers and practitioners interested in analyzing, designing, and evaluating digital countermeasures to combat misinformation that may potentially benefit from our taxonomy. To assist the target groups, several characteristics are particularly relevant as they provide information on (1) the intervention design, (2) the form of user interaction, and (3) the timing of the intervention. Categories were complemented and adjusted iteratively while identifying and reading new papers. For instance, when reading multiple papers that differed regarding the time of intervention, this category was included, and all relevant papers were categorized accordingly. Additionally, minor modifications to the categories were made during the process of reading and categorizing the articles when deemed necessary. Further, we looked at how user-centered interventions were categorized in other contexts with sensible information (e.g., cybersecurity) in systematic reviews [58]. The resulting final table can be found in the electronic supplement (see Table 2). A study can be sorted into several categories, and the subcategories are generally not mutually exclusive (e.g., some interventions may combine the intervention categories 'highlighting design' and '(binary) label' and others compare a 'correction' with 'showing indicators').

3 RESULTS: THE LANDSCAPE OF USER-CENTERED MISINFORMATION INTERVENTIONS

In this section, a detailed analysis of the literature review is presented. First, we give an overview regarding *methodologies* used by studies on user interventions (see Section 3.1). We then provide a *taxonomy of interventions to assist users in dealing with misinformation* by categorizing and clustering the identified research sample in distinct dimensions (see Section 3.2). Furthermore, we highlight how *transparency* (see Section 3.3) is used to assist users in dealing autonomously with misinformation, present the concept of *digital nudging* (see Section 3.4) as a trending digital countermeasure, and finally discuss the impact and perceptions of reviewed misinformation interventions (see Section 3.5).

3.1 Methodological Characteristics

To provide an overview of research methods typically used in the field of user-centered interventions to assist in dealing with misinformation, details of the respective study designs were collected. All studies were published between 2011 and 2024. First, we were interested in the different concepts included under the umbrella term 'misleading information'. In total, 149 publications referred to either *misinformation, disinformation or misleading information*. Furthermore, 5 publications were specifically interested in *rumors*, and 10 publications that addressed the concept of *news credibility*. Other publications referred to myths, propaganda, or controversial topics. Out of a total of 172 included papers, 17 present exclusively *conceptual* ideas of interventions. In contrast, the remaining studies collected empirical data in the form of *laboratory experiments* (14 publications), *online experiments* (106 publications), *field studies* (9 publications), *surveys* (28 publications), and *interviews* (20 publications). In our study, we understand a field study as an evaluation type that specifically observes the natural behavior of participants in a real-world scenario, in contrast to experiments that encompass a controlled setting designed by the researchers. In the context of misinformation, research experiments rarely take place within an actual lab of the researcher (lab experiment) but typically remotely in an online setting (online experiment), for instance, as a link to the researcher's experimental website, as these experiments often do not require physical presence for additional hardware items. In some cases, there is a combination, e.g., of survey and interview or of laboratory experiment and online experiment within one publication. Regarding sample size, the empirical studies range from small groups of participants (<20 e.g., [25, 32, 62, 106]) to large-scaled representative groups with far over 1,000 participants (e.g., [11, 95, 165]). You can find a visualization of sample sizes in Figure 2. A closer look at the participants reveals a clear bias, with the majority of students reporting having U.S. adults and college

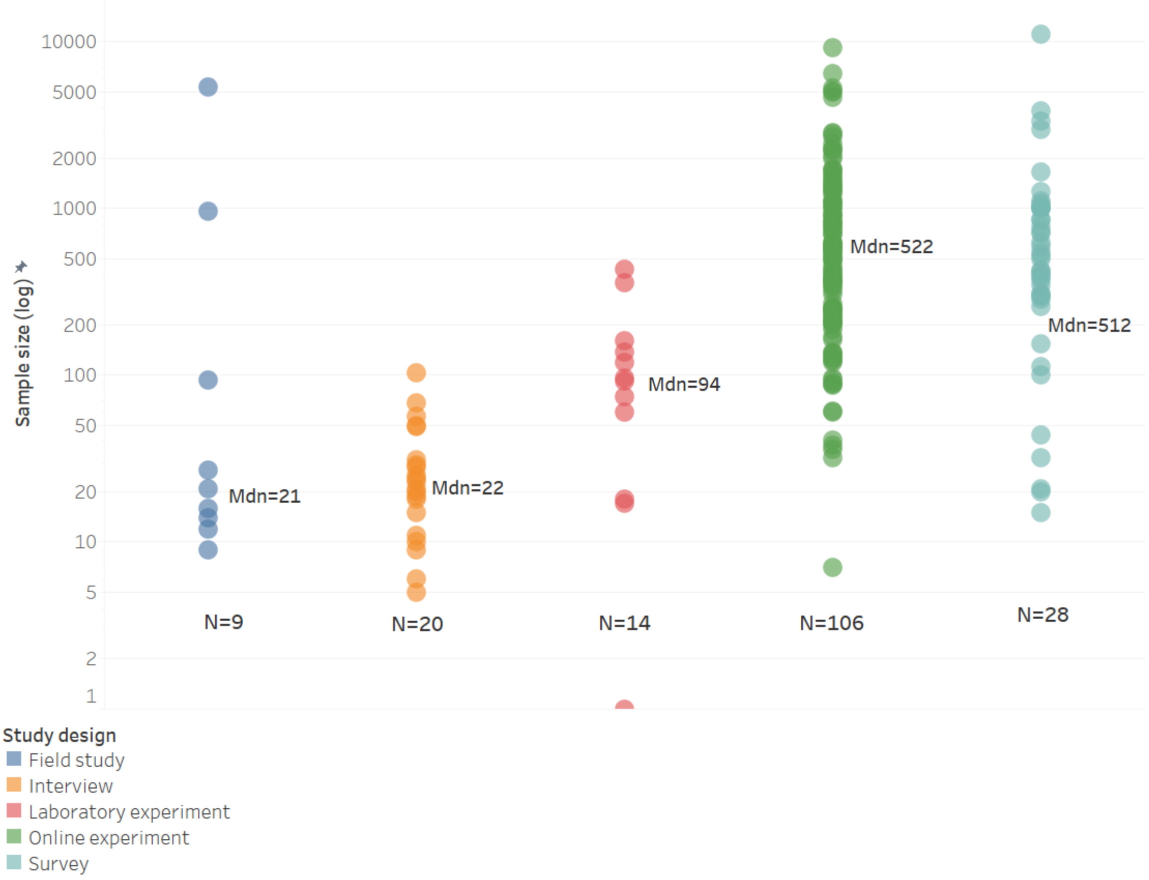


Fig. 2. Sample sizes (log) of the individual studies broken down by study type including the median (Mdn). Note that a publication often contains multiple user studies.

students as participants. However, there are also isolated studies that either address a very specific (vulnerable) target group (e.g., teenagers [12, 74], marginalized communities [191], or blind/low vision social media users [168]) or make a comparison between several countries (e.g., [4]).

While 71 publications generate their interventions or concepts generically for all online content and platforms (category *General*), others are developed and evaluated for specific platforms (see Figure 3 for temporal distribution regarding platforms). Nevertheless, transferability to other platforms is often not excluded. 36 publications address interventions for *Facebook*, 33 publications for *Twitter/X*, and 3 publications for *Instagram*. Another 21 publications deal with platforms that do not fall into one of the categories already mentioned (e.g., Reddit [24, 201], TikTok [69, 74], messengers like Telegram [76], websites [111], arguments over an audio speaker [42], text documents [60], messenger forwards [134]) and therefore were categorized as *Other*. This corresponds to known research biases that show a focus on much-researched platforms such as Twitter/X. This is often justified by the already developed data situation and easier linkage to existing literature. Especially the great relevance of misinformation on newer social media platforms

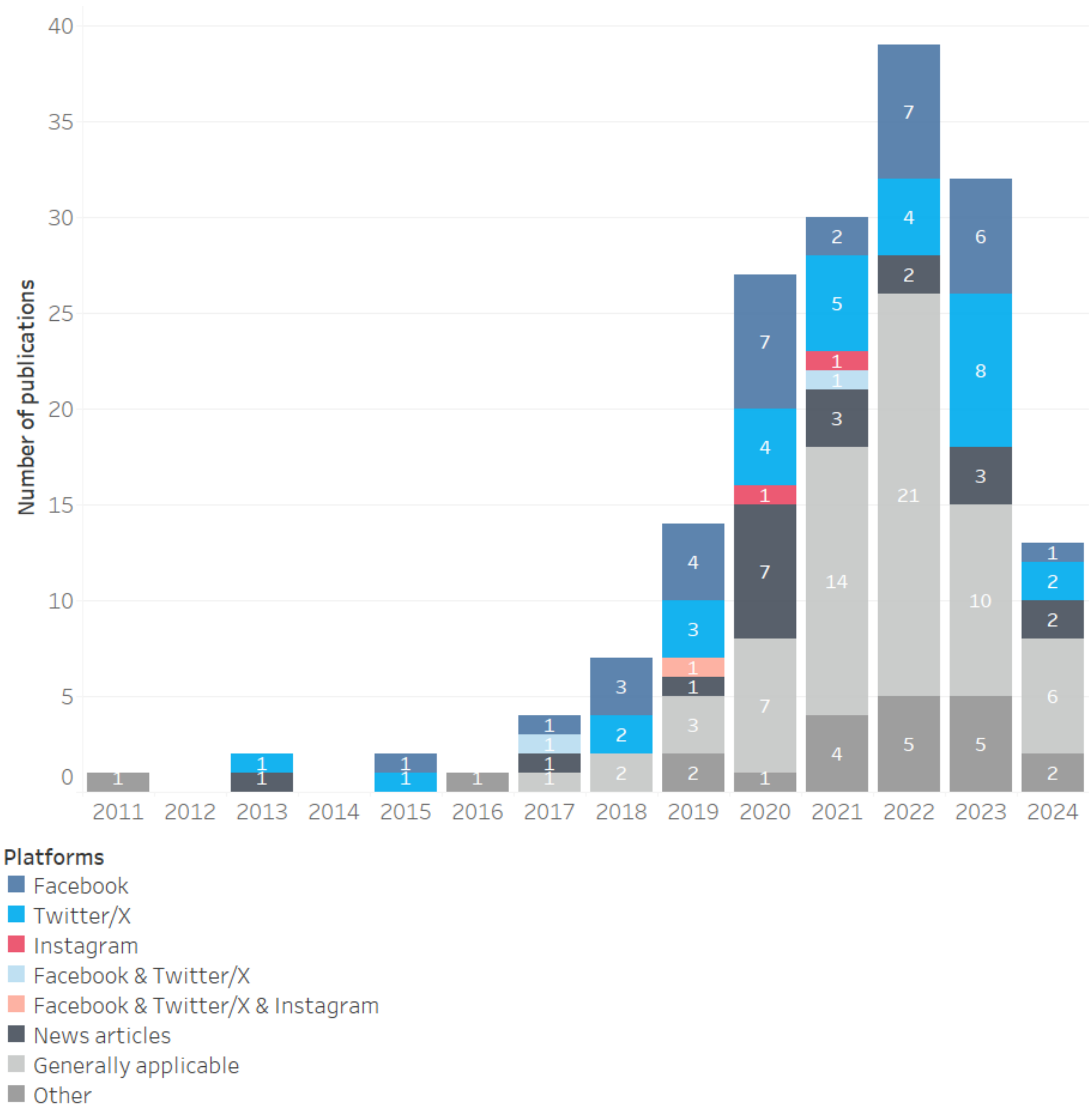


Fig. 3. Number of papers published according to their addressed platform.

like TikTok in crises like the Russian-Ukrainian war shows that there is still a great need for research. When looking more closely at the addressed content format also see Figure 9 in the Appendix), we can see that most publications focus on *social media posts* (91 publications), 49 on *articles or text in general* and only a few on *images* (8 publications) and *videos* (9 publications) while we observe a growing relevance of misinformation of exactly these formats. Additionally, there are a few exceptions that address a very specific format, such as *audio* (e.g., [42]) or misleading graphs [208].

A minority of the publications describe interventions that go beyond a low-fidelity prototype (e.g., in the form of screenshots) to include an actual implementation. Thus, *no implementations* can be read from 123 publications, while 14 publications deal with *browser plugins or browser extensions* (e.g., [18, 99]). Fourteen publications describe the implementation of a *custom platform* (e.g., [6]), and 3 publications show a *game-based implementation*. Elaborate tools are an important part of mitigating the spread of misinformation and can be part of a holistic solution. An example is 'Verifi!' [97], which provides an interface for dealing with misinformation on Twitter (now X). The system consists of five display options, allowing for easy comparison between how real and questionable news sources report on a subject, for example, by comparing the words or images used. Another example would be 'Prta' [121], which provides the user with a tool that takes a text or URL as input and highlights propaganda techniques.

3.2 A Taxonomy of User-centered Misinformation Interventions

The wide range of addressed concepts, platforms, and research areas shows that, on the one hand, a large number of conceptual ideas and empirical findings already exist for digitally supporting users in dealing with misinformation; on the other hand, these often differ fundamentally. In order to distinguish existing approaches from each other and to cluster commonalities, we have derived a taxonomy based on the identified literature. Therefore, we performed an in-depth analysis of interventions. For the interpretation of the following results, it is important to notice that publications can contain multiple interventions – in total, there were 237 interventions within the 172 publications. Those interventions were analyzed individually regarding the taxonomy characteristics, while previously reported methodological findings are valid for the entire publication and, therefore, did not distinguish between individual interventions. In the following, the literature-based categories of the taxonomy are explained in detail (see also Table 1 and Figure 4):

3.2.1 Intervention Design. The identified interventions on user-centered misinformation interventions vary greatly in their starting point. The digital support approaches and concepts are as diverse as the possibilities for protecting users from the effects of misinformation (e.g., deleting problematic content, warning, or strengthening media literacy). In an iterative process, nine intervention designs were identified based on the literature. Interventions could be assigned to multiple intervention designs, as they often used combinations. The majority of interventions propose or evaluate *correction/debunking* of misleading contents (66 stand-alone interventions and 30 interventions in combination with other intervention designs within 80 publications) and often represents a quite natural behavior of social media usage rather than an artificially generated technical countermeasure. For instance, many publications in that context evaluate whether corrections by users in the comment section of a post are effective in reducing belief in misleading content (e.g., [120]). Some of those interventions include a link to fact-checking websites, where the misleading content is debunked. This can be implemented both naturally by users in the comment section posting debunking links or digital interventions automatically exposing users to debunking (e.g., link to correcting source or user rebuttal within a comment/reply to a social media post or exposure to automatically generated counterfactual explanations [40]). Thus, interventions vary in terms of who is the arbiter of credibility assessment. While some are expert-based or rely on algorithm decisions, others rely on crowdsourcing of the community [47]. For instance, a browser extension allows users to suggest alternative headlines as a crowdsourced odd case for corrections, which are then presented to other users, empowering them to more actively participate in news consumption [88]. The dimension of who decides what is wrong or right within corrections and other intervention types was not systematically covered by our taxonomy but constitutes a relevant research area that has already been addressed by several studies [73, 88, 197]. Many interventions

Table 1. A taxonomy of user-centered misinformation interventions.

Category	Definition	Intervention examples	Publications
Intervention design	The intervention design distinguishes different countermeasures after the successful detection of misinformation. This includes general actions such as deleting content as well as concrete design decisions to encourage a learning effect.		
Warning	Interventions that give an explicit warning that the content is (potentially) misleading	Warning label; stop sign; "This post was disputed"	[7, 8, 16, 18, 20, 28, 38, 45, 56, 57, 59, 62, 64, 69, 70, 72, 92, 104, 105, 107, 109, 115, 116, 122, 126, 127, 129, 131, 133, 135, 142, 145, 157, 161, 165, 166, 168, 170, 188, 192, 202, 208, 216]
Correction/debunking	Interventions that correct/debunk misinformation	Naturally occurring or artificially generated user comments or comments from officials that correct misinformation; Links to fact-checking websites; expert sources; corrected headlines by users	[3, 11, 14, 16, 22–24, 31, 32, 39–41, 43, 44, 47, 49–52, 63, 69–71, 82, 83, 88, 88, 90, 98, 102, 104, 107–110, 113, 114, 117, 118, 120, 125, 129, 134, 140–142, 150, 155, 159, 160, 162, 169, 174–180, 182, 183, 185, 186, 188, 189, 191, 193–200, 202, 204, 207, 208, 210, 216, 217]
Showing indicators	Interventions that display indicators for misinformation to achieve transparency	showing how old a video actually is; color relevant words for misinformation classification; generic tips to detect misinformation; infographic	[13, 14, 19, 24, 31, 37, 46, 55, 60, 65, 67, 70, 72, 74, 76, 77, 80, 96, 97, 103, 105, 111, 121, 145, 146, 153, 155, 163, 169, 170, 192, 201, 208, 213, 214]
(Binary) labels	Interventions that label content as misinformation or true information; often binary	Tagging post as true or false; thumbs up/thumbs down; "Prediction: It is Fake News!"; traffic light symbols	[9, 17, 24, 25, 45, 51, 56, 64, 78, 86, 89, 95, 99, 107–109, 111, 115, 116, 119, 126, 131, 132, 134, 135, 142, 146, 155, 157, 161, 163, 168–170, 176, 182, 185, 190, 201, 216]
Highlighting design	Interventions that visually highlight relevant parts of a post for misinformation classification	highlight relevant words by color or size; color code tweets according to accuracy; highlight propaganda techniques using colors	[6, 9, 18–20, 60, 74, 76, 77, 80, 86, 89, 96, 103, 111, 113, 121, 153, 169, 170, 182, 208, 213, 218]
Visibility reduction	Interventions that reduce the visibility of misinformation visually	Reducing opacity or size	[8, 18, 20, 69, 103, 104, 107, 142, 157, 170]
Removal	Interventions that hide or remove misinformation	deleting misinformation	[157]
Complicate sharing	Interventions that include additional user effort before allowing to share misinformation	additional confirmation before sharing; require users to assess accuracy before sharing	[6, 87, 103, 104, 192]
Specific visualization	Interventions that use creative visualizations of relevant information	visualizing sentiment and controversy score of news articles; visualizing fact-checker decisions; platform based on social network analysis visualization; aggregate authentication measures; visualization of number of unvaccinated children with measles as fear correction; visualizing fact-checker decisions; infographic	[34, 46, 85, 88, 88, 89, 93, 97, 98, 101, 111, 133, 140, 144, 146, 148, 154, 163, 170, 203, 214, 216, 218]
User interaction	Interventions require varying degrees of interaction with the countermeasure		
Active	Active interventions require users to actively interact with a countermeasure	click to confirm before sharing; overlay on Facebook/Twitter (now X) post; pop-up	[2, 6–8, 28, 31, 35, 45, 57, 64, 69, 72, 87, 89, 99, 103, 104, 107, 109, 137, 142, 148, 157, 162, 167, 184, 192, 195]
Passive	Passive interventions can be potentially ignored while using social media	A label below a social media post; tooltip; correction in comment section; warning next to a post	[3, 9, 11, 13, 14, 16, 18–20, 22, 23, 25, 28, 38, 39, 41, 43–52, 55, 56, 59, 60, 62–64, 67, 69–71, 74, 76, 77, 80, 82, 83, 85, 88, 88, 90, 92, 93, 95, 96, 98, 100, 102–105, 108–110, 113–117, 119, 120, 122, 125–127, 129, 131–136, 139–141, 144–146, 148, 150, 153, 154, 157, 159–162, 165–170, 174, 175, 177–180, 182, 183, 185–189, 191, 193–195, 197–199, 199–202, 204, 207, 208, 210, 213, 214]
Timing	Digital misinformation interventions can address varying points in time within the social media usage		
Pre exposure	Interventions that take place immediately before the exposure to (mis)information	accuracy nudge before a news-sharing task; Pro-Truth pledge to engage in more pro-social behavior; narrative fear appeal message to encourage health experts to correct health misinformation online; general warning message about misleading articles; generic infographic; protective message: "Warning! Note: fake news can occur on Facebook. [...]"	[2, 25, 28, 31, 35, 38, 39, 41, 45, 46, 57, 64, 67, 69, 70, 72, 97, 103, 104, 109, 126, 136, 137, 148, 161, 162, 179, 184, 187, 195, 196, 210]
During exposure	Interventions that take place during the usage of social media (and during the exposure to misinformation)	algorithmic corrections next to a post; user comments underneath a post; warnings; adding "Rated False" tag to article headline; credibility labels; highlighting indicators for propaganda; wearable reasoner giving AI-based feedback on claims	[3, 6–9, 13, 14, 16, 18–20, 22–25, 38, 40, 42, 45, 47, 48, 51, 55, 56, 62, 64, 69–71, 76, 77, 80, 81, 83, 85, 86, 88, 88–90, 92, 95, 96, 100, 102, 104, 105, 107, 109, 110, 113, 115, 116, 118, 119, 122, 125–127, 129, 131–133, 135, 140–142, 145, 146, 153, 161, 165–168, 170, 174, 175, 177, 182, 185, 186, 193, 194, 196–198, 201, 202, 208, 213, 214, 216, 218]
Post exposure	Interventions occurring after seeing misinformation	warnings after exposure to misinformation; responses by official health authority; debunking text based on debunking handbook including plausible scientific explanations to close gaps in mental models	[11, 25, 28, 39, 41, 43, 44, 49, 50, 52, 59, 63, 64, 82, 98, 108, 110, 117, 119, 120, 134, 136, 139, 150, 159, 160, 162, 178–180, 183, 188, 189, 191, 195, 199, 200, 204, 207, 210]
At the moment of sharing	Interventions taking place directly at the moment of sharing misinformation	Encouraging to reflect on content before sharing; endorsing accuracy prompt: "I think this news is accruate" placed into sharing button; behavioral nudges using checkboxes to indicate whether a heading is accurate and to tag reasons via checklist at posting time; report of linguistic analysis as immediate feedback when sharing and possibility to cancel a tweet within 30 seconds	[6, 30, 87, 89, 104, 192]
On request of the user	Interventions that take place detached from social media platforms and have to be reached out to actively	web-app based on social network analysis for user exploration; Android application where user can enter URL or text for credibility assessment; system to analyze articles or URLs via interface or API	[17, 34, 37, 74, 78, 93, 97, 99, 101, 103, 111, 121, 144, 154, 155, 163, 169, 176, 190, 203]

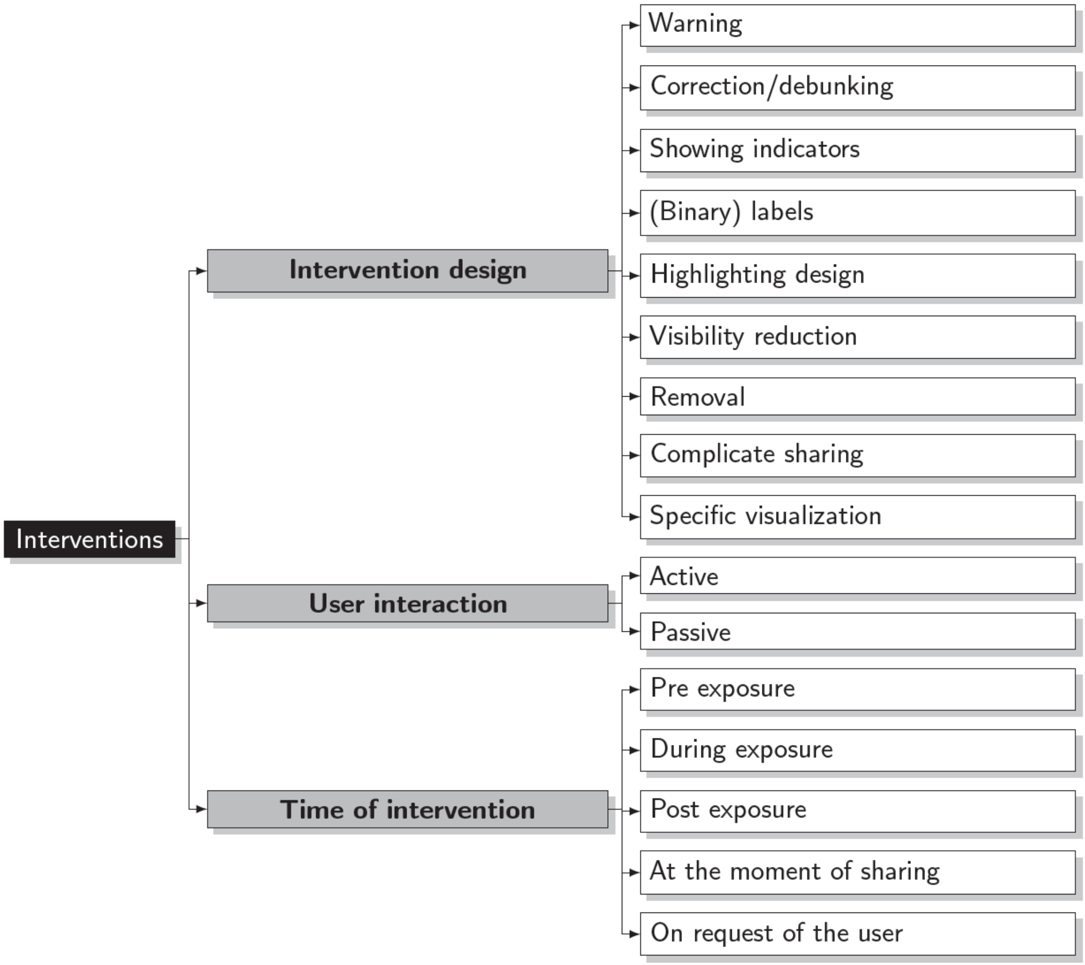


Fig. 4. A taxonomy for user-centered misinformation interventions.

do not only expose users to content debunking or rebuttals but give an explicit *warning* that the content is (potentially) misleading (19 stand-alone interventions and 39 interventions in combination within 42 publications). Those warnings reach from warning labels like stop signs to textual warnings, e.g., “This post was disputed!”.

Misinformation interventions can have different objectives. One of these objectives is to strengthen media literacy. In these types of interventions, concrete assistance in the form of indicators, for example, is typical. By *showing indicators* that support users in evaluating the credibility of content, the aim is to achieve a learning effect. Thirty-seven interventions correspond to this intervention design (including 7 as stand-alone interventions; 35 publications), for example by showing how old a video actually is [170] or by deriving words in the text that were particularly relevant for automatic detection as misleading [13]. Other intervention designs of this type compile more generic tips that users can apply to detect misleading content [46, 67, 70]. Research on indicators for misinformation, for instance, from the perspective of journalists as annotators [215], can be considered a significant foundation to inform indicator-based interventions. This intervention design is especially relevant, as studies have shown that users prefer transparent approaches where there is a potential learning effect [104]. However, in contrast to showing indicators for misleading content in a nuanced way, 48 interventions take the approach of assigning *(binary) labels* to contents (including 12 stand-alone interventions; 41 publications). This can be implemented, for example, by tagging content with a true or false tag, or thumbs up, thumbs down [17]. Other interventions give a probability in percent that content is misleading [99], and thus extend the framework of binary labels with richer information or provide more nuanced labels e.g., using traffic light colors. Often, these labels have in common that they do not provide a transparent explanation. However, binary and more nuanced labels do not, per definition, rule out a user-centered approach as they may be sensibly combined with explanations and can be applied for simplification of a complex underlying rating system.

Similarly, other ideas are concerned with increasing transparency. Some approaches use *highlighting design* as an intervention design which aims at facilitating potential learning effects. Twenty-four interventions (24 publications), for example, visually highlight relevant words for automatic classification within a social media post by color or size. For instance, Bhuiyan et al. [18] color code tweets on Twitter (now X) according to their computed accuracy. In contrast, Martino et al. [121] highlight propaganda techniques (e.g., exaggeration, loaded language, or oversimplification) detected within a text using different colors. Often, interventions that show indicators use some kind of highlighting design to do that, resulting in a common combination of both intervention designs. Indeed, despite highlighting components of content with colors, more *specific visualizations* (23 interventions; 22 publications) can be considered a distinct intervention design. Visualization is a very effective way to provide information as it provides *"the highest bandwidth channel from computer to the human"* [206, p. 2] and is used for interventions in different contexts [75]. Within the literature sample, there is a diverse set of creative visualizations of information. For example, Kim et al. [101] visualize the sentiment and controversy score of news articles within a conceptual study. In contrast, Park et al. [133] visualize fact-checker decisions regarding textual rumors. Moreover, Schmid et al. [163] developed and evaluated a platform based on social network analysis of contents on Twitter/X for users to proactively assess misinformation through visualization, and Chen et al. [34] designed visualizations of filter bubbles, exposing users visually with topics, sources, and opinions outside of their own bubble.

While previous intervention designs tended to provide verified feedback together with the problematic content, one study evaluates the effect of the *removal of misinformation* by hiding or removing a questionable post altogether [157]. Similarly but less rigorous is the attempt of *visibility reduction* (14 interventions in 10 publications), e.g., by reducing opacity or size. While we focus on visibility reduction that takes place visually, there are other (often network-based) approaches not covered by our more narrow understanding of user-centered misinformation interventions, as long as a user study demonstrating a direct impact on users is not included. For instance, studies reduce the visibility of misinformation in an algorithmic approach by reducing its flow [84]. Similarly, Epstein et al. [54] examine how layperson crowdsourcing of source credibility may be applied as input to social media ranking algorithms with promising results, leaving the potential for future research to investigate how this approach may be implemented regarding user feedback. Many intervention designs aim at preventing negative effects on people when confronted with misinformation or educating them to detect those contents themselves, as presented in the previous intervention designs. However, this can be extended to specifically preventing the spread of problematic content altogether, for instance, via *complicating sharing* (5 interventions in 5 publications). This may be implemented, for example, by including an additional confirmation before sharing or by nudging users to assess the accuracy of the content as they share it [87].

These user-centered approaches stand in contrast to network-based approaches to prevent the spread of misinformation through computational techniques as described by Chen et al. [36].

Many publications used multiple intervention designs, for instance as combinations or comparing interventions of different types against each other [70]. For example, showing indicators for misinformation often comes with some sort of highlighting design or a specific visualization (of the indicators).

While most interventions could be assigned to one or more of the intervention designs listed above, 50 interventions additionally used an intervention design that did not clearly fit the scheme (including 30 stand-alone interventions; 39 publications) while not appearing often enough as a distinct intervention design to represent an own type within the taxonomy. There are particularly unusual approaches, such as the development of wearable glasses that provide audio feedback on the truth of content [42] or a study that evaluates the effect of priming participants by letting them rate the accuracy of a headline before exposure to more potentially misleading content as a nudge to think more sufficiently [137] or on a similar basis, an explanation prompt that lets users explain why headlines were true or false [139]. Among the *other* category, there are nine interventions giving diverse kinds of information about misinformation and its detection immediately before exposure, e.g., in the form of an infographic [2], a video tutorial [12], a text about negative consequences of misinformation [38], a debiasing message [41], an awareness training [126], or a Pro-Truth Pledge [184]. Three interventions display a star rating or score, for instance, regarding credibility or sentiment [48, 100, 101]. Two interventions explicitly state to have integrated gamification elements [4, 176]. You can find a visualization of intervention designs in Figure 5.

3.2.2 User interaction. We compared whether an intervention required users to *actively* interact with the countermeasure (e.g., having to click to confirm sharing) or whether they could *passively* ignore the countermeasure (e.g., a label below a post). Furthermore, some interventions cannot clearly be labeled as active or passive as the actual form of implementation is not (yet) defined (30 interventions in 27 publications). For instance, some approaches make a more generic proposal of a warning without stating if it can be ignored or not. Other approaches present an intervention that takes place only at the request of the user, for example, as a separate smartphone app (e.g., [17]) and, thus, is neither active nor passive during the actual usage of social media. In cases where the intervention is designed as a mandatory one-time experience (e.g., [2, 137, 184]), users do not have to interact with the intervention during the following usage of social media but definitely once before the usage. Thus, those interventions are classified as active. We found that the majority of interventions are passive (175 interventions in 129 publications) while only 32 interventions (in 28 publications) deal with active interventions. A representation of the number of interventions regarding different interaction types can be found in Figure 6.

3.2.3 Time of intervention. Misinformation interventions can address different points in time in the context of social media use. While countermeasures that take place long before the actual usage of social media (e.g., trainings, educational games for school lessons, inoculation) were excluded from analysis, we included countermeasures that take place immediately before the exposure (e.g., short messages when logging in to a social media platform). We found that 37 interventions take place *pre-exposure* to misinformation (including 5 interventions coming with a combination of timings; 32 publications). For example, Pennycook et al. [137] nudged participants of a large-scale online study to think about accuracy before getting a news-sharing task. They asked participants to rate a headline's accuracy before exposing them to multiple other headlines and measuring their sharing intentions. Indeed, they found that giving a simple accuracy induction resulted in increased sharing discernment [137]. The majority of interventions (120 interventions in 92 publications) however, is designed to take place immediately *during* the exposure to misinformation while users

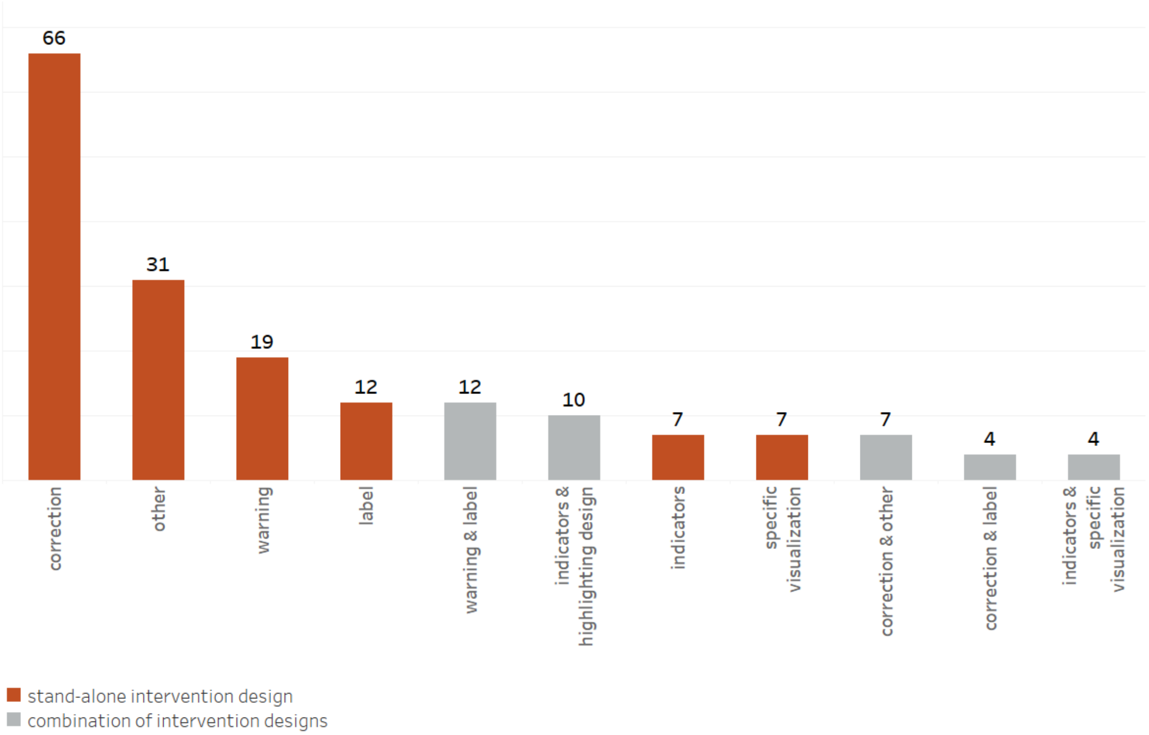


Fig. 5. Number of interventions addressing common intervention designs while differentiating between approaches that use a single intervention design versus multiple types in combination. Instances <4 were excluded. A more detailed breakdown of the category 'other' can be found in Section 3.2.1.

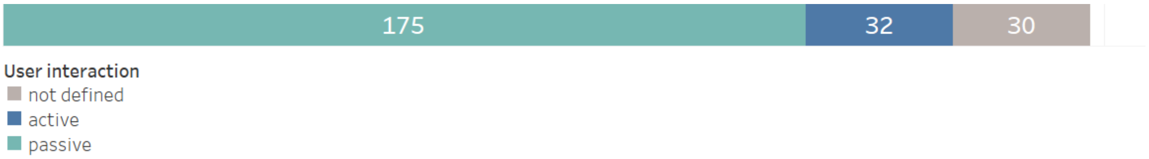


Fig. 6. Number of interventions regarding each type of user interaction. Publications can contain multiple interventions with different user interaction types.

are engaging with a social media platform or other content and encounter misinformation. This includes most of the corrections, warnings, labels, and highlighting approaches. For example, Bode and Vraga [23] compare algorithmic corrections via related articles from Snopes with social corrections via user comments underneath a Facebook post referring to the same debunking link on Snopes.

Forty-nine interventions in 42 publications deal with countermeasures *post-exposure*, e.g., Grady et al. [64] compared warnings after exposure to an article with warnings during other points in time. To specifically combat the spread of misinformation, a few interventions intervene directly *at the moment of sharing* (6 interventions in 6 publications; e.g.,

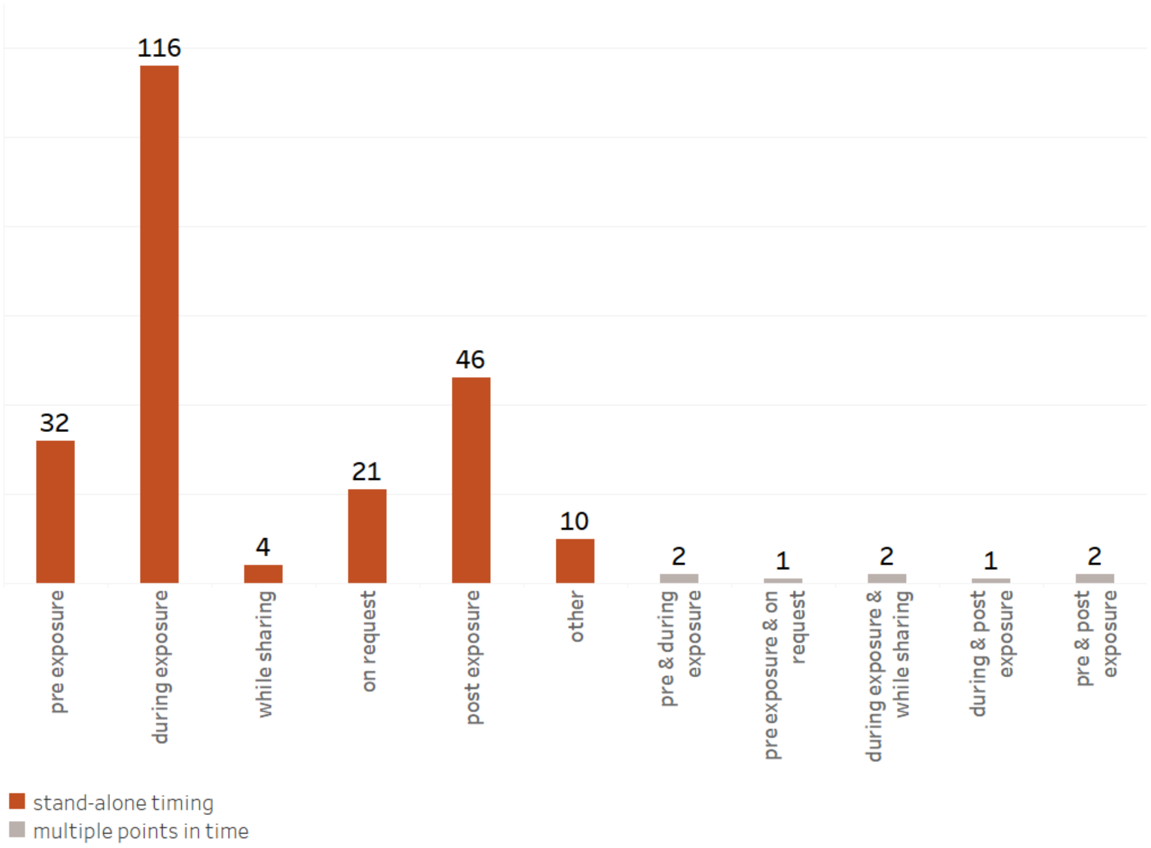
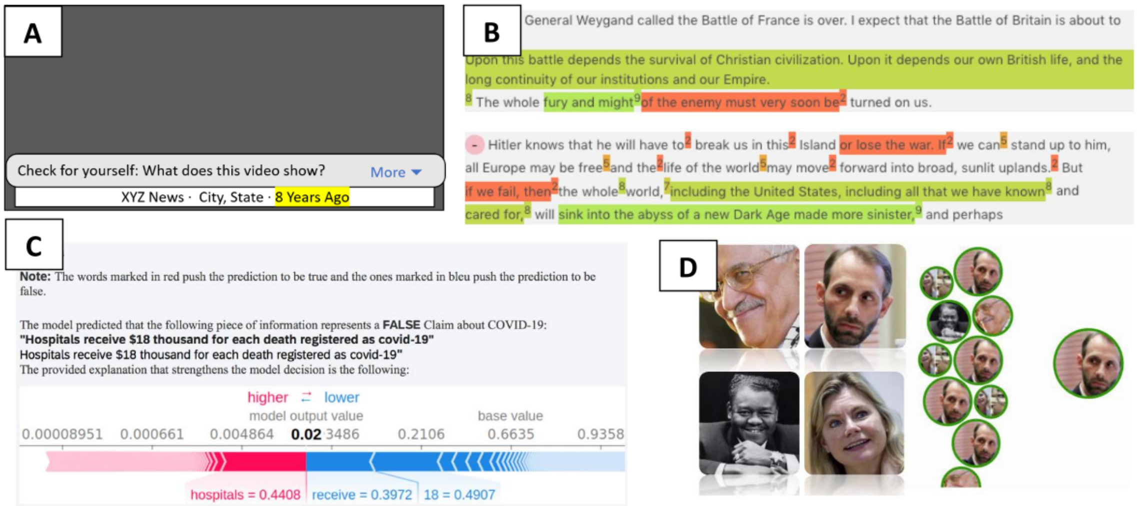


Fig. 7. Number of interventions addressing different points in time.

[192]). These presuppose that the user is about to share misinformation and he or she has already slightly passed the timing dimension 'during exposure'. Detached from the actual social media platforms, some approaches offer their own platforms. Accordingly, the intervention here takes place *at the request of the user* at any time (21 interventions in 19 publications; e.g., [17]), when users have to actively reach out to the intervention on a separate platform. A few interventions could not be assigned to one of those points in time (10 interventions in 8 publications). For example, Furuta and Suzuki [60] present a countermeasure to take place during article creation. While the majority of interventions specifically take place at one exact moment in time, eight interventions are designed to take place at multiple points in time, combining, for example, a pre-bunking message with a warning during exposure to misinformation. See Figure 7 for a visualization of the number of publications for each timing of the interventions.

3.3 How is transparency used to facilitate users autonomously dealing with misinformation?

As highlighted in Section 3.2.1, digital countermeasures in the context of misinformation can have different objectives. While some interventions aim to reduce the spread of such content itself, others aim to communicate the findings of the digital countermeasures to the end users, sometimes resulting in an environment that facilitates the strengthening of media literacy skills. Within our systematic literature review, our special focus of interest is on transparent approaches that offer some form of explanation, as opposed to (binary) labels without explanations or deletion of problematic content. Transparency can be achieved by different distinct intervention designs. One very common way of explanation is exposure to corrections or debunking. Indeed, the majority of publications within our scope deal with some sort of correction or debunking. While some approaches investigate the effect of user corrections within comment sections of social media posts, others focus on corrections of authorities. Interestingly, there is a very specific scientific discourse on the effectiveness of corrections, which is conducted in different disciplines. Corrections and debunking can be considered a central part of combating misinformation online. This type of intervention provides an opportunity to thoroughly confront officially refuted content with facts. Often, this takes the form of a more detailed article, which backs up its corrections with official sources. Official fact-checking websites, which are linked by the intervention, are usually used for this purpose. On the other hand, there are approaches that aim for transparency through media literacy training, for example, in the form of showing indicators and using a highlighting design of those (see examples in Figure 8). Here it is examined which components of a social media post comprehensibly indicate that it is misleading content. We discussed this type of intervention in more detail in Section 3.2.1. In addition, while labeling content as false or true without explanations typically comes as a top-down approach not addressing users' needs for transparency, labeling interventions can indeed provide comprehensibility and transparency when applied as a simplification of an otherwise too complex rating system as a combination with additional explanations.



Check for yourself: What does this video show?

XYZ News · City, State · 8 Years Ago

General Weygand called the Battle of France is over. I expect that the Battle of Britain is about to

Upon this battle depends the survival of Christian civilization. Upon it depends our own British life, and the long continuity of our institutions and our Empire.

[8] The whole fury and might[9] of the enemy must very soon be[2] turned on us.

- Hitler knows that he will have to[2] break us in this[2] Island or lose the war. If[2] we can[5] stand up to him, all Europe may be free[5] and the[2] life of the world[5] may move[2] forward into broad, sunlit uplands.[2] But if we fail, then[2] the whole[8] world,[7] including the United States, including all that we have known[8] and cared for,[8] will sink into the abyss of a new Dark Age made more sinister,[9] and perhaps

Note: The words marked in red push the prediction to be true and the ones marked in bleu push the prediction to be false.

The model predicted that the following piece of information represents a **FALSE** Claim about COVID-19:
"Hospitals receive $18 thousand for each death registered as covid-19"
Hospitals receive $18 thousand for each death registered as covid-19"
The provided explanation that strengthens the model decision is the following:

Fig. 8. Four exemplary interventions using transparent design to various degrees. A: An intervention showing the publish date of a video as indicator by Sherman et al. [170]. B: An intervention highlighting parts of a text containing propaganda techniques by Martino et al. [121]. C: An intervention used to affirm and refute claims using explainable machine learning by Ayoub et al. [13]. D: The image comparison view as part of a larger system designed by Karduni et al. [97]. The screenshots were taken from the respective papers.

3.4 Nudging as an ambivalent trending countermeasure

During the content analysis of the identified relevant literature, one specific form of user intervention particularly caught the eye: 22 publications refer to their form of intervention as a digital nudge. A nudge is defined as an intervention that *"alters people's behavior in a predictable way without significantly changing their economic incentives"* [181, p. 6]. It is a concept that has already been applied to many contexts, such as cybersecurity and health. The concept of nudging is controversially discussed in research. Thus, under certain circumstances, it represents a potential for subconscious manipulation also in harmful directions. Like many digital countermeasures, digital nudges, while promising, often offer little transparency and may run the risk of steering users in the wrong direction in ambiguous situations if they are not critically engaged with but rather trusted blindly. For instance, Lu et al. [119] show that AI-based credibility indicators can be used to steer participants in a certain direction, even if the AI is wrong. Since we did not specifically review which of the 237 interventions within 172 publications actually fit the definition of a nudge, we would nevertheless like to provide an overview of the publications that refer to their interventions as nudges themselves. Nudging has been applied by other reviews as a category of misinformation interventions on an individual level itself, complementing countermeasures like boosting, debunking, and content labeling [151]. In our review, we understand nudging as a concept that can be applied in diverse intervention designs. Some publications present intervention as "accuracy nudges" and introduce concepts in which users are specifically nudged to reflect on the accuracy of the content and to act more thoughtfully accordingly (e.g., [9, 30, 95, 127, 137]). For example, Capraro and Celadin [30] report promising results that indicate positive effects on sharing behavior when using an accuracy prompt. Similarly, von der Weth et al. [192] developed 'ShareAware' as a nudge for more conscious posting and sharing behavior. In a different approach, Andi and Akesson [7] developed a social norm-based nudge to effect sharing behavior by exposing participants to the message: "*[...] Most responsible people think twice before sharing content with their friends and followers*". In contrast, there are attempts to nudge users not only away from misinformation [48] but towards the consumption of credible news (e.g., [62, 81] or a habit of assessing accuracy of information [18, 20, 87]. In that context, Thornhill et al. [182] developed a nudge to steer users into fact-checking news online.

3.5 Impacts and perceptions of digital misinformation interventions

Our systematic review revealed a wide range of intervention designs addressing various types of user interaction and timings. For future research, it is important to determine which interventions are most promising and should be given more consideration. This paper aims to provide a comprehensive overview of misinformation interventions from diverse disciplines. These interventions strongly differ in their target of behavior change, such as improving credibility assessment, reducing the sharing of misinformation, helping users distinguish between misinformation and credible content, or decreasing the overall flow of misinformation on social media. The focus is on the design characteristics of the intervention and the user-centered evaluation method, including qualitative and quantitative evaluations. Many studies evaluate the efficacy of an intervention within a specific context (social media platforms, user groups, format of content etc.) and in comparison to a specific condition (control group without intervention, state-of-art intervention of the specific platform etc.). Others derive rich qualitative insights, e.g., into how users perceive an intervention in terms of concepts like trust, reactance, or comprehensibility.

While our review does not include a meta-analysis to derive statistical evidence of efficacy, we provide some initial insights into what appeared to be promising based on a qualitative overview, complemented by statistical effect size information that was explicitly stated in the corresponding publications. However, it is important to note that due to the nature of our review it does not allow for direct comparisons or objective evaluations of which interventions worked best or worst. Instead, it provides initial insights from a broad interdisciplinary perspective. We present our overview of central findings regarding (positive and negative) impacts and perceptions of interventions in each publication in Table 3 (electronic supplement). There, we summarize the beneficial effects of interventions that were mostly collected quantitatively, beneficial perceptions of interventions that were derived from qualitative studies, and insights on measures that were not effective or even resulted in counterproductive or unintended effects. Taking a closer look at the studies, they identify measures and characteristics that do or do not impact efficacy - sometimes with contradictory findings that demonstrate a necessity for further investigations. For instance, the timing of the correction does sometimes but not always seem to matter [41, 150], and there are indications that efficacy is sometimes but not always impacted by whether the correction is narrative or non-narrative [49, 108, 159]. There are further controversial findings on whether transparent information and explanations have a significant impact on efficacy (e.g., rather yes: [63, 104]; rather not: [120]). However, when considering findings on the role of transparency and explanations over all publications, the general tendency (including qualitative insights) indicates its impact on efficacy, user perception, and acceptance as promising.

Further controversial findings discuss which intervention mechanism/source type (social versus algorithmic correction or warning by citizens versus news agency or (e.g., health) experts) matters in terms of efficacy [e.g., 23, 71, 82, 117, 119], for example, by emphasizing potential unintended over-reliance on AI predictions even if they are not correct [119]. Other studies evaluate the impact of modality (e.g., images, videos, voice messages) of interventions on efficacy [134, 175, 183, 208, 216]. For instance, Pasquetto et al. [134] found that audio files were more effective in correcting beliefs than text or videos and Karduni et al. [97] revealed that corrections using images are more effective in correcting myths than corrections without images, independent of the image type (machine-technical image, expert image, diagram). When looking into the effect sizes (e.g., Cohen's d, (partial) η^2, Spearman's r and ρ) explicitly stated in the publications, they are small (e.g., (partial) η^2<0.06 or d<0.5) for the majority of publications [e.g., 12, 194], and medium to large ((partial) η^2>0.06) in fewer cases [e.g., 196]. As the interpretation is highly dependent on the research design, a future meta-analysis may complement our findings with statistical comparisons of efficacy that aim at controlling influencing factors revolving around the context of data in more narrowly defined domains. Comparability is often not possible due to the very diverse settings in which studies take place, addressing different social media platforms, formats of content, and participants (e.g., students with potentially higher levels of media literacy, elderly, adolescents, representative studies in different countries), and applying a variety of research designs to measure efficacy, e.g., asking participants to state whether they would share specific content versus asking them to rate the credibility [66]. Indeed, finding a consensus in research to measure the efficacy of misinformation interventions has been emphasized as an important step towards more successful interventions, and possible frameworks have been proposed [66]. Some meta-analyses have already reported on the efficacy of specific misinformation interventions like corrections, where deriving a subgroup of studies with a similar research design and conditions is sometimes achievable, allowing for comparisons or comparable interventions in different contexts. For instance, Chan and Albarracin [33] conducted a meta-analysis on the efficacy of corrections/debunking in the context of scientific misinformation, examining over 200 effect sizes and revealing that corrections are more successful when detailed. Still, in general, the debunking effect was not significant. Given the overall estimated lower impact of corrections/debunking and the strong research focus of the majority of studies on this type of intervention revealed in a related meta-analysis [21], which was confirmed in our review, this suggests scholars should not disregard other intervention types that might be less studied but more promising.

Due to the publication bias, most studies report statistically significant or qualitatively promising results. Only a few exceptions exclusively report what did not work in general [9, 31, 103, 167] or for specific user groups [109, 168]. For instance, Aslett et al. [9] report that dynamic source reliability labels placed in-feed did not reduce misperceptions, and Caramancion [31] demonstrates how preventive infographics had trivial to no effect. Despite non-efficacy of interventions, in some cases, studies reveal other unintended or counterproductive effects such as over-correction and other spill-over effects on accurate content [55, 72], over-reliance on interventions [119], increased belief in misinformation under certain circumstances like subjective messages or repeated exposure to content [11, 140, 178], priming of general mistrust in authentic content due to warnings [188], or lowered perception of extremeness due to stance labels on political ideology [62].

4 DISCUSSION

In Section 3 we have systematically categorized a variety of existing misinformation interventions to assist in dealing with misinformation online, providing concrete examples of identified dimensions. The analysis of our systematic literature review underscores the impression that a variety of diverse approaches have emerged in recent years and continue to emerge. It can be observed that these often differ significantly in their characteristics. In this section, we discuss and summarize our findings regarding our research questions.

4.1 RQ1: What are the typical methodological characteristics of existing studies on misinformation interventions?

Methodologically, studies of misinformation interventions differ in various dimensions, although several emphases and typical patterns are also apparent. Due to the inclusion and exclusion criteria of our review, the sample contains mainly studies that collect empirical data and only a few publications with an exclusively conceptual approach. Typically, publications on misinformation interventions evaluate novel or already established interventions in user studies, often in comparison to other existing approaches. A particular focus is on online experiments with a collection of quantitative and qualitative data, as this method is suitable for large-scale samples and a controlled environment. In order to examine the interventions in a realistic environment and to minimize biases, more evaluation in the form of field studies would be desirable for future studies. It is striking that mainly U.S. adults and college students are surveyed as study participants, while specific (vulnerable) target groups such as teenagers, persons of older age, or non-native speakers are largely neglected. This can be explained by the better accessibility of different user groups and represents a common problem known from other user studies in contexts of human-computer interaction and similar disciplines.

Not surprisingly, most publications deal with Facebook and Twitter/X as social media platforms or address news articles in general (see Section 3.1). Considering current and emergent social media platforms like TikTok and Instagram as image- and video-based platforms is still largely missing within the research landscape. However, the impact of those platforms and content types has shown to be highly relevant. Looking closer at our publication sample, we can note that there are already isolated publications for addressing exceptional formats such as image, video and audio. We can see that there is a positive correlation between the addressed formats 'video' and 'image'. Indeed, three out of five interventions for video formats are specifically addressing images as well (e.g., [12]).

4.2 RQ2: How do existing forms of user-centered misinformation interventions assist users in dealing with misinformation online?

Looking closely at misinformation interventions, one notices a publication emphasis on corrections and debunking. This form of intervention can be artificially controlled or occur naturally in the form of user comments. It is striking that corrections/debunking are examined in great detail in the literature from a wide variety of perspectives and with a focus on the smallest details, e.g., regarding timing or repetition [39, 43]. This finding is also supported by the review study by Chen et al. [36], who identified fact-based corrections as "the most common type of corrective communication strategy", classifying it as a part of message-based approaches. Often, corrections from the official side are based on thorough and elaborate journalistic work. For example, social media articles are linked to an official correction once the content has been thoroughly checked manually by experts. In the fast pace of social media and especially in emerging crisis situations, there is an overflow of accurate and misinformation that needs to be reacted to quickly. This is where expert-based corrections as digital countermeasures sometimes reach their limits as stand-alone interventions. Other approaches pursue corrections based on the assessment of users themselves. While there is no expert review here, active user participation in news consumption is facilitated and studies reveal promising findings. For instance, participants preferred suggested headlines by laypersons that corrected the original ones [88]. Indeed, the effects of corrections by laypersons versus experts have been studied in prior work [73, 197] and constitute a relevant dimension of interventions beyond this work's scope. While many correction interventions provide users with additional (fact-based) knowledge and can thus create transparency, other types of interventions aim to increase transparency and thus media competence, for example, through linguistic or content-related indicators. An advantage of those interventions is the scalability of using automatic detection algorithms in real-time during emerging crisis situations and on large data sets, often based on machine learning approaches. However, when automatically showing indicators such as a missing verification seal or semantic propaganda techniques, the final decision on whether content is misinformation or not either lies with the user or is taken over by the algorithm based on (potentially biased) training data, missing the expert knowledge of professional fact-checkers. Transparent misinformation interventions, independent of their implementation as correction or display of automatically detected indicators or other types, may offer the opportunity to counteract reactance of end users in contrast to approaches that lack an explanation and, in some cases, facilitate a feeling of censorship, paternalism, and loss of control.

In contrast to transparent approaches, there are also fewer educational interventions with the goal of reducing the consumption of misinformation through removal or visibility reduction. Both intervention designs can be considered helpful when considering the bias of people remembering content itself without a potentially shown correction or warning when exposed to misleading content [64]. On the other hand, this intervention design may lead to (a feeling of) censorship and a resulting migration to other platforms or tools that take less rigorous action against problematic content. While deleting/censoring dangerous or explicit content is a legitimate and important responsibility of social media platform operators, applying this solution of deletion to all problematic content, such as disinformation and misinformation, would not only lead to a migration of users to less restrictive platforms. In particular, it would represent a missed opportunity for media literacy education, some of which can be achieved through transparent digital countermeasures as a complement to school lessons.

In order to develop misinformation interventions in a user-centered way and to be able to achieve an actual effect, the early inclusion of the needs and requirements of different target groups is indispensable. A particular challenge is the accessibility of people who have no trust in official bodies. In this context, limits certainly emerge as to who can be reached at all by the corresponding technical tools. In order to avoid reactance, the timing of the intervention certainly plays a role in addition to the transparency of approaches. In our systematic literature review, we identified interventions that can be used at the user's request and others that are permanently present during the normal use of social media. It is an interesting research question: which point in time or which regularity of an intervention is suitable for which target groups? Considering the broad variety of misinformation interventions, we hope to provide a helpful overview of existing forms. We propose our taxonomy as a starting point to systematically capture intervention categories and identify relevant dimensions. It is intended to provide researchers with a framework to develop new interventions, to pool knowledge from different disciplines for the promotion of cross-disciplinary research, and to reveal promising research directions.

4.3 RQ3: Which trends and chances for future research can be derived from the existing literature?

In this paper, we have systematically analyzed 172 publications with 237 user interventions to assist in dealing with misinformation online. Our findings reveal current trends and movements in human-computer interaction, psychology, information systems, and communication sciences. As potential avenues for future research, we propose the following questions and interests:

(1) Are approaches for specific platforms transferable to other new platforms? With regard to particularly relevant contexts of use, it is also necessary to consider current and emergent social media platforms. Social media platforms are constantly changing. For some time now, there has been a noticeable trend toward TikTok, and Facebook, in particular, is losing a great deal of its importance, especially among younger people. In order not to have to reinvent the wheel again and again, studies on the transferability of findings to new types of platforms are important. While the majority of studies surveyed much-researched platforms such as Twitter (now X) and Facebook (e.g., [20, 102, 122, 129]; see Figure 3 and Section 3.1), there has been little research on misinformation on image- and video-based platforms such as Instagram and TikTok. Given the usage rates of these media, particularly among youth, and the increasing relevance of the platforms for misinformation (e.g., concerning the Russian-Ukrainian war), addressing this research gap is considered particularly relevant. At the same time, there are major obstacles to overcome here, especially with regard to the availability of labeled datasets, as they typically already exist for Twitter/X but are very time-consuming and complex to establish for video data.

(2) How can collected findings and technical approaches for text-based interventions be applied to emerging video- and image-based misinformation? With regard to transferability to new platforms, transferability to other information channels is also particularly central. Can text-based user interventions (e.g., [60]) be adapted for video- and image-based channels? How can new indicators and measures for detecting misinformation (e.g., image reverse search) be integrated into misinformation interventions? Challenges researchers are confronted with include the fast-evolving trends in social media. For instance, emotion-evoking content features on TikTok might solely constitute a characteristic of the platform's content while it might be considered a more valuable indicator for misinformation in other modalities.

(3) How can chances of digital misinformation interventions be effectively combined with the advantages of human experts? Fully automated mechanisms, e.g., for machine learning-based detection, can handle large amounts of data better than humans. In contrast, trained humans as experts (e.g., journalists [176]) can handle specific case decisions better than algorithms when the boundary between true and false is blurred and information is missing. How can human expert knowledge be used within digital countermeasures without losing the performance of the automatic tool? In which steps of the countermeasure can human intervention be integrated?

(4) How can automatic detection be combined with user-centered feedback? Automatic detection is often a black-box procedure and, therefore, cannot explain its decision-making. While efforts in explainable artificial intelligence already reveal promising results to make detection approaches more transparent without the human directly in the loop, they are still challenged with how to present these explainable outputs in a way that is valuable for a layperson, especially for people with low media literacy. How can the advantages of accurate automatic detection be combined with transparent and comprehensible explanations as user-centered feedback (see Section 3.3; initial attempts e.g., by Schmid et al. [163])?

(5) How can media literacy be encouraged without exposing users to misinformation? Approaches to increasing media literacy are often based on a display of misinformation with additional reference to comprehensible indicators or debunking. Nevertheless, the user is still exposed to the misinformation in this case. Studies suggest that even with a simultaneous warning, the misinformation content may be remembered at a later point in time, which speaks for less exposure to misinformation [64] and constitutes an effect that is controversially discussed in literature [158]. However, users tend to feel reactance and paternalism when content is hidden or deleted [104]. How can media literacy be trained within misinformation interventions without continuing to expose users to misinformation?

(6) Can vulnerable people profit from the general findings of participants with high media literacy? How can we reach vulnerable people with official tools? The bias toward U.S. adults and college students as study participants continues to be striking. Since not all individuals are equally affected by misinformation [171], but rather particularly vulnerable groups exist, the inclusion of individuals with lower levels of media literacy in the iterative design and evaluation process of the interventions is essential. Initial approaches are already moving in this direction [12]. Still, researchers are often challenged with conducting user studies outside of the university bubble with convenience samples of participants, as particularly studies with children and teenagers in the context of misinformation come with additional ethical questions and recruiting challenges. For instance, confronting adolescents with misleading information during a user study is a sensible task that needs thorough consideration. How can the findings be applied to vulnerable people (see Section 3.1; e.g., older people [155], children, teenagers [12, 173], non-native speakers)? How can these target groups be meaningfully integrated into the iterative design and evaluation process?

With our systematic literature review we hope to provide a starting point for cross-disciplinary debates and knowledge exchange, as well as an inspiration for future research.

5 LIMITATIONS & CONCLUSION

First, the large amount of publications in the area of interest, including a variety of different disciplines, was a challenge to deal with. Therefore, we focused specifically on approaches that take place within the real-time usage of social media and excluded approaches (especially educational trainings, games, and presentations) that take place long before the actual usage of social media [e.g., 173]. However, those approaches may provide additional insights into effective and user-centered interventions and are, therefore, suggested for future research. In addition, we excluded studies on psychological or social phenomena (e.g., norms) to receive a reasonable number of publications that allows for a thorough focus on research regarding the design and evaluation of digital interventions. These studies are valuable to consider when designing interventions tailored to specific persona and are suggested for future reviews.

Second, our approach takes a broad perspective on types of misleading information referred to as ‘misinformation’ as an umbrella term and encompassing unintentionally and intentionally misleading information as well as related phenomena (e.g., rumors, conspiracy theories). Table 2 roughly demonstrates in clusters which concept was used in each paper. While we excluded papers that contained a related term but understood it as a phenomenon not fitting within our broad definition (e.g., eyewitnesses remembering something inaccurately as 'misinformation'), we did not perform an in-depth analysis of how each term was defined and utilized in each paper. This is a limitation that may have impacted our screening phase. In addition, there might be odd cases of papers not identified within our systematic screening phase as they use different terms to address the topic of misinformation not included in our search term. For instance, a study by Ennals et al. [53] was brought to our attention during the review phase that refers to 'disputed claims' and was, thus, not detected.

Third, within our work, we thoroughly categorized the publications regarding multiple characteristics, involving two researchers with expert knowledge in that field of research. As the publications provide information on our categories in varying detail, we cannot exclude the possibility that some interventions were classified differently than they were intended by the authors themselves.

Fourth, while our work examines misinformation interventions from multiple perspectives, there are additional significant dimensions that have not yet been covered in this systematic study and are suggested for future work. In particular, in the context of user-centered interventions, looking at who is the arbiter of content credibility (e.g., decentralized decisions by the crowd versus experts or algorithmic decisions) is an important dimension that has significant effects on intervention perception and impact.

Misinformation remains a threat to the democratic order and the cohesion of society, and the fight against it remains important. It is a central goal to empower users in dealing with the overabundance of information online, especially during emerging crises. Digital misinformation interventions are one of several starting points to address that challenge, complementing professional journalistic work and media literacy training at schools. In this work, we have given an overview of existing countermeasures and have developed a taxonomy in order to systematize misinformation intervention research. Finally, we hope that this work – being a first step towards the systematization of misinformation intervention research – serves as an inspiration for future research and facilitates cross-disciplinary exchange of knowledge.

REFERENCES

[1] Zhila Aghajari, Eric Baumer, and Dominic Difranzo. 2023. Reviewing Interventions to Address Misinformation: The Need to Expand Our Vision Beyond an Individualistic Focus. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (Jan. 2023), 1–34. https://doi.org/10.1145/3579520

[2] Jon Agley, Yunyu Xiao, Esi Thompson, Xiwei Chen, and Lilian Golzarri-Arroyo. 2021. Intervening on Trust in Science to Reduce Belief in COVID-19 Misinformation and Increase COVID-19 Preventive Behavioral Intentions: Randomized Controlled Trial. *Journal of Medical Internet Research* 23, 10 (Oct. 2021), e32425. https://doi.org/10.2196/32425

[3] Michael Aird, Ulrich Ecker, Briony Swire, Adam Berinsky, and Stephan Lewandowsky. 2018. Does Truth Matter to Voters? The Effects of Correcting Political Misinformation in an Australian Sample. *Royal Society Open Science* 5, 12 (Dec. 2018), 180593. https://doi.org/10.1098/rsos.180593

[4] Malik Almaliki. 2019. Misinformation-Aware Social Media: A Software Engineering Perspective. *IEEE Access* 7 (2019), 182451–182458. https://doi.org/10.1109/ACCESS.2019.2960270

[5] Malik Almaliki. 2019. Online Misinformation Spread: A Systematic Literature Map. In *Proceedings of the 2019 3rd International Conference on Information System and Data Mining (ICISDM 2019)*. Association for Computing Machinery, New York, NY, USA, 171–178. https://doi.org/10.1145/3325917.3325938

[6] Zaid Amin, Nazlena Mohamad Ali, and Alan F. Smeaton. 2021. Visual Selective Attention System to Intervene User Attention in Sharing COVID-19 Misinformation. *International Journal of Advanced Computer Science and Applications* 12, 10 (2021), 36–41. https://doi.org/10.14569/IJACSA.2021.0121005

[7] Simge Andi and Jesper Akesson. 2020. Nudging Away False News: Evidence from a Social Norms Experiment. *Digital Journalism* 9, 1 (Nov. 2020), 106–125. https://doi.org/10.1080/21670811.2020.1847674

[8] Alberto Ardevol-Abreu, 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 (Sept. 2020), 1–13. https://doi.org/10.3145/epi.2020.sep.07

[9] Kevin Aslett, Andrew M. Guess, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2022. News Credibility Labels Have Limited Average Effects on News Diet Quality and Fail to Reduce Misperceptions. *Science Advances* 8, 18 (May 2022), eabl3844. https://doi.org/10.1126/sciadv.abl3844

[10] Shubham Atreja, Libby Hemphill, and Paul Resnick. 2023. Remove, Reduce, Inform: What Actions Do People Want Social Media Platforms to Take on Potentially Misleading Content? *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 291 (Oct. 2023), 33 pages. https://doi.org/10.1145/3610082

[11] Kevin Autry and Shea Duarte. 2021. Correcting the Unknown: Negated Corrections May Increase Belief in Misinformation. *Applied Cognitive Psychology* 35, 4 (July 2021), 960–975. https://doi.org/10.1002/acp.3823

[12] Carl-Anton Werner Axelsson, Mona Guath, and Thomas Nygren. 2021. Learning How to Separate Fake from Real News: Scalable Digital Tutorials Promoting Students' Civic Online Reasoning. *Future Internet* 13, 3 (March 2021), 60. https://doi.org/10.3390/fi13030060

[13] Jackie Ayoub, X. Jessie Yang, and Feng Zhou. 2021. Combat COVID-19 Infodemic Using Explainable Natural Language Processing Models. *Information Processing & Management* 58, 4 (July 2021), 102569. https://doi.org/10.1016/j.ipm.2021.102569

[14] Ingrid Bachmann and Sebastian Valenzuela. 2023. Studying the Downstream Effects of Fact-Checking on Social Media: Experiments on Correction Formats, Belief Accuracy, and Media Trust. *Social Media + Society* 9, 20563051231179694 (June 2023), 1–13. https://doi.org/10.1177/20563051231179694

[15] Joseph B. Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S. Schafer, Emma S. Spiro, Kate Starbird, and Jevin D. West. 2022. Combining Interventions to Reduce the Spread of Viral Misinformation. *Nature Human Behaviour* 6, 10 (Oct. 2022), 1372–+. https://doi.org/10.1038/s41562-022-01388-6

[16] Dipto Barman and Owen Colan. 2023. Does Explanation Matter? An Exploratory Study on the Effects of Covid–19 Misinformation Warning Flags on Social Media. In *2023 10th International Conference on Behavioural and Social Computing (BESC)*. IEEE, Larnaca, Cyprus, 1–7. https://doi.org/10.1109/BESC59560.2023.10386371

[17] Ranojoy Barua, Rajdeep Maity, Dipankar Minj, Tarang Barua, and Ashish Kumar Layek. 2019. F-NAD: An Application for Fake News Article Detection Using Machine Learning Techniques. In *2019 IEEE Bombay Section Signature Conference (IBSSC)*. IEEE, Mumbai, India, 1–6. https://doi.org/10.1109/IBSSC47189.2019.8973059

[18] Md Momen Bhuiyan, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021. NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 427:1–427:30. https://doi.org/10.1145/3479571

[19] MD Momen Bhuiyan, Hayden Whitley, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021. Designing Transparency Cues in Online News Platforms to Promote Trust: Journalists' &amp; Consumers' Perspectives. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 395:1–395:31. https://doi.org/10.1145/3479539

[20] Md Momen Bhuiyan, Kexin Zhang, Kelsey Vick, Michael A. Horning, and Tanushree Mitra. 2018. FeedReflect: A Tool for Nudging Users to Assess News Credibility on Twitter. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '18)*. Association for Computing Machinery, New York, NY, USA, 205–208. https://doi.org/10.1145/3272973.3274056

[21] Robert A. Blair, Jessica Gottlieb, Brendan Nyhan, Laura Paler, Pablo Argote, and Charlene J. Stainfield. 2024 FEB 2024. Interventions to Counter Misinformation: Lessons from the Global North and Applications to the Global South. *Current Opinion in Psychology* 55, 101732 (2024 FEB 2024), 1–10. https://doi.org/10.1016/j.copsyc.2023.101732

[22] Leticia Bode and Emily K. Vraga. 2015. In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media. *Journal of Communication* 65, 4 (Aug. 2015), 619–638. https://doi.org/10.1111/jcom.12166

[23] L Bode and EK Vraga. 2018. See Something, Say Something: Correction of Global Health Misinformation on Social Media. *Health Communication* 33, 9 (2018), 1131–1140. https://doi.org/10.1080/10410236.2017.1331312

[24] Lia Bozarth, Jane Im, Christopher Quarles, and Ceren Budak. 2023. Wisdom of Two Crowds: Misinformation Moderation on Reddit and How to Improve This Process—A Case Study of COVID-19. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 155 (April 2023), 33 pages. https://doi.org/10.1145/3579631

[25] Nadia M. Brashier, Gordon Pennycook, Adam J. Berinsky, and David G. Rand. 2021. Timing Matters When Correcting Fake News. *Proceedings of the National Academy of Science of the United States of America* 118, 5 (Feb. 2021), e2020043118. https://doi.org/10.1073/pnas.2020043118

[26] Adrian M. P. Braşoveanu and Răzvan Andonie. 2019. Semantic Fake News Detection: A Machine Learning Perspective. In *Advances in Computational Intelligence*, Ignacio Rojas, Gonzalo Joya, and Andreu Catala (Eds.). Vol. 11506. Springer International Publishing, Cham, 656–667. https://doi.org/10.1007/978-3-030-20521-8_54

[27] Sonja Buchegger and Jean-Yves Le Boudec. 2003. The Effect of Rumor Spreading in Reputation Systems for Mobile Ad-hoc Networks. In *WiOpt'03: Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*. IEEE, Sophia Antipolis, France, 1–10.

[28] Klara Austeja Buczel, Adam Siwiak, Malwina Szpitalak, and Romuald Polczyk. 2024. How Do Forewarnings and Post-Warnings Affect Misinformation Reliance? The Impact of Warnings on the Continued Influence Effect and Belief Regression. *Memory & Cognition* 52, 2 (2024), 1–17. https://doi.org/10.3758/s13421-024-01520-z

[29] Danielle Caled and Mário J. Silva. 2021. Digital Media and Misinformation: An Outlook on Multidisciplinary Strategies against Manipulation. *Journal of Computational Social Science* 5, 1 (May 2021), 123–159. https://doi.org/10.1007/s42001-021-00118-8

[30] Valerio Capraro and Tatiana Celadin. 2022. "I Think This News Is Accurate": Endorsing Accuracy Decreases the Sharing of Fake News and Increases the Sharing of Real News. *Personality and Social Psychology Bulletin* 0, 0 (2022), 1–11. https://doi.org/10.1177/01461672221117691

[31] Kevin Matthe Caramancion. 2022. Quantification of Infographic Intervention Effect on Mis/Disinformation Vulnerability. In *2022 IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT)*. IEEE, Lviv, Ukraine, 297–300. https://doi.org/10.1109/CSIT56902.2022.10000489

[32] Aimee Challenger, Petroc Sumner, and Lewis Bott. 2022. COVID-19 Myth-Busting: An Experimental Study. *Bmc Public Health* 22, 1 (Jan. 2022), 131. https://doi.org/10.1186/s12889-021-12464-3

[33] Man-pui Sally Chan and Dolores Albarracin. 2023 SEP 2023. A Meta-Analysis of Correction Effects in Science-Relevant Misinformation. *Nature Human Behaviour* 7, 9 (2023 SEP 2023), 1514–1525. https://doi.org/10.1038/s41562-023-01623-8

[34] Guangyu Chen, Paolo Ciuccarelli, and Sara Colombo. 2022. VisualBubble: Exploring How Reflection-Oriented User Experiences Affect Users' Awareness of Their Exposure to Misinformation on Social Media. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 289, 7 pages. https://doi.org/10.1145/3491101.3519615

[35] Liang Chen and Hongjie Tang. 2022. Examining the Persuasion Process of Narrative Fear Appeals on Health Misinformation Correction. *Information, Communication & Society* 26, 15 (Oct. 2022), 1–19. https://doi.org/10.1080/1369118X.2022.2128849

[36] Sijing Chen, Lu Xiao, and Akit Kumar. 2022. Spread of Misinformation on Social Media: What Contributes to It and How to Combat It. *Computers in Human Behavior* 141, 2023 (Dec. 2022), 107643. https://doi.org/10.1016/j.chb.2022.107643

[37] Tosti H. C. Chiang, Chih-Shan Liao, and Wei-Ching Wang. 2022. Impact of Artificial Intelligence News Source Credibility Identification System on Effectiveness of Media Literacy Education. *Sustainability* 14, 8 (April 2022), 4830. https://doi.org/10.3390/su14084830

[38] Katherine Clayton, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, Morgan Sandhu, Rachel Sang, Rachel Scholz-Bright, Austin T. Welch, Andrew G. Wolff, Amanda Zhou, and Brendan Nyhan. 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 (Dec. 2020), 1073–1095. https://doi.org/10.1007/s11109-019-09533-0

[39] Michael Craig and Santosh Vijaykumar. 2023. One Dose Is Not Enough: The Beneficial Effect of Corrective COVID-19 Information Is Diminished If Followed by Misinformation. *Social Media + Society* 9, 20563051231161298 (April 2023), 1–14. https://doi.org/10.1177/20563051231161298

[40] Shih-Chieh Dai, Yi-Li Hsu, Aiping Xiong, and Lun-Wei Ku. 2022. Ask to Know More: Generating Counterfactual Explanations for Fake Claims. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*. Association for Computing Machinery, New York, NY, USA, 2800–2810. https://doi.org/10.1145/3534678.3539205

[41] Yue Dai. 2021. The Effects of Message Order and Debiasing Information in Misinformation Correction. *International Journal of Communication* 15, 2021 (2021), 1039–1059.

[42] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2020. Wearable Reasoner: Towards Enhanced Human Rationality Through A Wearable Device With An Explainable AI Assistant. In *Proceedings of the Augmented Humans International Conference (AHs '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3384657.3384799

[43] Nora Denner, Benno Viererbl, and Thomas Koch. 2023. Effects of Repeated Corrections of Misinformation on Organizational Trust: More Is Not Always Better. *International Journal of Strategy Communication* 17, 1 (Jan. 2023), 39–53. https://doi.org/10.1080/1553118X.2022.2135098

[44] Saoirse Connor Desai and Stian Reimers. 2023. Does Explaining the Origins of Misinformation Improve the Effectiveness of a given Correction? *Memory & Cognition* 51, 2 (2023), 422–436. https://doi.org/10.3758/s13421-022-01354-7

[45] Tom Dobber, Sanne Kruikemeier, Fabio Votta, Natali Helberger, and Ellen P. Goodman. 2023. The Effect of Traffic Light Veracity Labels on Perceptions of Political Advertising Source and Message Credibility on Social Media. *Journal of Information Technology and Politics* (2023), 1–16. https://doi.org/10.1080/19331681.2023.2224316

[46] Shawn Domgaard and Mina Park. 2021. Combating Misinformation: The Effects of Infographics in Verifying False Vaccine News. *Health Education Journal* 80, 8 (Dec. 2021), 974–986. https://doi.org/10.1177/00178969211038750

[47] Chiara Patricia Drolsbach and Nicolas Pröllochs. 2023. Diffusion of Community Fact-Checked Misinformation on Twitter. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 267 (Oct. 2023), 22 pages. https://doi.org/10.1145/3610058

[48] Megan Duncan. 2020. What's in a Label? Negative Credibility Labels in Partisan News. *Journalism & Mass Communication Quarterly* 99, 2 (Oct. 2020), 390–413. https://doi.org/10.1177/1077699020961856

[49] Ulrich K. H. Ecker, 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 (Dec. 2020), 1–26. https://doi.org/10.1186/s41235-020-00266-x

[50] Ullrich K. H. Ecker, Joshua L. Hogan, and Stephan Lewandowsky. 2017. Reminders and Repetition of Misinformation: Helping or Hindering Its Retraction?: Journal of Applied Research in Memory and Cognition. *Journal of Applied Research in Memory and Cognition* 6, 2 (June 2017), 185–192. https://doi.org/10.1037/h0101809

[51] UKH Ecker, S Lewandowsky, and M Chadwick. 2020. Can Corrections Spread Misinformation to New Audiences? Testing for the Elusive Familiarity Backfire Effect. *Cognitive Research: Principles and Implications* 5, 1 (Aug. 2020), 1–25. https://doi.org/10.1186/s41235-020-00241-6

[52] UKH Ecker, S Lewandowsky, B Swire, and D Chang. 2011. Correcting False Information in Memory: Manipulating the Strength of Misinformation Encoding and Its Retraction. *Psychonomic Bulletin & Review* 18, 3 (June 2011), 570–578. https://doi.org/10.3758/s13423-011-0065-1

[53] Rob Ennals, Beth Trushkowsky, and John Mark Agosta. 2010. Highlighting Disputed Claims on the Web. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, Raleigh North Carolina USA, 341–350. https://doi.org/10.1145/1772690.1772726

[54] Ziv Epstein, Gordon Pennycook, and David Gertler Rand. 2020. Will the Crowd Game the Algorithm? Using Layperson Judgments to Combat Misinformation on Social Media by Downranking Distrusted Sources. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. https://doi.org/10.31234/osf.io/z3s5k

[55] K. J. Kevin Feng, Nick Ritchie, Pia Blumenthal, Andy Parsons, and Amy X. Zhang. 2023. Examining the Impact of Provenance-Enabled Media on Trust and Accuracy Perceptions. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 270 (Oct. 2023), 42 pages. https://doi.org/10.1145/3610061

[56] Kathrin Figl, Samuel Kiessling, and Ulrich Remus. 2023. Do Symbol and Device Matter? The Effects of Symbol Choice of Fake News Flags and Device on Human Interaction with Fake News on Social Media Platforms. *Computers in Human Behavior* 144, 107704 (2023), 1–16. https://doi.org/10.1016/j.chb.2023.107704

[57] Frans Folkvord, Freek Snelting, Doeschka Anschutz, Tilo Hartmann, Alexandra Theben, Laura Gunderson, Ivar Vermeulen, and Francisco Lupiáñez-Villanueva. 2022. Effect of Source Type and Protective Message on the Critical Evaluation of News Messages on Facebook: Randomized Controlled Trial in the Netherlands. *Journal of Medical Internet Research* 24, 3 (March 2022), e27945. https://doi.org/10.2196/27945

[58] Anjuli Franz, Verena Zimmermann, Gregor Albrecht, Katrin Hartwig, Christian Reuter, Alexander Benlian, and Joachim Vogt. 2021. SoK: Still Plenty of Phish in the Sea — A Taxonomy of User-Oriented Phishing Interventions and Avenues for Future Research. In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. USENIX, Virtual Event, 339–358.

[59] Melanie Freeze, Mary Baumgartner, Peter Bruno, Jacob R. Gunderson, Joshua Olin, Morgan Quinn Ross, and Justine Szafran. 2021. Fake Claims of Fake News: Political Misinformation, Warnings, and the Tainted Truth Effect. *Political Behavior* 43, 4 (Dec. 2021), 1433–1465. https://doi.org/10.1007/s11109-020-09597-3

[60] Tomoya Furuta and Yu Suzuki. 2021. A Fact-checking Assistant System for Textual Documents. In *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, Tokyo, Japan, 243–246. https://doi.org/10.1109/MIPR51284.2021.00046

[61] Vijaya Gadde and Matt Derella. 2020. An Update on Our Continuity Strategy during COVID-19.

[62] Mingkun Gao, Ziang Xiao, Karrie Karahalios, and Wai-Tat Fu. 2018. To Label or Not to Label: The Effect of Stance and Credibility Labels on Readers' Selection and Perception of News Articles. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 55:1–55:16. https://doi.org/10.1145/3274324

[63] Anat Gesser-Edelsburg, Alon Diamant, Rana Hijazi, and Gustavo S. Mesch. 2018. Correcting Misinformation by Health Organizations during Measles Outbreaks: A Controlled Experiment. *PLOS ONE* 13, 12 (Dec. 2018), e0209505. https://doi.org/10.1371/journal.pone.0209505

[64] Rebecca Hofstein Grady, Peter Ditto, and Elizabeth Loftus. 2021. Nevertheless, Partisanship Persisted: Fake News Warnings Help Briefly, but Bias Returns with Time. *Cognitive Research-Principles and Implications* 6, 1 (July 2021), 1–16. https://doi.org/10.1186/s41235-021-00315-z

[65] Sukeshini Grandhi, Linda Plotnick, and Starr Roxanne Hiltz. 2021. By the Crowd and for the Crowd: Perceived Utility and Willingness to Contribute to Trustworthiness Indicators on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 5, GROUP (July 2021), 1–24. https://doi.org/10.1145/3463930

[66] Brian Guay, Adam J. Berinsky, Gordon Pennycook, and David Rand. 2023 AUG 2023. How to Think about Whether Misinformation Interventions Work. *Nature Human Behaviour* 7, 8 (2023 AUG 2023), 1231–1233. https://doi.org/10.1038/s41562-023-01667-w

[67] Andrew M. Guess, Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A Digital Media Literacy Intervention Increases Discernment between Mainstream and False News in the United States and India. *Proceedings of the Nationale Academy of Science of the United States of America* 117, 27 (July 2020), 15536–15545. https://doi.org/10.1073/pnas.1920498117

[68] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2021. The Future of False Information Detection on Social Media: New Perspectives and Trends. *Comput. Surveys* 53, 4 (July 2021), 1–36. https://doi.org/10.1145/3393880

[69] Chen Guo, Nan Zheng, and Chengqi (John) Guo. 2023. Seeing Is Not Believing: A Nuanced View of Misinformation Warning Efficacy on Video-Sharing Social Media Platforms. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 294 (Oct. 2023), 35 pages. https://doi.org/10.1145/3610085

[70] Michael Hameleers. 2022. Separating Truth from Lies: Comparing the Effects of News Media Literacy Interventions and Fact-Checkers in Response to Political Misinformation in the US and Netherlands. *Information, Communication & Society* 25, 1 (Jan. 2022), 110–126. https://doi.org/10.1080/1369118X.2020.1764603

[71] Michael Hameleers, 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 (March 2020), 281–301. https://doi.org/10.1080/10584609.2019.1674979

[72] Michael Hameleers and Toni van der Meer. 2023. Striking the Balance between Fake and Real: Under What Conditions Can Media Literacy Messages That Warn about Misinformation Maintain Trust in Accurate Information? *Behaviour & Information Technology* (2023), 1–14. https://doi.org/10.1080/0144929X.2023.2267700

[73] Aniko Hannak, Drew Margolin, Brian Keegan, and Ingmar Weber. 2014. Get Back! You Don't Know Me like That: The Social Mediation of Fact-Checking Interventions in Twitter Conversations. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8. AAAI, Michigan, USA, 187–196.

[74] Katrin Hartwig, Tom Biselli, Franziska Schneider, and Christian Reuter. 2024. From Adolescents' Eyes: Assessing an Indicator-Based Intervention to Combat Misinformation on TikTok. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–20. https://doi.org/10.1145/3613904.3642264

[75] Katrin Hartwig and Christian Reuter. 2022. Nudging Users towards Better Security Decisions in Password Creation Using Whitebox-Based Multidimensional Visualisations. *Behaviour & Information Technology* 41, 7 (May 2022), 1357–1380. https://doi.org/10.1080/0144929X.2021.1876167

[76] Katrin Hartwig, Ruslan Sandler, and Christian Reuter. 2024. Navigating Misinformation in Voice Messages: Identification of User-Centered Features for Digital Interventions. *Risk, Hazards & Crisis in Public Policy* 0, 0 (2024), 1–33. https://doi.org/10.1002/rhc3.12296

[77] Katrin Hartwig, Stefka Schmid, Tom Biselli, Helene Pleil, and Christian Reuter. 2024. Misleading information in crises: exploring content-specific indicators on Twitter from a user perspective. *Behaviour & Information Technology* (2024).

[78] Sarah Hawa, Lanita Lobo, Unnati Dogra, and Vijaya Kamble. 2021. Combating Misinformation Dissemination through Verification and Content Driven Recommendation. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. Curran Associates, Tirunelveli, India, 917–924. https://doi.org/10.1109/ICICV50876.2021.9388406

[79] Stefan Helmstetter and Heiko Paulheim. 2018. Weakly Supervised Learning for Fake News Detection on Twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, Barcelona, 274–277. https://doi.org/10.1109/ASONAM.2018.8508520

[80] Hendrik Heuer and Elena Leah Glassman. 2022. A Comparative Evaluation of Interventions against Misinformation: Augmenting the WHO Checklist. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 241, 21 pages. https://doi.org/10.1145/3491102.3517717

[81] Benjamin D. Horne, Mauricio Gruppi, and Sibel Adali. 2019. Trustworthy Misinformation Mitigation with Soft Information Nudging. In *2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE, Los Angeles, CA, USA, 245–254. https://doi.org/10.1109/TPS-ISA48467.2019.00039

[82] Yan Huang and Weirui Wang. 2022. When a Story Contradicts: Correcting Health Misinformation on Social Media through Different Message Formats and Mechanisms. *Information Communication & Society* 25, 8 (2022), 1192–1209.

[83] Dulcie Irving, Robbie W. A. Clark, Stephan Lewandowsky, and Peter J. Allen. 2022. Correcting Statistical Misinformation about Scientific Findings in the Media: Causation versus Correlation. *Journal of Experimental Psychology-Applied* 28, 1 (2022), 1–9. https://doi.org/10.1037/xap0000408

[84] Matthew O. Jackson, Suraj Malladi, and David McAdams. 2022. Learning through the Grapevine and the Impact of the Breadth and Depth of Social Networks. *Proceedings of the National Academy of Sciences* 119, 34 (Aug. 2022), e2205549119. https://doi.org/10.1073/pnas.2205549119

[85] Farnaz Jahanbakhsh and David R Karger. 2024. A Browser Extension for In-Place Signaling and Assessment of Misinformation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–21. https://doi.org/10.1145/3613904.3642473

[86] Farnaz Jahanbakhsh, Yannis Katsis, Dakuo Wang, Lucian Popa, and Michael Muller. 2023. Exploring the Use of Personalized AI for Identifying Misinformation on Social Media. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 105, 27 pages. https://doi.org/10.1145/3544548.3581219

[87] Farnaz Jahanbakhsh, Amy X. Zhang, Adam J. Berinsky, Gordon Pennycook, David G. Rand, and David R. Karger. 2021. Exploring Lightweight Interventions at Posting Time to Reduce the Sharing of Misinformation on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 18:1–18:42. https://doi.org/10.1145/3449092

[88] Farnaz Jahanbakhsh, Amy X. Zhang, Karrie Karahalios, and David R. Karger. 2022. Our Browser Extension Lets Readers Change the Headlines on News Articles, and You Won't Believe What They Did! *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 530 (Nov. 2022), 33 pages. https://doi.org/10.1145/3555643

[89] Farnaz Jahanbakhsh, Amy X. Zhang, and David R. Karger. 2022. Leveraging Structured Trusted-Peer Assessments to Combat Misinformation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 524:1–524:40. https://doi.org/10.1145/3555637

[90] M. Rosie Jahng, Elizabeth Stoycheff, and Annisa Rochadiat. 2021. They Said It's "Fake": Effects of Discounting Cues in Online Comments on Information Quality Judgments and Information Authentication. *Mass Communication and Society* 24, 4 (July 2021), 527–552. https://doi.org/10.1080/15205436.2020.1870143

[91] Kamila Janmohamed, Nathan Walter, Kate Nyhan, Kaveh Khoshnood, Joseph D. Tucker, Natalie Sangngam, Frederick L. Altice, Qinglan Ding, Allie Wong, Zachary M. Schwitzky, Chris T. Bauch, Munmun De Choudhury, Orestis Papakyriakopoulos, and Navin Kumar. 2021 DEC 2 2021. Interventions to Mitigate COVID-19 Misinformation: A Systematic Review and Meta-Analysis. *Journal of Health Communication* 26, 12 (2021 DEC 2 2021), 846–857. https://doi.org/10.1080/10810730.2021.2021460

[92] Jay Jennings and Natalie Jomini Stroud. 2021. Asymmetric Adjustment: Partisanship and Correcting Misinformation on Facebook. *New Media & Society* 0, 0 (June 2021), 146144482110217. https://doi.org/10.1177/14614448211021720

[93] Youngseung Jeon, Jaehoon Kim, Sohyun Park, Yunyong Ko, Seongeun Ryu, Sang-Wook Kim, and Kyungsik Han. 2024. HearHere: Mitigating Echo Chambers in News Consumption through an AI-based Web System. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 1–34. https://doi.org/10.1145/3637340

[94] Jennifer Jerit and Yangzi Zhao. 2020. Political Misinformation. *Annual Review of Political Science* 23, 1 (May 2020), 77–94. https://doi.org/10.1146/annurev-polisci-050718-032814

[95] Chenyan Jia, Alexander Boltz, Angie Zhang, Anqing Chen, and Min Kyung Lee. 2022. Understanding Effects of Algorithmic vs. Community Label on Perceived Accuracy of Hyper-partisan Misinformation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 371:1–371:27. https://doi.org/10.1145/3555096

[96] Gargi Joshi, Ananya Srivastava, Bhargav Yagnik, Mohammed Hasan, Zainuddin Saiyed, Lubna A. Gabralla, Ajith Abraham, Rahee Walambe, and Ketan Kotecha. 2023. Explainable Misinformation Detection across Multiple Social Media Platforms. *IEEE access : practical innovations, open solutions* 11 (2023), 23634–23646. https://doi.org/10.1109/ACCESS.2023.3251892

[97] Alireza Karduni, Isaac Cho, Ryan Wesslen, Sashank Santhanam, Svitlana Volkova, Dustin L Arendt, Samira Shaikh, and Wenwen Dou. 2019. Vulnerable to Misinformation? Verifi!. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 312–323. https://doi.org/10.1145/3301275.3302320

[98] Sabrina Heike Kessler and Eva Bachmann. 2022. Debunking Health Myths on the Internet: The Persuasive Effect of (Visual) Online Communication. *Journal of Public Health-Heidelberg* 30, 8 (Aug. 2022), 1823–1835. https://doi.org/10.1007/s10389-022-01694-3

[99] Yash Khivasara, Yash Khare, and Tejas Bhadane. 2020. Fake News Detection System Using Web-Extension. In *2020 IEEE Pune Section International Conference (PuneCon)*. IEEE, India, 119–123. https://doi.org/10.1109/PuneCon50868.2020.9362384

[100] Antino Kim, 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 Managment Information Systems* 36, 3 (July 2019), 931–968. https://doi.org/10.1080/07421222.2019.1628921

[101] Paul Kim, Ziyu Fan, Lance Fernando, Jacques Sham, Crystal Sun, Yixin Sun, Brian Wright, Xi Yang, Nicholas Ross, and Diane Myung-kyung Woodbridge. 2019. Controversy Score Calculation for News Articles. In *2019 First International Conference on Transdisciplinary AI (TransAI)*. IEEE, CA, USA, 56–63. https://doi.org/10.1109/TransAI46475.2019.00018

[102] Sojung Claire Kim, Emily K. Vraga, and John Cook. 2021. An Eye Tracking Approach to Understanding Misinformation and Correction Strategies on Social Media: The Mediating Role of Attention and Credibility to Reduce HPV Vaccine Misperceptions. *Health Communication* 36, 13 (Nov. 2021), 1687–1696. https://doi.org/10.1080/10410236.2020.1787933

[103] Yeongdae Kim, Takane Ueno, Katie Seaborn, Hiroki Oura, Jacqueline Urakami, and Yuto Sawa. 2023. Exoskeleton for the Mind: Exploring Strategies against Misinformation with a Metacognitive Agent. In *Proceedings of the Augmented Humans International Conference 2023 (AHs '23)*. Association for Computing Machinery, New York, NY, USA, 209–220. https://doi.org/10.1145/3582700.3582725

[104] Jan Kirchner and Christian Reuter. 2020. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–27. https://doi.org/10.1145/3415211

[105] Timo K. Koch, Lena Frischlich, and Eva Lermer. 2023. Effects of Fact-Checking Warning Labels and Social Endorsement Cues on Climate Change Fake News Credibility and Engagement on Social Media. *Journal of Applied Social Psychology* 53, 6 (2023), 495–507. https://doi.org/10.1111/jasp.12959

[106] Nadejda Komendantova, Love Ekenberg, Mattias Svahn, Aron Larsson, Syed Iftikhar Hussain Shah, Myrsini Glinos, Vasilis Koulolias, and Mats Danielson. 2021. A Value-Driven Approach to Addressing Misinformation in Social Media. *Humanities and Social Sciences Communications* 8, 1 (Jan. 2021), 1–12. https://doi.org/10.1057/s41599-020-00702-9

[107] Sarah E. Kreps and Douglas L. Kriner. 2022. The COVID-19 Infodemic and the Efficacy of Interventions Intended to Reduce Misinformation. *Public Opinion Quarterly* 86, 1 (March 2022), 162–175. https://doi.org/10.1093/poq/nfab075

[108] Jiyoung Lee. 2022. The Effect of Web Add-on Correction and Narrative Correction on Belief in Misinformation Depending on Motivations for Using Social Media. *Behaviour & Information Technology* 41, 3 (2022), 629–643.

[109] Jiyoung Lee and Kim Bissell. 2023. User Agency-Based versus Machine Agency-Based Misinformation Interventions: The Effects of Commenting and AI Fact-Checking Labeling on Attitudes toward the COVID-19 Vaccination. *New Media & Society* (2023), 1–21. https://doi.org/10.1177/14614448231163228

[110] Jiyoung Lee and Kim Bissell. 2024. Correcting Vaccine Misinformation on Social Media: The Inadvertent Effects of Repeating Misinformation within Such Corrections on COVID-19 Vaccine Misperceptions. *Current Psychology* (2024), 1–13. https://doi.org/10.1007/s12144-024-05651-z

[111] Seongmin Lee, Sadia Afroz, Haekyu Park, Zijie J. Wang, Omar Shaikh, Vibhor Sehgal, Ankit Peshin, and Duen Horng Chau. 2022. MisVis: Explaining Web Misinformation Connections via Visual Summary. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 228, 6 pages. https://doi.org/10.1145/3491101.3519711

[112] Yang-Jun Li, Jens Marga, Christy Cheung, Xiao-Liang Shen, and Matthew Lee. 2022. Health Misinformation on Social Media: A Systematic Literature Review and Future Research Directions. *AIS Transactions on Human-Computer Interaction* 14, 2 (June 2022), 116–149. https://doi.org/10.17705/1thci.00164

[113] Raymond Liaw, Ari Zilnik, Mark Baldwin, and Stephanie Butler. 2013. Maater: Crowdsourcing to Improve Online Journalism. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*. ACM Press, Paris, France, 2549. https://doi.org/10.1145/2468356.2468828

[114] Helen M Lillie, Chelsea L Ratcliff, Andy J King, Manusheela Pokharel, and Jakob D Jensen. 2024. Using Narratives to Correct Politically Charged Health Misinformation and Address Affective Belief Echoes. *Journal of Public Health* (April 2024), 1–7. https://doi.org/10.1093/pubmed/fdae050

[115] Gionnieve Lim and Simon T. Perrault. 2023. Effects of Automated Misinformation Warning Labels on the Intents to like, Comment and Share Posts. In *Proceedings of the 11th International Conference on Human-Agent Interaction (HAI '23)*. Association for Computing Machinery, New York, NY, USA, 299–305. https://doi.org/10.1145/3623809.3623856

[116] Gionnieve Lim and Simon T. Perrault. 2023. Intents and Motivations to Like, Comment and Share Posts With Warnings of Misinformation. In *Proceedings of the 35th Australian Computer-Human Interaction Conference*. ACM, Wellington New Zealand, 108–113. https://doi.org/10.1145/3638380.3638390

[117] Xingyu Liu, Li Qi, Laurent Wang, and Miriam J. Metzger. 2023. Checking the Fact-Checkers: The Role of Source Type, Perceived Credibility, and Individual Differences in Fact-Checking Effectiveness. *Communication Research* (2023), 1–28. https://doi.org/10.1177/00936502231206419

[118] Kuan-Chieh Lo, Shih-Chieh Dai, Aiping Xiong, Jing Jiang, and Lun-Wei Ku. 2021. All the Wiser: Fake News Intervention Using User Reading Preferences. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. ACM, Virtual Event Israel, 1069–1072. https://doi.org/10.1145/3437963.3441696

[119] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2022. The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–27. https://doi.org/10.1145/3555562

[120] Cameron Martel, 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 (2021), 120–133. https://doi.org/10.17645/mac.v9i1.3519

[121] Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, and Preslav Nakov. 2020. Prta: A System to Support the Analysis of Propaganda Techniques in the News. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 287–293. arXiv:2005.05854

[122] Paul Mena. 2020. Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. *Policy and Internet* 12, 2 (June 2020), 165–183. https://doi.org/10.1002/poi3.214

[123] Martin Mende, Valentina O. Ubal, Marina Cozac, Beth Vallen, and Christopher Berry. 2024 JAN 2024. Fighting Infodemics: Labels as Antidotes to Mis- and Disinformation?! *Journal of Public Policy and Marketing* 43, 1, SI (2024 JAN 2024), 31–52. https://doi.org/10.1177/07439156231184816

[124] Yisroel Mirsky and Wenke Lee. 2022. The Creation and Detection of Deepfakes: A Survey. *Comput. Surveys* 54, 1 (Jan. 2022), 1–41. https://doi.org/10.1145/3425780

[125] Won-Ki Moon, Myojung Chung, and S. Mo. Jones-Jang. 2023. How Can We Fight Partisan Biases in the COVID-19 Pandemic? AI Source Labels on Fact-checking Messages Reduce Motivated Reasoning. *Mass Communication and Society* 26, 4 (2023), 646–670. https://doi.org/10.1080/15205436.2022.2097926

[126] Patricia L. Moravec, Antino Kim, and Alan R. Dennis. 2020. Appealing to Sense and Sensibility: System 1 and System 2 Interventions for Fake News on Social Media. *Information Systems Research* 31, 3 (Sept. 2020), 987–1006. https://doi.org/10.1287/isre.2020.0927

[127] Elmie Nekmat. 2020. Nudge Effect of Fact-Check Alerts: Source Influence and Media Skepticism on Sharing of News Misinformation in Social Media. *Social Media + Society* 6, 1 (Jan. 2020), 8–14. https://doi.org/10.1177/2056305119897322

[128] Brendan Nyhan and Jason Reifler. 2010. When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior* 32, 2 (June 2010), 303–330. https://doi.org/10.1007/s11109-010-9112-2

[129] Pinar Ozturk, Huaye Li, and Yasuaki Sakamoto. 2015. Combating Rumor Spread on Social Media: The Effectiveness of Refutation and Warning. In *2015 48th Hawaii International Conference on System Sciences*. IEEE, HI, USA, 2406–2414. https://doi.org/10.1109/HICSS.2015.288

[130] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *Systematic Reviews* 10, 1 (March 2021), 89. https://doi.org/10.1186/s13643-021-01626-4

[131] Orestis Papakyriakopoulos and Ellen Goodman. 2022. The Impact of Twitter Labels on Misinformation Spread and User Engagement: Lessons from Trump's Election Tweets. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*. Association for Computing Machinery, New York, NY, USA, 2541–2551. https://doi.org/10.1145/3485447.3512126

[132] Saumya Pareek and Jorge Goncalves. 2024. Peer-Supplied Credibility Labels as an Online Misinformation Intervention. *International Journal of Human-Computer Studies* 188 (Aug. 2024), 1–17. https://doi.org/10.1016/j.ijhcs.2024.103276

[133] Sungkyu Park, Jamie Yejean Park, Hyojin Chin, Jeong-han Kang, and Meeyoung Cha. 2021. An Experimental Study to Understand User Experience and Perception Bias Occurred by Fact-checking Messages. In *Proceedings of the Web Conference 2021*. ACM, Ljubljana Slovenia, 2769–2780. https://doi.org/10.1145/3442381.3450121

[134] Irene V. Pasquetto, Eaman Jahani, Shubham Atreja, and Matthew Baum. 2022. Social Debunking of Misinformation on WhatsApp: The Case for Strong and In-group Ties. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (March 2022), 1–35. https://doi.org/10.1145/3512964

[135] Gordon Pennycook, 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 (Nov. 2020), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478

[136] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. 2021. Shifting Attention to Accuracy Can Reduce Misinformation Online. *Nature* 592, 7855 (April 2021), 590–595. https://doi.org/10.1038/s41586-021-03344-2

[137] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson J. Lu, and David G. Rand. 2020. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science* 31, 7 (July 2020), 770–780. https://doi.org/10.1177/0956797620939054

[138] Gordon Pennycook and David G. Rand. 2022. Accuracy Prompts Are a Replicable and Generalizable Approach for Reducing the Spread of Misinformation. *Nature Communications* 13, 2333 (April 2022), 1–12. https://doi.org/10.1038/s41467-022-30073-5

[139] Raunak M. Pillai and Lisa K. Fazio. 2023. Explaining Why Headlines Are True or False Reduces Intentions to Share False Information. *Collabra-Psychology* 9, 87617 (2023), 1–11. https://doi.org/10.1525/collabra.87617

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

[141] Ethan Porter, Yamil Velez, and Thomas J. Wood. 2022. Factual Corrections Eliminate False Beliefs about COVID-19 Vaccines. *Public Opinion Quarterly* 86, 3 (2022), 762–773. https://doi.org/10.1093/poq/nfac034

[142] Ethan Porter and Thomas J. Wood. 2022. Political Misinformation and Factual Corrections on the Facebook News Feed: Experimental Evidence. *Journal of Politics* 84, 3 (July 2022), 1812–1817. https://doi.org/10.1086/719271

[143] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 231–240. https://doi.org/10.18653/v1/P18-1022

[144] Pardis Pourghomi, Fadi Safieddine, Wassim Masri, and Milan Dordevic. 2017. How to Stop Spread of Misinformation on Social Media: Facebook Plans vs. Right-Click Authenticate Approach. In *2017 International Conference on Engineering & MIS (ICEMIS)*. IEEE, Monastir, 1–8. https://doi.org/10.1109/ICEMIS.2017.8272957

[145] Clara Pretus, Ali M. Javeed, Diana Hughes, Kobi Hackenburg, Manos Tsakiris, Oscar Vilarroya, and Jay J. Van Bavel. 2024. The ¡i¿Misleading¡/I¿ Count: An Identity-Based Intervention to Counter Partisan Misinformation Sharing. *Philisophical Transactions of the Royal Society B- Biological Sciences* 379, 20230040 (2024), 1–9. https://doi.org/10.1098/rstb.2023.0040

[146] Toby Prike, Lucy H. Butler, and Ullrich K. H. Ecker. 2024. Source-Credibility Information and Social Norms Improve Truth Discernment and Reduce Engagement with Misinformation Online. *Scientific Reports* 14, 1 (March 2024), 1–11. https://doi.org/10.1038/s41598-024-57560-7

[147] Toby Prike and Ullrich K. H. Ecker. 2023 DEC 2023. Effective Correction of Misinformation. *Current Opinion in Psychology* 54, 101712 (2023 DEC 2023), 1–6. https://doi.org/10.1016/j.copsyc.2023.101712

[148] Sijia Qian, Cuihua Shen, and Jingwen Zhang. 2023. Fighting Cheapfakes: Using a Digital Media Literacy Intervention to Motivate Reverse Search of out-of-Context Visual Misinformation. *Journal of Computer-Mediated Communication* 28, 1 (Jan. 2023), 1–12. https://doi.org/10.1093/jcmc/zmac024

[149] Dipti P. Rana, Isha Agarwal, and Anjali More. 2018. A Review of Techniques to Combat The Peril of Fake News. In *Proceedings of the 2018 4th International Conference on Computing Communication and Automation (ICCCA)*. Greater Noida, India, 1–7. https://doi.org/10.1109/CCAA.2018.8777676

[150] Patrick R. Rich 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 (Sept. 2020), 310–322. https://doi.org/10.1016/j.jarmac.2020.04.001

[151] Jon Roozenbeek, Eileen Culloty, and Jane Suiter. 2023. Countering Misinformation Evidence, Knowledge Gaps, and Implications of Current Interventions. *European Psychologist* 28, 3 (July 2023), 189–205. https://doi.org/10.1027/1016-9040/a000492

[152] Jon Roozenbeek and Sander van der Linden. 2019. Fake News Game Confers Psychological Resistance against Online Misinformation. *Palgrave Communications* 5, 1 (Dec. 2019), 65. https://doi.org/10.1057/s41599-019-0279-9

[153] Margie Ruffin, Gang Wang, and Kirill Levchenko. 2022. Explaining Why Fake Photos Are Fake: Does It Work? *Proc. ACM Hum.-Comput. Interact.* 7, GROUP, Article 8 (Dec. 2022), 22 pages. https://doi.org/10.1145/3567558

[154] Fadi Safieddine, Wassim Masri, and Pardis Pourghomi. 2016. Corporate Responsibility in Combating Online Misinformation. *International Journal of Advanced Computer Science and Applications* 7, 2 (2016), 126–132. https://doi.org/10.14569/IJACSA.2016.070217

[155] Nina Sakhnini and Debaleena Chattopadhyay. 2022. A Review of Smartphone Fact-Checking Apps and Their (Non) Use Among Older Adults. In *Adjunct Publication of the 24th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '22)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3528575.3551448

[156] Emily Saltz, Soubhik Barari, Claire Leibowicz, and Claire Wardle. 2021. Misinformation Interventions Are Common, Divisive, and Poorly Understood. *Harvard Kennedy School Misinformation Review* 2, 5 (Oct. 2021), 1–25. https://doi.org/10.37016/mr-2020-81

[157] Emily Saltz, Claire R Leibowicz, and Claire Wardle. 2021. Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–6. https://doi.org/10.1145/3411763.3451807

[158] Jasmyne A. Sanderson, Vanessa Bowden, Briony Swire-Thompson, Stephan Lewandowsky, and Ullrich K. H. Ecker. 2023 SEP 2023. Listening to Misinformation While Driving: Cognitive Load and the Effectiveness of (Repeated) Corrections. *Journal of Applied Research in Memory and Cognition* 12, 3 (2023 SEP 2023), 325–334. https://doi.org/10.1037/mac0000057

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

[160] Leonie Schaewitz and Nicole Kramer. 2020. Combating Disinformation: Effects of Timing and Correction Format on Factual Knowledge and Personal Beliefs. In *Multidisciplinary International Symposium on Disinformation in Open Online Media*, Max VanDuijn, Mike Preuss, Viktoria Spaiser, Frank Takes, and Suzan Verberne (Eds.), Vol. 12259. Springer, Cham, 233–245. https://doi.org/10.1007/978-3-030-61841-4_16

[161] Lisa Scharrer, Vanessa Pape, and Marc Stadtler. 2022. Watch Out: Fake! How Warning Labels Affect Laypeople's Evaluation of Simplified Scientific Misinformation. *Discourse Processes* 59, 8 (Sept. 2022), 575–590. https://doi.org/10.1080/0163853X.2022.2096364

[162] Philipp Schmid and Cornelia Betsch. 2022. Benefits and Pitfalls of Debunking Interventions to Counter mRNA Vaccination Misinformation during the COVID-19 Pandemic. *Science Communication* 44, 5 (Oct. 2022), 531–558. https://doi.org/10.1177/10755470221129608

[163] Stefka Schmid, Katrin Hartwig, Robert Cieslinski, and Christian Reuter. 2022. Digital Resilience in Dealing with Misinformation on Social Media during COVID-19 A Web Application to Assist Users in Crises. *Information Systems Frontiers* 2022 (2022), 1–23. https://doi.org/10.1007/s10796-022-10347-5

[164] Guido Schryen, Gerit Wagner, Alexander Benlian, and Guy Paré. 2020. A Knowledge Development Perspective on Literature Reviews: Validation of a New Typology in the IS Field. *Communications of the Association for Information Systems* 46, 7 (2020), 134–186. https://doi.org/10.17705/1CAIS.04607

[165] Haeseung Seo, Aiping Xiong, and Dongwon Lee. 2019. Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation. In *Proceedings of the 10th ACM Conference on Web Science*. ACM, Boston Massachusetts USA, 265–274. https://doi.org/10.1145/3292522.3326012

[166] Filipo Sharevski, Amy Devine, Peter Jachim, and Emma Pieroni. 2022. Meaningful Context, a Red Flag, or Both? Preferences for Enhanced Misinformation Warnings among US Twitter Users. In *Proceedings of the 2022 European Symposium on Usable Security (EuroUSEC '22)*. Association for Computing Machinery, New York, NY, USA, 189–201. https://doi.org/10.1145/3549015.3555671

[167] Filipo Sharevski and Donald Gover. 2021. Two Truths and a Lie: Exploring Soft Moderation of COVID-19 Misinformation with Amazon Alexa. In *The 16th International Conference on Availability, Reliability and Security*. ACM, Vienna Austria, 1–9. https://doi.org/10.1145/3465481.3470017

[168] Filipo Sharevski and Aziz N Zeidieh. 2023. "I Just Didn't Notice It:" Experiences with Misinformation Warnings on Social Media amongst Users Who Are Low Vision or Blind. In *Proceedings of the 2023 New Security Paradigms Workshop (NSPW '23)*. Association for Computing Machinery, New York, NY, USA, 17–33. https://doi.org/10.1145/3633500.3633502

[169] Zien Sheikh Ali, Watheq Mansour, Fatima Haouari, Maram Hasanain, Tamer Elsayed, and Abdulaziz Al-Ali. 2023. Tahaqqaq: A Real-Time System for Assisting Twitter Users in Arabic Claim Verification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 3170–3174. https://doi.org/10.1145/3539618.3591815

[170] Imani N. Sherman, Jack W. Stokes, and Elissa M. Redmiles. 2021. Designing Media Provenance Indicators to Combat Fake Media. In *24th International Symposium on Research in Attacks, Intrusions and Defenses*. ACM, San Sebastian Spain, 324–339. https://doi.org/10.1145/3471621.3471860

[171] Anu Shrestha and Francesca Spezzano. 2020. Online Misinformation: From the Deceiver to the Victim. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '19)*. Association for Computing Machinery, New York, NY, USA, 847–850. https://doi.org/10.1145/3341161.3343536

[172] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD explorations newsletter* 19, 1 (Sept. 2017), 15.

[173] Yvonne Skipper, Daniel Jolley, and Joseph Reddington. 2023 NOV 2023. 'But Wait, That Isn't Real': A Proof-of-Concept Study Evaluating 'Project Real', a Co-Created Intervention That Helps Young People to Spot Fake News Online. *British Journal of Development Psychology* 41, 4 (2023 NOV 2023), 371–384. https://doi.org/10.1111/bjdp.12456

[174] Ciarra N. Smith and Holly H. Seitz. 2019. Correcting Misinformation About Neuroscience via Social Media. *Science Communication* 41, 6 (Dec. 2019), 790–819. https://doi.org/10.1177/1075547019890073

[175] Yunya Song, Sai Wang, and Qian Xu. 2022. Fighting Misinformation on Social Media: Effects of Evidence Type and Presentation Mode. *Health Education Research* 37, 3 (2022), 185–198. https://doi.org/10.1093/her/cyac011

[176] Catherine Sotirakou, Theodoros Paraskevas, and Constantinos Mourlas. 2022. Toward the Design of a Gamification Framework for Enhancing Motivation Among Journalists, Experts, and the Public to Combat Disinformation: The Case of CALYPSO Platform. In *HCI in Games (Lecture Notes in Computer Science)*, Xiaowen Fang (Ed.). Springer International Publishing, Cham, 542–554. https://doi.org/10.1007/978-3-031-05637-6_35

[177] M. Connor Sullivan. 2019. Leveraging Library Trust to Combat Misinformation on Social Media. *Library & Information Science Research* 41, 1 (Jan. 2019), 2–10. https://doi.org/10.1016/j.lisr.2019.02.004

[178] Yuko Tanaka and Rumi Hirayama. 2019. Exposure to Countering Messages Online: Alleviating or Strengthening False Belief? *Cyberpsychology, Behavior, and Social Networking* 22, 11 (Nov. 2019), 742–746. https://doi.org/10.1089/cyber.2019.0227

[179] Yuko Tanaka, Yasuaki Sakamoto, and Toshihiko Matsuka. 2013. Toward a Social-Technological System That Inactivates False Rumors through the Critical Thinking of Crowds. In *2013 46th Hawaii International Conference on System Sciences*. IEEE, Wailea, HI, USA, 649–658. https://doi.org/10.1109/HICSS.2013.557

[180] Ran Tao, Jianing Li, Liwei Shen, and Sijia Yang. 2023. Hope over Fear: The Interplay between Threat Information and Hope Appeal Corrections in Debunking Early COVID-19 Misinformation. *Social Science & Medicine* 333, 116132 (2023). https://doi.org/10.1016/j.socscimed.2023.116132

[181] Thaler, Richard H. and Sunstein, Cass R. 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness.* Penguin Books, London, United Kingdom.

[182] Calum Thornhill, Quentin Meeus, Jeroen Peperkamp, and Bettina Berendt. 2019. A Digital Nudge to Counter Confirmation Bias. *Frontiers in Big Data* 2 (June 2019), 11. https://doi.org/10.3389/fdata.2019.00011

[183] Yu-Chia Tseng, Nanyi Bi, Yung-Ju Chang, and Chien Wen (Tina) Yuan. 2022. Investigating Correction Effects of Different Modalities for Misinformation about COVID-19. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing (CSCW'22 Companion)*. Association for Computing Machinery, New York, NY, USA, 54–58. https://doi.org/10.1145/3500868.3559455

[184] Gleb Tsipursky, Fabio Votta, and Kathryn M Roose. 2018. Fighting Fake News and Post-Truth Politics with Behavioral Science: The Pro-Truth Pledge. *Behavior and Social Issues* 27, 1 (2018), 47–70.

[185] Marina Tulin, Michael Hameleers, Claes de Vreese, Michael Opgenhaffen, and Ferre Wouters. 2024. Beyond Belief Correction: Effects of the Truth Sandwich on Perceptions of Fact-Checkers and Verification Intentions. *Journalism Practice* (2024), 1–21. https://doi.org/10.1080/17512786.2024.2311311

[186] Melissa Tully, 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 (Oct. 2020), 1–12. https://doi.org/10.1177/2056305120978377

[187] Melissa Tully, Emily K. Vraga, and Leticia Bode. 2020. Designing and Testing News Literacy Messages for Social Media. *Mass Communication and Society* 23, 1 (Jan. 2020), 22–46. https://doi.org/10.1080/15205436.2019.1604970

[188] Toni G. L. A. van der Meer, Michael Hameleers, and Jakob Ohme. 2023. Can Fighting Misinformation Have a Negative Spillover Effect? How Warnings for the Threat of Misinformation Can Decrease General News Credibility. *Journalism Studies* 24, 6 (2023), 803–823. https://doi.org/10.1080/1461670X.2023.2187652

[189] Tony G. L. A. van der Meer 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 (April 2020), 560–575. https://doi.org/10.1080/10410236.2019.1573295

[190] Abigail T. Velasco, Allen Roi C. Cortez, John Meynard B. Camay, Ian Michael C. Giba, and Marlon A. Diloy. 2023. Factit: A Fact-Checking Browser Extension. In *2023 IEEE 12th International Conference on Educational and Information Technology (ICEIT)*. IEEE, Chongqing, China, 342–347. https://doi.org/10.1109/ICEIT57125.2023.10107833

[191] Yamil R. Velez, Ethan Porter, and Thomas J. Wood. 2023. Latino-Targeted Misinformation and the Power of Factual Corrections. *Journal of Politics* 85, 2 (2023). https://doi.org/10.1086/722345

[192] Christian von der Weth, Jithin Vachery, and Mohan Kankanhalli. 2020. Nudging Users to Slow Down the Spread of Fake News in Social Media. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, London, United Kingdom, 1–6. https://doi.org/10.1109/ICMEW46912.2020.9106003

[193] Emily K. Vraga and Leticia Bode. 2018. I Do Not Believe You: How Providing a Source Corrects Health Misperceptions across Social Media Platforms. *Information, Communication & Society* 21, 10 (Oct. 2018), 1337–1353. https://doi.org/10.1080/1369118X.2017.1313883

[194] Emily K. Vraga, Leticia Bode, and Melissa Tully. 2021. The Effects of a News Literacy Video and Real-Time Corrections to Video Misinformation Related to Sunscreen and Skin Cancer. *Health Communication* 37, 13 (2021), 1622–1630.

[195] Emily K. Vraga, Sojung Claire Kim, John Cook, and Leticia Bode. 2020. Testing the Effectiveness of Correction Placement and Type on Instagram. *The International Journal of Press/Politics* 25, 4 (Oct. 2020), 632–652. https://doi.org/10.1177/1940161220919082

[196] Emily K. Vraga, Melissa Tully, and Leticia Bode. 2021. Assessing the Relative Merits of News Literacy and Corrections in Responding to Misinformation on Twitter. *New Media & Society* 24, 10 (2021), 2354–2371. https://doi.org/10.1177/1461444821998691

[197] EK Vraga and Leticia Bode. 2017. Using Expert Sources to Correct Health Misinformation in Social Media. *Science Communication* 39, 5 (Oct. 2017), 621–645. https://doi.org/10.1177/1075547017731776

[198] EK Vraga, L Bode, and M Tully. 2022. Creating News Literacy Messages to Enhance Expert Corrections of Misinformation on Twitter. *Communication Research and Practice* 49, 2 (2022), 245–267.

[199] EK Vraga, SC Kim, and J Cook. 2019. Testing Logic-based and Humor-based Corrections for Science, Health, and Political Misinformation on Social Media. *Jourbal of Broadcasting & Electronic Media* 63, 3 (July 2019), 393–414. https://doi.org/10.1080/08838151.2019.1653102

[200] Christopher N. Wahlheim, Timothy R. Alexander, and Carson D. Peske. 2020. Reminders of Everyday Misinformation Statements Can Enhance Memory for and Beliefs in Corrections of Those Statements in the Short Term. *Psychological Science* 31, 10 (Oct. 2020), 1325–1339. https://doi.org/10.1177/0956797620952797

[201] Franz Waltenberger, Simon Höferlin, and Michael Froehlich. 2023. Reddit Insights: Improving Online Discussion Culture by Contextualizing User Profiles. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 245, 6 pages. https://doi.org/10.1145/3544549.3585671

[202] Austin Horng-En Wang. 2022. PM Me the Truth? The Conditional Effectiveness of Fact-Checks across Social Media Sites. *Social Media + Society* 8, 098347 (April 2022), 1–16. https://doi.org/10.1177/20563051221098347

[203] Ran Wang, Kehan Du, Qianhe Chen, Yifei Zhao, Mojie Tang, Hongxi Tao, Shipan Wang, Yiyao Li, and Yong Wang. 2022. RumorLens: Interactive Analysis and Validation of Suspected Rumors on Social Media. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 232, 7 pages. https://doi.org/10.1145/3491101.3519712

[204] Weirui Wang and Yan Huang. 2021. 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* 43, 2 (April 2021), 170–198. https://doi.org/10.1177/1075547020974384

[205] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. *Social Science & Medicine* 240 (Nov. 2019), 112552. https://doi.org/10.1016/j.socscimed.2019.112552

[206] Colin Ware. 2012. *Information Visualization: Perception for Design* (4 ed.). Elsevier, Cambridge.

[207] Victoria Westbrook, Duane T. Wegener, and Mark W. Susmann. 2023. Mechanisms in Continued Influence: The Impact of Misinformation Corrections on Source Perceptions. *Memory & Cognition* 51, 6 (2023), 1317–1330. https://doi.org/10.3758/s13421-023-01402-w

[208] Winnifred Wijnker, Ionica Smeets, Peter Burger, and Sanne Willems. 2022. Debunking Strategies for Misleading Bar Charts. *JCOM - Journal of Science Communication* 21, 7 (2022), 1–27. https://doi.org/10.22323/2.21070207

[209] Tom Wilson, Kaitlyn Zhou, and Kate Starbird. 2018. Assembling Strategic Narratives: Information Operations as Collaborative Work within an Online Community. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–26. https://doi.org/10.1145/3274452

[210] Reed M. Wood, Marie Juanchich, Mark Ramirez, and Shenghao Zhang. 2023. Promoting COVID-19 Vaccine Confidence through Public Responses to Misinformation: The Joint Influence of Message Source and Message Content. *Social Science & Medicine* 324, 115863 (May 2023), 1–11. https://doi.org/10.1016/j.socscimed.2023.115863

[211] Thomas Wood and Ethan Porter. 2019. The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. *Political Behavior* 41, 1 (March 2019), 135–163. https://doi.org/10.1007/s11109-018-9443-y

[212] Fan Xu, Victor S. Sheng, and Mingwen Wang. 2021. A Unified Perspective for Disinformation Detection and Truth Discovery in Social Sensing: A Survey. *Comput. Surveys* 55, 1 (Nov. 2021), 6:1–6:33. https://doi.org/10.1145/3477138

[213] Wan Yit Yong, Rajesh Jaiswal, and Fernando Perez Tellez. 2023. Explainability in NLP Model: Detection of Covid-19 Twitter Fake News. In *Proceedings of the 2023 Conference on Human Centered Artificial Intelligence: Education and Practice (HCAIep '23)*. Association for Computing

Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3633083.3633212

[214] Himanshu Zade, Megan Woodruff, Erika Johnson, Mariah Stanley, Zhennan Zhou, Minh Tu Huynh, Alissa Elizabeth Acheson, Gary Hsieh, and Kate Starbird. 2023. Tweet Trajectory and AMPS-based Contextual Cues Can Help Users Identify Misinformation. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 103 (April 2023), 27 pages. https://doi.org/10.1145/3579536

[215] Amy X. Zhang, Martin Robbins, Ed Bice, Sandro Hawke, David Karger, An Xiao Mina, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, Emmanuel Vincent, and Jennifer Lee. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*. ACM Press, Lyon, France, 603–612. https://doi.org/10.1145/3184558.3188731

[216] Yuqi Zhang, Bin Guo, Yasan Ding, Jiaqi Liu, Chen Qiu, Sicong Liu, and Zhiwen Yu. 2022. Investigation of the Determinants for Misinformation Correction Effectiveness on Social Media during COVID-19 Pandemic. *Information Processing & Management* 59, 102935 (2022), 1–16. https://doi.org/10.1016/j.ipm.2022.102935

[217] Wenqing Zhao. 2019. Misinformation Correction across Social Media Platforms. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, Las Vegas, NV, USA, 1371–1376. https://doi.org/10.1109/CSCI49370.2019.00256

[218] Chengbo Zheng and Xiaojuan Ma. 2022. Evaluating the Effect of Enhanced Text-Visualization Integration on Combating Misinformation in Data Story. In *2022 IEEE 15th Pacific Visualization Symposium (PacificVis)*. IEEE, Tsukuba, Japan, 141–150. https://doi.org/10.1109/PacificVis53943.2022.00023

[219] Xinyi Zhou and Reza Zafarani. 2021. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *Comput. Surveys* 53, 5 (Sept. 2021), 1–40. https://doi.org/10.1145/3395046

[220] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2019. Detection and Resolution of Rumours in Social Media: A Survey. *Comput. Surveys* 51, 2 (March 2019), 1–36. https://doi.org/10.1145/3161603

A APPENDIX

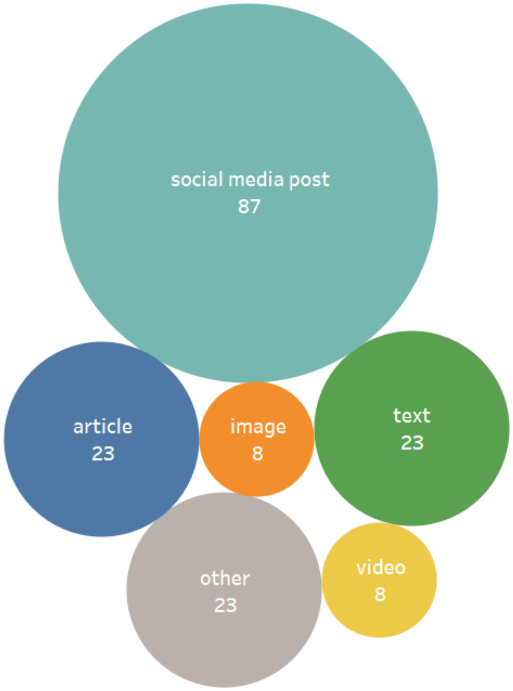


Fig. 9. Number of papers addressing specific formats.

B ELECTRONIC SUPPLEMENT

Table 2. Categorization of 172 publications regarding user intervention characteristics and methodological aspects.

Author	Year	Sample Size	Lab study	Online experiment	Field study	Survey	Interviews	Conceptual	Facebook	Twitter/X	General	Other	Social media posts	Articles	Text	Images	Video	Other	Warning	Correction/debunking	Showing indicators	(Binary) label	Highlighting design	Visibility reduction	Removal	Complicate sharing	Specific visualization	Other	Active	Passive	Neither/unclear	Pre exposure	During	At the moment of sharing	On request	Post exposure	Other	Mis-/Disinformation	Rumors	News credibility	Other	Browser extension/Plugin	Own platform	Game	Other
Author	**Year**	**Sample Size**	**Evaluation Type**						**Platform**				**Format**						**Intervention Design**										**Intera.**			**Timing**						**Concept**				**Implem.**			
Agley et al. [2]	2021	1,000		●							●							●										●	●			●						●						●	
Aird et al. [3]	2018	370	●	●							●							●		●										●			●								●			●	
Almaliki [4]	2019	100				●					●		●															●		●							●	●							●
Amin et al. [6]	2021	38		●								●						●					●			●			●				●	●				●					●		
Andi and Akesson [7]	2020	1,003				●					●		●						●										●				●					●							●
Ardevol-Abreu et al. [8]	2020	N1=31 N2=350				●	●				●		●						●					●					●				●					●							●
Aslett et al. [9]	2022	N1=3,862 N2=3,337 N3=968			●	●						●						●				●	●							●			●							●		●			
Autry and Duarte [11]	2021	N1=357 N2=75	●								●							●		●										●						●		●						●	
Axelsson et al. [12]	2021	N1=90 N2=119		●							●		●			●	●											●			●						●	●							●
Ayoub et al. [13]	2021	244		●							●				●						●									●			●					●					●		
Bachmann and Valenzuela [14]	2023	1,472		●					●				●							●	●									●			●					●							●
Barman and Colan [16]	2023	348		●							●		●						●	●										●			●					●							●
Barua et al. [17]	2019	-						●				●			●							●									●				●			●					●		
Bhuiyan et al. [18]	2021	N1=430 N2=12	●		●					●			●						●				●	●						●			●							●		●			
Bhuiyan et al. [19]	2021	31	●				●					●		●							●		●							●			●							●					●
Bhuiyan et al. [20]	2018	16			●					●			●						●				●	●						●			●							●		●			

Continued from previous page

Author	Year	Sample Size	Lab study	Online experiment	Field study	Survey	Interviews	Conceptual	Facebook	Twitter/X	General	Other	Social media posts	Articles	Text	Images	Video	Other	Warning	Correction/debunking	Showing indicators	(Binary) label	Highlighting design	Visibility reduction	Removal	Complicate sharing	Specific visualization	Other	Active	Passive	Neither/unclear	Pre exposure	During	At the moment of sharing	On request	Post exposure	Other	Mis-/Disinformation	Rumors	News credibility	Other	Browser extension/Plugin	Own platform	Game	Other
Author	**Year**	**Sample Size**	**Evaluation Type**						**Platform**				**Format**						**Intervention Design**										**Intera.**			**Timing**						**Concept**				**Implem.**			
Bode and Vraga [22]	2015	N1=524 N2=500				●			●				●							●										●			●					●							●
Bode and Vraga [23]	2018	136		●					●				●							●										●			●					●				●			
Bozarth et al. [24]	2023	18					●					●	●							●	●	●									●		●					●							●
Brashier et al. [25]	2021	2,683		●							●		●									●								●		●	●			●		●							●
Buczel et al. [28]	2024	337		●							●			●					●										●	●		●				●		●							●
Capraro and Celadin [30]	2022	N1=550 N2=558 N3=550 N4=372		●					●				●															●			●			●				●							●
Caramancion [31]	2022	327		●							●							●		●	●							●	●			●						●							●
Challenger et al. [32]	2022	N1=1,291 N2=2,084		●							●							●		●											●						●				●				●
Chen et al. [34]	2022	10					●		●				●														●				●				●			●							●
Chen and Tang [35]	2022	415		●							●							●										●	●			●						●							●
Chiang et al. [37]	2022	60		●							●			●							●										●				●					●			●		
Clayton et al. [38]	2020	2,994				●			●				●						●									●		●		●	●					●							●
Craig and Vijaykumar [39]	2023	231		●								●			●					●								●		●		●				●		●							●
Dai [41]	2021	350		●							●				●					●								●		●		●				●		●							●
Dai et al. [40]	2022	N1=425 N2=625				●					●				●					●											●		●					●							●
Danry et al. [42]	2020	18	●									●						●										●			●		●					●							●
Denner et al. [43]	2023	211		●					●				●							●										●						●		●							●
Desai and Reimers [44]	2023	365		●						●			●							●										●						●		●							●
Dobber et al. [45]	2023	1,054		●								●					●		●			●							●	●		●	●								●				●
Domgaard and Park [46]	2021	250		●							●			●							●						●			●		●						●							●
Drolsbach and Pröllochs [47]	2023	7		●						●			●							●										●			●					●							●
Duncan [48]	2020	390				●						●			●													●		●			●							●					●
Ecker et al. [49]	2020	2,279		●								●		●						●										●						●		●							●

Continued from previous page

Author	Year	Sample Size	Lab study	Online experiment	Field study	Survey	Interviews	Conceptual	Facebook	Twitter/X	General	Other	Social media posts	Articles	Text	Images	Video	Other	Warning	Correction/debunking	Showing indicators	(Binary) label	Highlighting design	Visibility reduction	Removal	Complicate sharing	Specific visualization	Other	Active	Passive	Neither/unclear	Pre exposure	During	At the moment of sharing	On request	Post exposure	Other	Mis-/Disinformation	Rumors	News credibility	Other	Browser extension/Plugin	Own platform	Game	Other
Author	**Year**	**Sample Size**	**Evaluation Type**						**Platform**				**Format**						**Intervention Design**										**Intera.**			**Timing**						**Concept**				**Implem.**			
Ecker et al. [50]	2017	60	●									●		●						●										●						●					●				●
Ecker et al. [52]	2011	N1=161 N2=138	●									●		●						●										●						●					●				●
Ecker et al. [51]	2020	1718		●							●		●							●		●								●			●					●							●
Feng et al. [55]	2023	595		●						●						●	●				●									●			●					●							●
Figl et al. [56]	2023	256		●					●				●						●			●								●			●					●							●
Folkvord et al. [57]	2022	307		●					●				●						●									●	●			●						●							●
Freeze et al. [59]	2021	434		●								●		●					●											●						●		●							●
Furuta and Suzuki [60]	2021	-						●				●			●						●		●							●							●	●							●
Gao et al. [62]	2018	122		●							●		●						●									●		●			●					●		●			●		
Gesser-Edelsburg et al. [63]	2018	243	●	●					●				●							●										●						●					●				●
Grady et al. [64]	2021	418				●					●		●						●			●							●	●		●	●			●		●							●
Grandhi et al. [65]	2021	376				●					●		●								●										●						●			●					●
Guess et al. [67]	2020	N1=9,190 N2=4,669 N3=6,439	●	●					●				●								●									●		●						●							●
Guo et al. [69]	2023	28					●					●					●		●	●				●					●	●		●	●					●							●
Hameleers [70]	2020	1,091		●							●			●					●	●	●									●		●	●					●							●
Hameleers et al. [71]	2020	1,404		●						●			●							●										●			●					●							●
Hameleers and van der Meer [72]	2023	1,105		●		●					●				●				●		●								●			●						●							●
Hartwig et al. [74]	2024	N1=21 N2=18				●	●					●					●				●		●							●					●			●							●
Hartwig et al. [77]	2024	N1=44 N2=23				●	●			●			●								●		●							●			●					●							●
Hartwig et al. [76]	2024	20					●					●						●			●		●							●			●					●							●
Hawa et al. [78]	2021	-						●				●			●							●									●				●			●					●		
Heuer and Glassman [80]	2022	N1=188 N2=208		●							●			●							●		●					●		●			●					●							●
Horne et al. [81]	2019	-						●			●		●															●			●		●					●							●
Huang and Wang [82]	2020	N1=235 N2=235		●					●				●							●										●						●		●							●
Irving et al. [83]	2022	129		●							●				●					●										●			●					●							●

Continued from previous page

Author	Year	Sample Size	Lab study	Online experiment	Field study	Survey	Interviews	Conceptual	Facebook	Twitter/X	General	Other	Social media posts	Articles	Text	Images	Video	Other	Warning	Correction/debunking	Showing indicators	(Binary) label	Highlighting design	Visibility reduction	Removal	Complicate sharing	Specific visualization	Other	Active	Passive	Neither/unclear	Pre exposure	During	At the moment of sharing	On request	Post exposure	Other	Mis-/Disinformation	Rumors	News credibility	Other	Browser extension/Plugin	Own platform	Game	Other
Author	**Year**	**Sample Size**	**Evaluation Type**						**Platform**				**Format**						**Intervention Design**										**Intera.**			**Timing**						**Concept**				**Implem.**			
Jahanbakhsh and Karger [85]	2024	32		●		●					●		●									●					●			●			●					●				●			
Jahanbakhsh et al. [89]	2022	N1=154 N2=14			●	●					●		●									●	●				●		●				●	●				●					●		
Jahanbakhsh et al. [87]	2021	1,668				●					●		●													●			●					●				●							●
Jahanbakhsh et al. [86]	2023	61		●						●			●									●	●					●			●		●					●							●
Jahanbakhsh et al. [88]	2022	N1=27 N2=312			●	●					●			●						●							●			●			●					●				●			
Jahng et al. [90]	2021	205		●							●			●						●										●			●					●							●
Jennings and Stroud [92]	2021	N1=1,262 N2=1,586		●					●				●						●											●			●					●							●
Jeon et al. [93]	2024	N1=6 N2=94		●	●		●				●			●													●			●					●						●		●		
Jia et al. [95]	2022	1,677		●							●		●									●								●			●					●							●
Joshi et al. [96]	2023	-						●			●		●								●		●							●			●					●							●
Karduni et al. [97]	2019	5					●			●			●								●						●				●	●			●			●		●			●		
Kessler and Bachmann [98]	2022	700		●							●							●		●							●			●						●					●				●
Khivasara et al. [99]	2020	-						●				●			●							●							●						●			●				●			
Kim et al. [100]	2019	N1=590 N2=299				●			●				●															●		●			●					●		●					●
Kim et al. [101]	2019	-						●				●	●														●	●			●				●						●		●		
Kim et al. [102]	2021	92	●							●			●							●										●			●					●							●
Kim et al. [103]	2023	N1=17 N2=57 N3=49	●				●			●			●								●		●	●		●			●	●		●			●			●				●			
Kirchner and Reuter [104]	2020	N1=1,012 N2=15 N3=1,030				●			●				●						●	●				●		●			●	●		●	●	●				●							●
Koch et al. [105]	2023	571		●					●				●						●		●							●	●	●			●					●							●
Komendantova et al. [106]	2021	N1=103 N2=68 N3=50					●				●							●										●			●						●	●				●			
Kreps and Kriner [107]	2022	2,000		●							●							●	●	●		●		●					●				●					●							●
Lee [108]	2022	171		●					●				●							●		●								●						●		●				●			●
Lee et al. [111]	2022	-						●				●						●			●	●	●				●				●				●			●					●		
Lee and Bissell [109]	2023	377		●					●				●						●	●		●							●	●		●	●					●							●

Continued from previous page

Author	Year	Sample Size	Lab study	Online experiment	Field study	Survey	Interviews	Conceptual	Facebook	Twitter/X	General	Other	Social media posts	Articles	Text	Images	Video	Other	Warning	Correction/debunking	Showing indicators	(Binary) label	Highlighting design	Visibility reduction	Removal	Complicate sharing	Specific visualization	Other	Active	Passive	Neither/unclear	Pre exposure	During	At the moment of sharing	On request	Post exposure	Other	Mis-/Disinformation	Rumors	News credibility	Other	Browser extension/Plugin	Own platform	Game	Other
Author	**Year**	**Sample Size**	**Evaluation Type**						**Platform**				**Format**						**Intervention Design**										**Intera.**			**Timing**						**Concept**				**Implem.**			
Lee and Bissell [110]	2024	502		●					●				●							●										●			●			●		●							●
Liaw et al. [113]	2013	?		●								●		●						●			●							●			●					●				●			
Lillie et al. [114]	2024	469		●							●			●						●										●						●		●							●
Lim and Perrault [116]	2023	36		●							●		●						●			●								●			●					●							●
Lim and Perrault [115]	2023	200		●							●		●						●			●								●			●					●							●
Liu et al. [117]	2023	859		●						●			●							●										●						●		●							●
Lo et al. [118]	2021	89		●							●			●						●											●		●					●					●		
Lu et al. [119]	2022	N1=538 N2=1,098		●							●			●								●						●		●			●			●		●							●
Martel et al. [120]	2021	2,228		●					●	●			●							●										●						●		●							●
Martino et al. [121]	2020	-						●				●			●						●		●								●				●						●		●		
Mena [122]	2020	501		●					●				●						●											●			●					●							●
Moon et al. [125]	2022	354		●					●				●							●										●			●					●							●
Moravec et al. [126]	2020	398		●					●				●						●			●						●		●		●	●					●							●
Nekmat [127]	2020	929		●								●		●					●											●			●					●							●
Ozturk et al. [129]	2015	259				●				●			●						●	●										●			●						●						●
Papakyriakopoulos and Goodman [131]	2022	-						●		●			●						●			●								●			●					●							●
Pareek and Goncalves [132]	2024	96		●							●			●								●								●					●			●							●
Park et al. [133]	2021	11,145				●						●			●				●								●			●			●						●						●
Pasquetto et al. [134]	2022	N1=2,805 N2=25		●			●					●						●		●		●						●		●						●		●							●
Pennycook et al. [135]	2020	N1=5 271 N2=1,568		●					●				●						●			●								●			●					●							●
Pennycook et al. [137]	2020	N1=853 N2=856				●					●		●															●	●			●						●							●
Pennycook et al. [136]	2021	N>5,000			●	●				●			●															●		●		●				●		●							●
Pillai and Fazio [139]	2023	499		●														●										●		●						●		●							●

Continued from previous page

Author	Year	Sample Size	Lab study	Online experiment	Field study	Survey	Interviews	Conceptual	Facebook	Twitter/X	General	Other	Social media posts	Articles	Text	Images	Video	Other	Warning	Correction/debunking	Showing indicators	(Binary) label	Highlighting design	Visibility reduction	Removal	Complicate sharing	Specific visualization	Other	Active	Passive	Neither/unclear	Pre exposure	During	At the moment of sharing	On request	Post exposure	Other	Mis-/Disinformation	Rumors	News credibility	Other	Browser extension/Plugin	Own platform	Game	Other
Author	**Year**	**Sample Size**	**Evaluation Type**						**Platform**				**Format**						**Intervention Design**										**Intera.**			**Timing**						**Concept**				**Implem.**			
Pluviano et al. [140]	2017	120	●								●				●					●							●			●			●					●			●				●
Porter and Wood [142]	2022	N1=5,000 N2=2,000		●					●				●						●	●		●		●					●				●					●							●
Porter et al. [141]	2022	5,075		●							●		●							●										●			●					●							●
Pourghomi et al. [144]	2017	-						●	●				●			●											●			●					●			●							●
Pretus et al. [145]	2024	N1=1,709 N2=804		●						●			●						●		●							●		●			●					●							●
Prike et al. [146]	2024	415		●							●		●								●	●					●			●			●					●							●
Qian et al. [148]	2023	905		●							●					●											●		●	●		●						●							●
Rich and Zaragoza [150]	2020	N1=134 N2=134 N3=102		●								●		●						●										●						●		●							●
Ruffin et al. [153]	2022	N1=113 N2=543				●					●					●					●		●							●			●								●				●
Scharrer et al. [161]	2022	41		●							●				●				●			●								●		●	●					●							●
Safieddine et al. [154]	2016	-						●				●			●	●											●			●					●			●				●	●		
Sakhnini and Chattopadhyay [155]	2022	11					●				●							●		●	●	●									●				●			●					●		
Saltz et al. [157]	2021	N1=15 N2=23					●				●		●			●	●		●			●		●	●			●	●	●							●	●							●
Sangalang et al. [159]	2019	N1=385 N2=586		●								●			●					●										●						●		●							●
Schaewitz and Kramer [160]	2020	221		●								●		●						●										●						●		●							●
Schmid et al. [163]	2022	N1=9 N2=7		●			●			●			●								●	●					●				●				●			●					●		
Schmid and Betsch [162]	2022	N1=2,444 N2=817		●							●				●					●									●			●				●		●							●
Seo et al. [165]	2019	N1=522 N2=624		●							●		●						●											●			●					●							●
Sharevski and Gover [167]	2021	304				●				●								●										●	●	●			●					●							●
Sharevski and Zeidieh [168]	2023	29					●				●						●		●			●								●			●					●							●
Sharevski et al. [166]	2022	337		●						●			●						●											●			●					●							●
Sheikh Ali et al. [169]	2023	-						●		●					●					●	●	●	●							●					●			●							●

Continued from previous page

Author	Year	Sample Size	Lab study	Online experiment	Field study	Survey	Interviews	Conceptual	Facebook	Twitter/X	General	Other	Social media posts	Articles	Text	Images	Video	Other	Warning	Correction/debunking	Showing indicators	(Binary) label	Highlighting design	Visibility reduction	Removal	Complicate sharing	Specific visualization	Other	Active	Passive	Neither/unclear	Pre exposure	During	At the moment of sharing	On request	Post exposure	Other	Mis-/Disinformation	Rumors	News credibility	Other	Browser extension/Plugin	Own platform	Game	Other
Author	**Year**	**Sample Size**	**Evaluation Type**						**Platform**				**Format**						**Intervention Design**										**Intera.**			**Timing**						**Concept**				**Implem.**			
Sherman et al. [170]	2021	N1=24 N2=19 N3=1,456		●			●				●						●		●		●	●	●	●			●			●			●					●							●
Smith and Seitz [174]	2019	744		●					●									●		●										●			●					●			●				●
Song et al. [175]	2022	610		●					●				●							●										●			●					●							●
Sotirakou et al. [176]	2022	-						●			●							●		●		●						●			●				●			●					●		
Sullivan [177]	2019	N1=625 N2=600		●					●				●							●										●			●					●							●
Tanaka and Hirayama [178]	2019	164		●						●			●							●										●						●			●						●
Tanaka et al. [179]	2013	87		●						●			●							●										●		●				●			●						●
Tao et al. [180]	2023	836		●							●		●							●										●						●		●							●
Thornhill et al. [182]	2019	20				●				●					●					●		●	●							●			●					●					●		
Tseng et al. [183]	2022	210		●							●				●	●	●			●										●						●		●							●
Tsipursky et al. [184]	2018	21			●				●				●															●	●			●						●							●
Tulin et al. [185]	2024	752		●							●			●						●		●								●			●					●							●
Tully et al. [186]	2020	610		●						●			●							●										●			●					●							●
Tully et al. [187]	2020	N1=702 N2=787		●						●			●															●		●		●						●							●
van der Meer and Jin [189]	2020	700		●							●			●						●										●						●		●							●
van der Meer et al. [188]	2023	1,305		●							●			●					●	●								●		●						●		●							●
Velasco et al. [190]	2023	285				●					●				●							●									●				●			●				●			
Velez et al. [191]	2023	2,869		●							●		●							●										●						●		●							●
von der Weth et al. [192]	2020	-						●		●			●						●		●					●		●	●					●				●				●			
Vraga et al. [196]	2021	916		●						●			●							●								●		●			●					●							●
Vraga and Bode [193]	2018	1,384		●						●			●							●										●			●					●							●
Vraga and Bode [197]	2017	271		●					●	●			●							●										●			●					●							●
Vraga et al. [198]	2022	N1=1,207 N2=603		●						●			●							●								●		●			●					●							●

Continued from previous page

Author	Year	Sample Size	Lab study	Online experiment	Field study	Survey	Interviews	Conceptual	Facebook	Twitter/X	General	Other	Social media posts	Articles	Text	Images	Video	Other	Warning	Correction/debunking	Showing indicators	(Binary) label	Highlighting design	Visibility reduction	Removal	Complicate sharing	Specific visualization	Other	Active	Passive	Neither/unclear	Pre exposure	During	At the moment of sharing	On request	Post exposure	Other	Mis-/Disinformation	Rumors	News credibility	Other	Browser extension/Plugin	Own platform	Game	Other
Author	**Year**	**Sample Size**	**Evaluation Type**						**Platform**				**Format**						**Intervention Design**										**Intera.**			**Timing**						**Concept**				**Implem.**			
Vraga et al. [194]	2021	1348		●					●								●			●								●		●		●	●					●							●
Vraga et al. [199]	2019	406		●					●				●							●										●						●		●							●
Vraga et al. [195]	2020	1,005		●								●	●							●								●	●	●		●				●		●							●
Wahlheim et al. [200]	2020	96	●								●							●		●								●		●						●		●							●
Waltenberger et al. [201]	2023	9			●		●					●	●								●	●								●			●					●				●			
Wang and Huang [204]	2021	271		●								●	●							●										●						●		●							●
Wang [202]	2022	N1=601 N2=1,060		●		●			●			●	●	●					●	●										●			●					●							●
Wang et al. [203]	2022	1	●								●		●														●				●				●				●				●		
Westbrook et al. [207]	2023	N1=125 N2=138 N3=251		●							●			●						●										●						●		●							●
Wijnker et al. [208]	2022	441		●							●							●	●	●	●		●							●			●					●							●
Wood et al. [210]	2023	2,257		●							●			●						●										●		●				●		●							●
Yong et al. [213]	2023	-						●			●				●						●		●							●			●					●							●
Zade et al. [214]	2023	21					●			●			●								●						●			●			●					●							●
Zhang et al. [216]	2022	-						●				●	●						●	●		●					●				●		●					●							●
Zhao [217]	2019	252		●					●	●		●	●							●											●						●	●							●
Zheng and Ma [218]	2022	222		●							●							●					●				●				●		●					●							●

Table 3. Overview of effects and perceptions of reviewed misinformation interventions (conceptual studies without evaluation were excluded).

Source	**Beneficial Effects**	**Beneficial Perceptions**	**Not effective / counterproductive**
Agley et al. [2]	Exposure to infographics with scientific information slightly increases trust in science compared to exposure to control infographic.		Exposure to infographics with scientific information does not have direct or indirect effects on COVID-19 preventive behaviors.
Aird et al. [3]	Exposure to fact-checks corrects beliefs and affects voters' support when corrections outnumber affirmations compared to other correction ratios and for both sides of the political spectrum (η^2=0.13 (fact checks); η^2=0.01 (myth:fact ratio).		
Almaliki [4]		Users perceive interventions with gamification elements useful but preferences for elements vary.	
Amin et al. [6]	Interventions with Visual Selective Attention System can increase attentive behavior of COVID-19 misinformation sharing compared to pre-intervention (D-Scores similar to Cohen's d: Highest number of participants in category 'Neutral/ No Preference' (D-score=-0.15 to D-score=0.15)		
Andi and Akesson [7]	Social norm-based nudge decreases misinformation sharing behavior compared to non-application.		
Ardevol-Abreu et al. [8]			Warning labels to assess credibility are not regarded as central assessment measures by users.
Aslett et al. [9]			Providing dynamic, in-feed source reliability labels do not significantly improve news diet quality or reduce misperceptions (<0.08 change in SD of the pre-treatment measure).
Autry and Duarte [11]			Negated corrections and replacements lead to increased belief in misinformation for cases with no previous exposure to the target concept, relative to cases with exposure and cases with no treatment (η^2=0.22 (main effect of exposure); η^2=0.23(main effect of correction); η^2=0.18 (interaction between exposure and correction)).
Axelsson et al. [12]	Observational learning and feedback as intervention tools increase user performance of credibility assessment compared to the non-treatment control group (η^2=0.043)		
Ayoub et al. [13]	Additional employment of SHapley Additive exPlanations (SHAP) in NLP misinformation detection model and SHAP combined with source and evidence information increases user trust in misinformation detection compared to presenting output text only.		
Bachmann and Valenzuela [14]	Fact-checks are similarly effective at reducing people's misperceptions across message formats (transparency elements, arousing visuals) (d=0.51 (Study 1) and d=0.38 (Study 2))		Compared to control groups without intervention, users exposed to political fact-checks trust news less and perceive the media as more biased, especially after reading corrections debunking pro-attitudinal misinformation.
Barman and Colan [16]	Warning flags with and without explanation text from fact-checking websites reduce perceived accuracy of misinformation and intent to share. Explanatory texts could enhance the trustworthiness of the intervention.		
Bhuiyan et al. [18]	Credibility nudges as browser extension improve user's skills to distinguish news tweets' credibility compared to control group (d=0.296)		
Bhuiyan et al. [19]		Transparency cues (source and message credibility) on news websites increase consumer trust.	

Continued from previous page

Source	Beneficial Effects	Beneficial Perceptions	Not effective / counterproductive
Bhuiyan et al. [20]	Attention and reflection nudges enhance users' credibility assessment compared to control group	Attention and reflection nudges enhance users' credibility assessment (reread and rethink news; use external sources; actively participate in assessment) compared to control group.	
Bode and Vraga [22]	Exposure to corrective information decreases user misperceptions compared to pre-treatment and to the control group (η^2=0.052)		
Bode and Vraga [23]	Interventions with algorithmic or social corrections are equally effective in health misinformation corrections compared to control conditions without intervention for high and low conspiracy belief individuals (η^2=0.046 (interventions overall); η^2=0.016 (comparison between algorithmic and social correction))		
Bozarth et al. [24]		Almost half of participants (moderators on Reddit) preferred cues over labels from expert fact-checkers as they can help discern user intent. A quarter distrusts professional fact-checkers.	
Brashier et al. [25]	Debunking measures have a stronger long-term impact on users' fact-checking memory than prebunking, labeling, or no measures.		
Buczel et al. [28]	Warning before misinformation reduces reliance on it in short-term in comparison to no warning. Warning after misinformation had no effect (η^2=0.05 (forwarning vs. retraction only))		Reliance on misinformation increased for over 7 days although the memory of retraction continued.
Capraro and Celadin [30]	Accuracy endorsement prompt nudge reduces fake news sharing but also increases sharing of real news compared to simple fake alert and no-nudge (f=0.129 (two nudges); f=0.125 (two nudges, different UI); f=0.129 (comparison between endorsing accuracy condition and accuracy salience condition))		
Caramancion [31]			Preventive infographics have trivial to no effect on social media users
Chiang et al. [37]	AI news source credibility system positively affects users' information assessment and attitude towards media literacy learning.		
Challenger et al. [32]	Myth-busting formats, question-answer formats and fact-myth formats are more effective interventions than fact-only formats and control baseline in reducing COVID-19 misinformation agreement ratings.		
Chen and Tang [35]	Intervention with narrative fear appeal messages are effective in promoting health experts to correct online health misinformation for the public.		
Chen et al. [34]	Correct assessment of misinformation overall improved by VisualBubble. Participants became more willing to make assessments and more critical (effect sizes: Topic Filter: large (d=0.98 and 0.98); Opinion Filter: negligible (d=0.00) and medium (d=0.79); Source Filter: large (d=1.11 and 1.01))		Showed tendency to become over-skeptical
Clayton et al. [38]	Intervention with a general warning about misleading articles reduce the perceived accuracy of false headlines relative to a no-warning condition and 'rated false' tag is more effective than 'disputed' tag.(d=0.08 (general warning before seeing headlines); d=0.26 ('disputed' tag); d=0.38 ('rated false' tag))		
Craig and Vijaykumar [39]	Corrective infographic improved rating of misinformation as untruthful and reduced reported willingness to share it. Debunking may be short-lived if followed by misinformation. Effect can be maintained in presence of further corrective information (e.g., η^2=0.150, 0.109 and 0.079)		
Dai [41]	Timing of misinformation correction interventions (pre/post exposure) and addition of coherence message (debiasing/no addition) impacts effectiveness (η^2=0.087 (post exposure); η^2=0.047 (debiasing message); η^2=0.163 (time lapse))		

Continued from previous page

Source	Beneficial Effects	Beneficial Perceptions	Not effective / counterproductive
Dai et al. [40]		Most participants indicate that counterfactual explanations can accurately explain why a piece of news is fake and results suggest that the approach generates the most helpful explanations compared to state-of-the-art methods (human evaluation based on survey with young, well-educated participants).	
Danry et al. [42]		Wearable AI system with explainable feedback enhances rationality in evaluating information in comparison to non-explainable AI and control group	
Denner et al. [43]	A single correction and repeated corrections significantly increased organizational trust compared with no correction		Small negative effect of perceived persuasive intent on organizational trust after repeated corrections.
Desai and Reimers [44]			No evidence that corrections explaining the reason the misinformation was presented were more effective than a correction not accompanied by explanation
Dobber et al. [45]	Red and orange traffic light labels placed concurrently with in contrast to prior to the start of a political advertisement significantly affect credibility perception. Direct-to-consumer labels can be effective but it depends on timing and position.		
Domgaard and Park [46]	Interventions with info graphs increase user ability to identify vaccine-related misinformation compared to text-only intervention and no intervention.		
Drolsbach and Pröllochs [47]	Community fact-checked misinformation is less viral and receives fewer retweets than non-misleading posts.		
Duncan [48]	Credibility labels are effective on news validation when ideological perspective of the user match the ideology of the news brand but also in cases where they do not match.		
Ecker et al. [49]	Corrections are generally effective at influencing inferential reasoning but narrative corrections are not more effective than non-narrative		
Ecker et al. [50]	Corrections are more effective when they explicitly repeat the myth compared to corrections that do not repeat the misinformation (η^2=0.04 (memory); η^2=0.27 (inferential reasoning))		
Ecker et al. [52]	Strong corrections and cognitive load interventions, measured in different degrees of interventions or misinformation strength, can reduce (but never fully) the continued influence effect of strong misinformation, but even strong interventions are less effective on weak misinformation. (η^2=0.05 (strength of misinformation); η^2=0.41 (strength of correction); η^2=0.04 (strength of cognitive load on misinformation); η^2=0.07 (strength of cognitive load on correction);)		
Ecker et al. [51]	Misinformation corrections do not lead to familiarity backfire effects but instead lead to corrective effect in both, audiences unfamiliar to a misinformation and audiences familiar to the topic (i.a., η^2=0.024 (false claim inference across all conditions: no-exposure/fact-check with and without cognitive load); η^2=0.004 (fact check condition without cognitive load))		
Feng et al. [55]	Provenance has effect on credibility perception. Helped correct truth judgments towards deceptive media (qualitatively measured)		Over-corrected in some cases and shifted away from truth in some non-deceptive media
Figl et al. [56]	All evaluated flags lead to reduced perceived credibility. The semantic priming effect of different warning symbols (e.g., stop symbol associated with stopping behavior) makes a difference. Stronger warnings may be required on smartphones than on PCs.		

Continued from previous page

Source	Beneficial Effects	Beneficial Perceptions	Not effective / counterproductive
Folkvord et al. [57]	Interventions with source information positively affect the critical news evaluation compared to a control group with no intervention (e.g., η^2=0.05 for vaccination misinformation)		Inclusion of a protective warning message does not significantly affect critical evaluation (e.g., η^2<0.001 for vaccination and health insurance misinformation).
Freeze et al. [59]	General misinformation warnings which also contain invalid instances, in contrast to valid-only instances and control with no intervention, lead to a discarding of authentic information and to increased memory uncertainty.		
Gao et al. [62]			Stance labels on political ideologies intensify readers' selective exposure (tendency to look for agreeable opinions), and lower the perception of extremeness and criticality of misinformation. Credibility labels only have a limited effect on reducing selective exposure and misinformation identification.
Gesser-Edelsburg et al. [63]	Corrections of misinformation from health organizations are more effective for pro-vaccination as well as for vaccination-hesitant individuals when communication addresses full, transparent information and emotional aspects compared to 'common' one-dimensional, partial responses.	Additional qualitative analysis reinforces quantitative findings.	
Grady et al. [64]	Misinformation warnings for political news are effective in short-term to correct beliefs and eliminate partisan bias but in long-term corrected beliefs weaken and biases return.		
Grandhi et al. [65]		Users perceive trustworthiness indicators as useful for reducing uncertainty and for providing guidance on content interaction.	
Guess et al. [67]	Digital media literacy interventions increase user ability to discern between correct information and misinformation compared to control group without intervention (d=0.2 (US-based study); d=0.11 (India-based study))		
Guo et al. [69]	Specific contextual warnings for video-sharing platforms can alert users to be vigilant and are influenced by explicitness and risk level. In terms of accuracy judgment the interstitial warning and specific contextual warning were both considered effective.		
Hameleers [70]	A combination of media literacy- and fact-checking interventions are most effective in lowering perceived accuracy of political misinformation, compared to each intervention separately and control group without intervention.		
Hameleers and van der Meer [72]	General rather than issue-specific warnings about misinformation are more effective for participants with higher level of trust in the media.		The prebunking exposure to different warning interventions did not influence the truth rating of factually accurate information or misinformation. Observed negative spillover effects of prebunking warnings on truth rating of accurate information.
Hameleers et al. [71]	Multimodality (text-plus-visual) impacts credibility of disinformation but also of fact-checking interventions compared to disinformation and intervention with text-only and compared to control without intervention.		Source type (ordinary citizen, news agency) does not influence credibility level
Hartwig et al. [74]	In several instances, participants changed or consolidated their assessment of the information presented with the help of the indicators.	Participants found the indicators useful for practice and as a reminder to be more able to identify disinformation on their own in the future, without app support.	Adolescents tended to blindly in the intervention.
Hartwig et al. [77]	When topical, formal, and rhetorical indicators are presented with tweets, they improve users' perception and evaluation.	Approach is perceived as useful overall within the context of COVID-19 and the Russian war against Ukraine.	

Continued from previous page

Source	Beneficial Effects	Beneficial Perceptions	Not effective / counterproductive
Hartwig et al. [76]		When assessing the comprehensibility and perceived usefulness of features to assess a voice message's credibility, it received a mostly positive feedback especially on features that refer to the content itself.	
Heuer and Glassman [80]	Checklist with source labels is significantly better in influencing participants' performance on correct article ratings for the better.	Checklist that provides source labels was considered most helpful. The interactive checklist is perceived as more helpful than the written checklist.	
Horne et al. [81]	Soft information nudging/trust nudging has potential benefit of moving even extreme or conspiracy news consumers towards higher quality information (based on simulations)		
Huang and Wang [82]	Misinformation belief is impacted by the message format (narrative/non-narrative) and correction mechanism (social/algorithmic correction) (η^2=0.03 and 0.04 (message format);)		
Irving et al. [83]	Correction reduces number of references to misinformation (medium-to-large effect size) and was remembered and recalled (δ=0.64, 95% BCI [0.28, 0.99] (medium-to-large))		
Jahanbakhsh and Karger [85]		It helped them think about the news in a more analytical way or gauge their trust in a source. They liked being interactive with the news content and the ability to call out content they found biased or misleading.	Assessing took extra time and effort. Sometimes they found it hard to assess a piece of content. They want to think for themselves, unassisted by anyone.
Jahanbakhsh et al. [89]	Lightweight nudging interventions (checkboxes, checklists, free-text rationales) which provide accuracy assessment and rationale reduce misinformation sharing (but also sharing overall).		
Jahanbakhsh et al. [87]	Users perceive incorporation of three new user affordances into social media as useful tools to independent, user-friendly misinformation combat.	Qualitative examples reinforce quantitative findings.	
Jahanbakhsh et al. [86]	Personalized AI impacts users' judgment and grows larger over time, but is reduced when users provide reasoning for their assessment (e.g., exp(β)=1.60 for condition whether AI's prediction had a statistically significant effect on user agreeing with AI)		
Jahanbakhsh et al. [88]	Users perceived value in browser extension that allows to change headlines and used it to make various changes. In follow-up study: substantial number of alternative headlines were preferred especially if bias was removed or deceptions were corrected.		
Jahng et al. [90]	Discounting cues ('fake news' labels) in online comments negatively impact users' ability of veracity evaluation and increase need to authenticate information compared to control group without exposure to discounting cues. (i.a., η^2=0.041 (evaluation ability); η^2=0.057 (need to authenticate)))		
Jennings and Stroud [92]	Partisan affiliations impact likeliness to belief in misinformation, particularly about opposing parties (i.a., η^2=0.13 (user partisanship (P) and party-affiliation of misinformation target(M)); η^2=0.01 (P, M and fact-check condition (F))		Overall, independent from partisan affiliation, fact check interventions do not improve information evaluation compared to cases without intervention.
Jeon et al. [93]	Both the quantitative and qualitative results confirmed that HearHere has an impact on mitigating political polarization and broadening one's perspectives on news consumption.		
Jia et al. [95]	Interventions with misinformation labels (algorithm, community, third-party fact-checker, and no label) reduce credibility of misinformation for liberal users independent of post-ideology while only algorithm labels are effective in reducing ideology-consistent misinformation for conservative users (and all label types for opposing-ideology misinformation).		

Continued from previous page

Source	Beneficial Effects	Beneficial Perceptions	Not effective / counterproductive
Karduni et al. [97]		Visual analytic systems are a helpful tool to support the investigation of misinformation on social media and to enhance traditional (media literacy education) strategies.	
Karduni et al. [97]	Corrections of health-related misinformation with additional use of images is more effective in correcting myth belief than without images (η^2=0.117)		Image type (machine-technical image, expert image, diagram) does not influence persuasive effect.
Kim et al. [100]	Source rating mechanisms are effective interventions to correct users beliefs, whereby expert rating and user article rating are more effective than user source rating. Low ratings and no-ratings have a disproportional stronger effect on user skepticism than high ratings on user trust.		
Kim et al. [101]	Controversy score that provides additional information of opinions on topics and encourages further exploration can be a more effective tool to combat myth belief than approaches that seek to correct or standardize news opinions.		
Kim et al. [103]			No single strategy ((1) hiding content, allowing for explanations, and option to toggle view, (2) including an engagement option with the correction that allows for indicator details, (3) Placing agent next to share button that asks for accuracy and reasoning and presents statistics) was superior over the control. Study highlights necessity of transparency and clarity about intervention's logic and concerns about repeated exposure to misinformation and lack of user engagement.
Kim et al. [102]	Humorous interventions increase user attention to relevant corrections of misinformation, but non-humorous interventions outperform humorous interventions via higher credibility ratings. (η^2=0.19 humor)		
Kirchner and Reuter [104]	Warning-based interventions significantly effect perceived news accuracy but explanation-based approaches are most effective.	Warning-based interventions (with additional explanations) are more effective in correcting user beliefs than less transparent methods such as reduced post size and fact-checks in related articles.	
Komendantova et al. [106]		Stakeholders (journalists/fact-checkers, policymakers, citizens) require design tools for mitigating misinformation and prioritise information regarding actors behind misinformation and tracing the life cycle of misinforming posts. The most valued features across groups relate to timing and flow of misinformation.	
Koch et al. [105]	Warning labels reduced perceived credibility and lowered self-reported likelihood to amplify fake news (rather small effect).		Removing social endorsement cues (e.g., engagement counts) did not have an effect. Did not find a positive effect of warning labels on users' likelihood to elaborate on the fake news post.
Kreps and Kriner [107]	Compared to no intervention, 'false' tags only have a small effect on users' accuracy assessments while journalistic fact-checks are more effective in reducing misperceptions as well as sharing (independent of partisanship).		
Lee and Bissell [110]			Repeated exposure of myths within corrective information increased perceived familiarity about misinformation and increased misinformation credibility (partial η^2=.02 (effect of correction types on misinformation familiarity))

Continued from previous page

Source	Beneficial Effects	Beneficial Perceptions	Not effective / counterproductive
Lee [108]	Web add-on corrections generally decrease the belief in misinformation compared to no correction. For those who are motivated to use social media for specifically for social interaction, narrative corrections are most effective, compared to web add-on's and no corrections (η^2=0.025 (narrative correction for social interaction-motivated users))		Amongst users in general, narrative corrections are not more effective than web add-on corrections or no corrections.
Lee and Bissell [109]	Both commenting and AI fact-checking labels were effective at promoting positive attitudes toward vaccination compared to no intervention. Commenting intervention emerged as promising for suburban participants and the AI intervention was pronounced for urban populations (η^2=.03 (for difference in attitudes between three experimental groups))		Neither of the interventions showed salient effects with the rural population.
Liaw et al. [113]		The proposed system utilizes crowd-sourced corrections, such as in-line commentary and corrections which are ranked by the user to enhance comprehension of news.	
Lillie et al. [114]	The narrative corrective directly reduced misinformation belief compared with a didactic corrective and a no-correction control.		
Lim and Perrault [116]	Post engagement was generally dampened by the presence of warning labels.		Participants were more likely to share congruent posts, with or without labels, suggesting the need for other interventions to address political polarization effects.
Lim and Perrault [115]	The intent to comment and share was significantly lower for posts with a generic warning label than unlabeled posts. The knowledge, source, and propagation labels encouraged sharing instead. Partisanship effects were observed across the labels (partial η^2=0.016 for effect of warning labels on sharing intentions and 0.0077 on commenting intention)		
Liu et al. [117]	No differences in effectiveness across fact-checking sources (professional fact-checkers, mainstream news outlets, social media platforms, AI, crowd-sourcing; η^2=0.01) but sources perceived as more credible are more effective		
Lo et al. [118]		Indicates effectiveness of an fake news intervention module that co-works with a news recommendation system and guides users towards verified news.	
Lu et al. [119]			AI label nudges people into aligning their veracity belief in the news with the AI model's prediction regardless of its correctness compared to a control group (Control vs. AI-before: d= 0.17; Control vs. AI-after: d=0.15)
Martel et al. [120]			Hedging corrections or providing increased explanatory depth in corrections of misinformation had no impact on engagement with corrective messages on social media.
Martino et al. [121]		The Prta system raises awareness about the use of propaganda techniques in the news, promoting media literacy and critical thinking.	
Mena [122]	A warning label was effective in reducing the intention of a user to share misinformation on Facebook compared to a user who did not see the warning. (d=0.36)		
Moon et al. [125]	AI and user consensus (vs. human experts) source labels reduced partisan-based motivated reasoning in assessing fact-checking message credibility (η^2=0.0018 for pattern of motivated reasoning varied by fact-checking sources)		
Moravec et al. [126]	System 1 (automatic cognition) and System 2 (deliberate cognition) interventions both were effective and intervention combining both was twice as effective.		

Continued from previous page

Source	Beneficial Effects	Beneficial Perceptions	Not effective / counterproductive
Nekmat [127]	A fact-check alert was effective in reducing likelihood of sharing misinformation compared to non-exposure.		
Ozturk et al. [129]	A textual counter presented to a rumor was effective in decreasing the likelihood of sharing a tweet compared to the rumor only and rumor with information condition.		
Papakyriakopoulos and Goodman [131]	Textual overlap in labels reduces user interactions and stronger rebuttals reduced toxicity in comments.		Label placement did not change propensity of users to share and engage with labeled content but falsity of content did.
Pareek and Goncalves [132]	Credibility disputes raised by one's co-partisans significantly reduced belief in misinformation, irrespective of one's relationship closeness with the peer. A peer's knowledgeability may be more potent than trustworthiness in causing belief change, and trust can sometimes manifest even in the credibility judgement of distant peers, when perceived to have expertise or a fact-checking tendency.		
Park et al. [133]	When opposite fact-checking labels are shown, users who initially disapprove of a claim are less likely to change their views than those who initially approve of the same claim.	User interviews revealed that users are more likely to share claims with a Divided Evidence label than those with a Lack of Evidence label.	
Pasquetto et al. [134]	Audio files on WhatsApp were found to be more effective than text or video-based sources in correcting beliefs about misinformation and they were shared more frequently when communicated by someone close to the user.		
Pennycook et al. [135]	Warnings were effective in a modest reduction in perceived accuracy of false headlines, particularly for politically concordant headlines, relative to a control condition.		The presence of warnings caused untagged headlines to be seen as more accurate than in the control, even if they were false.
Pennycook et al. [137]	Simple accuracy reminders before sharing information on social media are effective in increasing truth discernment in participants' sharing intentions compared to a control group (d=0.142)		
Pennycook et al. [136]	Shifting the attention of the users on the accuracy of information can encourage them to share higher quality news (e.g., Pearson's r=0.71/0.67/0.61)		
Pillai and Fazio [139]	Participants were less likely to share false headlines in the explain prompt condition compared to control group (exceeded the necessary number of participants according to a priori power analysis; η^2=.03)		
Pluviano et al. [140]			Displaying a myth about vaccines causing autism alongside a factual correction resulted in an increase in belief in the myth over a 7 day time period (partial η^2=0.175)
Porter et al. [141]	Corrections eliminate effects of misinformation on beliefs about vaccine. Effect is robust to formatting changes in the presentation of corrections. Corrections without any formatting modifications are effective at reducing false beliefs with formatting variations playing a very minor role (fact-checks increase accuracy by 0.41 scale points on a four-point scale regardless of formatting; modifications to formatting increase accuracy only by 0.03 points.)		
Porter and Wood [142]	Fact-checks are effective in increasing factual accuracy on realistic simulations of social media platforms (Study 1 Correction Effect d=0.55; Study 2 d=0.79)		
Pretus et al. [145]	Adding a misleading count next to the like count reduced participants' reported likelihood to share inaccurate information by 25% compared to control condition. It was five times more effective as an accuracy nudge (misleading count compared to no intervention: d=0.20; misleading count compared to accuracy nudge: d=0.13).		

Continued from previous page

Source	Beneficial Effects	Beneficial Perceptions	Not effective / counterproductive
Prike and Ecker [147]	The social-norm intervention reduced belief in false claims and improved discrimination between true and false claims. It also had some positive impact on social media engagement. Credibility badges led to greater belief in true claims, lower belief in false claims, and improved discrimination. The credibility-badge intervention also had robust positive impacts on social media engagement, leading to increased flagging and decreased liking and sharing of false posts. Credibility badges and social norms can be effective interventions for counteracting online misinformation. Credibility badges were associated with larger effect sizes and more consistent results across the alternative analysis specifications (partial η^2=0.09 (credibility badge) and 0.01 for social norm).		
Qian et al. [148]	Active interventions significantly increased intention of using reverse image search tools compared to passive interventions and a control group.		Neither passive nor active interventions had an effect on credibility judgment or misinformation discernment.
Rich and Zaragoza [150]			When correcting misinformation, there was no evidence that the time of correction mattered for the efficacy of the correction and the participants corrected beliefs were not durable (durability of corrected belief η^2=0.43; time of correction η^2=0.02)
Ruffin et al. [153]	Simply highlighting and explaining manipulation in photos was not always effective but when it was, it did help make users less agreeing with intended messages (e.g., β=-0.58 of linear regression model for explaining the manipulation versus seeing the original image).		Intervention was not always effective. Explanation had negative effect on feeling/sentiment toward the subject/image
Sakhnini and Chattopadhyay [155]		Fact-checking apps should be sensitive to age-related, personal, and political biases	
Saltz et al. [157]		Findings suggest strong emotional reactions to misinformation labels in general, which are perceived as overly paternalistic, biased, and punitive.	
Sangalang et al. [159]	Narrative correctives (with or without emotional ending) can effectively reduce misinformation beliefs, while emotional corrective endings are better at correcting attitudes.		
Schaewitz and Kramer [160]	Detailed corrections presented alongside disinformation are more effective in better remembering facts compared to simple corrections (η^2=0.02)		The influence of detailed corrections on personal beliefs regarding the topic of the disinformation is counterproductive as more details in the correction seem to raise readers' concerns when corrections are presented together with the disinformation.
Scharrer et al. [161]	Warnings on top of a scientific message made laypeople hesitant about uncritically and confidently accepting the message as true. Participants agreed less with the claims and deemed the text to be less credible than without the warning (η^2=0.48)		Warnings cannot reduce or prevent boost in persuasiveness of easily understandable misinformation.
Schmid et al. [163]		A web app based on Social Network Analysis could effectively provide an overview of potentially misleading vs. non-misleading content on Twitter, which can be explored by users and enable foundational learning.	
Schmid and Betsch [162]	Text-based refutations effectively reduced belief in misinformation and immunized participants against impact in short-time (final power of 94.5% was reached to detect a small effect size. Credibility judgment after 2 months was slightly lower (d=0.04))	Unintended effects: lacking effect on intentions, backfire-effects among religious groups, biased judgments when omitting information about vaccine side effects	
Seo et al. [165]	Machine-Learning-Graph warning, indicating Source Reliability, Content Truthfulness and Picture/Video Truthfulness, was effective increased participants' sensitivity in differentiating fake from real news. (η^2=0.018)		
Sharevski and Zeidieh [168]			Warning labels as visual frictions are not accessible for low vision or blind users.

Continued from previous page

Source	Beneficial Effects	Beneficial Perceptions	Not effective / counterproductive
Sharevski et al. [166]		SPAM warning tags are promising and increase trust in soft moderation. Text-only variant tells participants more of what is going on and a text-and-flag variant gives more specifics and is tougher to refute as a large visual cue. Warning tag with improbable interpretation of facts (FFS) gave convincing options for users to pick why the context is fitting to the misinformation tweet. Left- and right-leaning participants positively rated the intervention.	
Sharevski and Gover [167]			The utterance of a warning cover before a Tweet containing valid information about COVID-19 vaccines by Alexa will not reduce the perceived accuracy of the spoken back Tweet's content relative to a no warning cover condition (d=0.018)
Sherman et al. [170]	A combination of expert and user insights is effective in defining interpretable warnings and design guidelines for communicating the provenance of video content to end-users.		results raise concerns around the potential for users to overgeneralize misinformation warnings regarding video or text information
Smith and Seitz [174]	Corrective mock Facebook news feeds were effective in reducing belief in neuroscience myths when shown immediately after the misinformation for those who held incorrect beliefs at pretest.		If participants held correct beliefs at pretest, a single exposure to misinformation (even when immediately corrected) was enough to have a negative impact on their beliefs.
Song et al. [175]	Image-only modality triggered significantly lower levels of message elaboration and heightened message credibility and increased engagement intentions (effect of evidence type on self-reported message elaboration: η^2=0.01. Effect of presentation mode on message elaboration: η^2=0.02)	Presence of statistical evidence in assertions reduced message elaboration and effects of message in correcting misperceptions, decreased perceived message credibility and lowered intentions to further engange with and disseminate the corrective message.	
Sullivan [177]			Libraries were not effective in correcting misconceptions about the flu vaccine through comments on social media.
Tanaka and Hirayama [178]	Objective countermessages reduced belief in rumors and subjective countermessages strengthened false beliefs (e.g., η^2=0.02. Post-hoc power analysis revealed adequate G*Power >0.80 at medium to large effect size levels).		Subjective countermessages even strengthened false beliefs
Tanaka et al. [179]	Displaying criticism of false information prior to rumors during a disaster response is effective in increasing proportion of responses aimed at stopping the spread of rumors compared to displaying the criticism after the rumor.		
Tao et al. [180]	All three types of corrections improved belief accuracy. Corrections incorporating hope appeals showed enhanced effectiveness when threat information was present in comparison to absent hope appeals (Power analysis reveals study can detect small effect sizes (f=0.11) with power of 80%. Hope appeal when threat was present versus absent: η^2=0.01)		
Thornhill et al. [182]	BalancedView, a proof-of-concept that shows news stories relevant to a tweet suggests that nudging users by providing context information may change the behavior of them towards that of informed news readers.		
Tseng et al. [183]	Corrective information in the form of text, images, or videos is effective in reducing participants' perceived credibility and potential action for misinformation, with videos being particularly effective in correcting text-based misinformation.		
Tsipursky et al. [184]	The Pro-Truth Pledge (PTP) has been shown to effectively reduce the sharing of misinformation and encourage truthful behavior on social media (d=-1.93).		

Continued from previous page

Source	Beneficial Effects	Beneficial Perceptions	Not effective / counterproductive
Tulin et al. [185]	Truth sandwich in fact-check had indirect benefits such as more positively perceived intentions of fact-checkers and less reactance to reading subsequent fact-checks compared to classic fact check that repeats false claim (small effect sizes, e.g., for classic fact-check: β=-.13		Truth sandwich was not effective in correcting false beliefs but had indirect benefits.
Tully et al. [186]	Users tend to provide accurate information in corrections, particularly after viewing other corrections. However, users are generally unlikely to respond to tweets containing misinformation (exposure to corrections: η^2=0.001; tone of corrections: η^2=0.010).		neither exposure to corrections nor tone of corrections increases the self-reported likelihood of responding to the misinformation tweet as compared to the misinformation-only condition
Tully et al. [187]	News literacy messages alter misinformation perceptions, however not with a single message (e.g., partial η^2=.0.006 for NL tweet leading participants to rate stories overall as less credible than texting tweet)		
van der Meer et al. [188]			Warnings can prime general distrust in authentic news.
van der Meer and Jin [189]	Corrective information is effective in debunking misinformation, and factual elaboration compared to simple rebuttal stimulates intentions to take protective action, with government agencies and news media being more effective in improving belief accuracy compared to social peers.		
Velasco et al. [190]		The browser extension that allows to insert text and creates a (binary) feedback based on logistic regression was rated highly acceptable in terms of functionality, reliability, usability, efficiency, maintainability, and portability.	
Velez et al. [191]	Fact-checks undo effects of misinformation on beliefs (large and significant effect of over .26 scale points change). No Backfire effect was observed.		
von der Weth et al. [192]	Nudging users toward more conscious posting and sharing behavior by using linguistic analysis to infer the factuality of content and the credibility of sources is effective in reducing the reach and speed of spread of misinformation.		
Vraga et al. [196]	User corrections of a meme containing misinformation are effective in reducing the credibility assessment of the misinformation post (η^2=0.077) and misperceptions (η^2=0.088)		Exposure to news literacy messages did not enhance the effectiveness of corrective responses or boost NL attitudes and may have generated cynicism.
Vraga and Bode [193]	Social corrections providing a source are effective compared to not giving a source (partial η^2=0.035)		
Vraga and Bode [197]	Misinformation correction by expert group is effective without loosing the groups credibility and trustworthiness in the context of a health topic (misinformation correction: partial η^2=0.009; trustworthiness: partial η^2=0.001; credibility: partial η^2=0.004)		misinformation corrections of a single user is not effective
Vraga et al. [198]	Expert organizations can be effective in successfully correcting misinformation on social media on two controversial health topics		
Vraga et al. [194]	User corrections in real-time partially reduce the effect of misinformation videos on beliefs (partial η^2=0.03 compared to no intervention) but not on intentions.		
Vraga et al. [199]	Logic-based and humor-based rhetorical corrections reduce misperceptions only for some topics (partial η^2=0.013).		
Vraga et al. [195]	Logic-focused (before and after misinformation) and fact-focused (after the misinformation) corrections reduce misperceptions, with logic-focused corrections appearing to reduce the credibility of misinformation and fact-focused corrections being more credible.		
Wahlheim et al. [200]	Reminders of misinformation are effective to diminish the negative effects of fake-news exposure short-term (d=0.29)		
Waltenberger et al. [201]	Contextualizing user profiles with data from previous contributions helped users contextualize posts, identify political tendencies, distinguish humor from problematic mindsets (qualitatively measured)		

Continued from previous page

Source	**Beneficial Effects**	**Beneficial Perceptions**	**Not effective / counterproductive**
Wang and Huang [204]	One sided narrative messages are more effective then two-sided ones for correcting misinformation on e-cigarettes		Effect disappeared if participants had smoked e-cigarettes before
Wang [202]	Participants accept unwelcome fact-checks on Facebook but welcome fact-checks on Line (private messaging app). Fact-checks help increase media literacy in open platforms and hamper media literacy in private messaging apps.		
Westbrook et al. [207]	External correction (news source labeling misinformation as false) influences perceptions of misinformation source. Perceptions of the misinformation source can cause changes in belief in misinformation. (a priori power analysis allowed for desired power of 0.8)		
Wijnker et al. [208]	All investigated correction methods for misleading graphs were effective for debunking misinformation directly after correction and reduces over time. Showing an accurate alternative graph was more effective than visual cues or text-based warning cues to activate graph literacy or warning messages for possible deceit.		
Wood et al. [210]	Debunking messages of healthcare professional lead to increase in beliefs about risks of vaccines in the UK but not the US. Messages from political authorities and discrediting messages had no effect. There is a joint importance of message source and messaging strategy regarding effectiveness of debunking (e.g., debunking by health experts reduced belief that vaccines cause severe side effects by 0.19 points on Likert scale)		
Zade et al. [214]		Tweet trajectory (e.g., unfamiliar activity invokes skepticism in following network) and contextual cues (e.g., profile description helps infer purpose of account) helped support users in assessing credibility (qualitatively evaluated).	
Zhao [217]	Participants exhibit a more positive attitude towards corrective messages and have higher vaccination certainty when such messages are present across multiple social media platforms, as opposed to only one platform.		
Zhang et al. [216]	Concise corrections are more effective than exhaustive ones. Graphical explanation has small positive effect (e.g., Spearman's ρ=0.126).		Textual explanations for why misinformation is wrong do not significantly affect effectiveness. Warnings in a tough tone make corrections worse. Textual and graphic warnings have negative associations with correction effectiveness.
Zheng and Ma [218]	Explanatory annotations and interactive linking in misinformation combining text and visualizations can significantly lower perceived credibility (e.g., d=-0.367). The effect to raise awareness is limited/marginal while linking was more effective than annotation (e.g., d=-0.367)		