Reviewing Interventions to Address Misinformation: The Need to Expand Our Vision Beyond an Individualistic Focus

ZHILA AGHAJARI, Lehigh University, USA
ERIC P. S. BAUMER, Lehigh University, USA
DOMINIC DIFRANZO, Lehigh University, USA

Prior work has identified a variety of factors that drive the way people identify and respond to misinformation. Such factors include confirmation bias, perceived credibility of the information source, individual media literacy, social norms, and others. This paper reviews the interventions designed to address misinformation and examines how various underlying mechanisms of response to misinformation are operationalized and implemented in the reviewed interventions. Key findings show that most prior work to address misinformation heavily focuses on individual pieces of misinformation and the actions individuals take in response to those individual pieces. These individualistic approaches, we argue, overlook the other drivers of responses to misinformation, such as individuals' prior beliefs and the social contexts in which misinformation is encountered. Additionally, the analysis shows that an individualistic focus on misinformation draws attention away from the systemic nature and consequences of misinformation. This paper argues that to overcome the limitation of individualistic approaches to addressing misinformation, future interventions need to expand their scope beyond individualistic approaches. As one way to do so, it discusses leveraging the impacts of community factors that impact the spread and impacts of misinformation. The paper concludes by using social norms as an example to illustrate how a focus on community factors might work in practice.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI; Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Misinformation, Misleading Content, Drivers of Misinformation, Social Media, Social Norms, Online Communities

ACM Reference Format:
Zhila Aghajari, Eric P. S. Baumer, and Dominic DiFranzo. 2023. Reviewing Interventions to Address Misinformation: The Need to Expand Our Vision Beyond an Individualistic Focus. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 87 (April 2023), 34 pages. https://doi.org/10.1145/3579520

1 Introduction

The rise of online misinformation is a growing problem that can impact political, economic, and social interactions across the globe. Informed by prior work [150, 224, 247], this paper considers any content that has the potential to mislead people about the truth, either inadvertently or purposefully, as misinformation. Misinformation has sweeping impacts, from its effects on the US 2016 and 2020 presidential elections [18, 48, 64, 193, 232], to its impacts on India's economy during 2019 coronavirus disease outbreaks [9], to its impacts on attitudes toward COVID-19

Authors' addresses: Zhila Aghajari, zha219@lehigh.edu, Lehigh University, 322 Building C, 113 Research Drive, Bethlehem, PA, USA; Eric P. S. Baumer, ericpsb@lehigh.edu, Lehigh University, 235 Building C, 113 Research Drive, Bethlehem, PA, USA; Dominic DiFranzo, djd219@lehigh.edu, Lehigh University, 328 Building C, 113 Research Drive, Bethlehem, PA, USA.

2573-0142/2023/4-ART87 $15.00
https://doi.org/10.1145/3579520

vaccines [82, 158, 201, 235]). In several cases, online misinformation even inspired different incidents of violent attacks in physical interactions [e.g., 4, 203].

This critical societal challenge has particular relevance for the CSCW community and the broader SIGCHI community [14, 19, 100, 127, 165]. Many of the technologies that we as a community have developed and studied—personalized recommendation systems, distributed social information sharing, automatic content curation, etc.—are the very technologies that exacerbate the rapid spread and detrimental effects of misinformation [56, 111, 141, 250, 255]. Perhaps, then, these same technologies can be strategically used and/or redesigned with interventions to combat misinformation [e.g., 2, 8, 13, 37, 140, 208, 236]. Some of these interventions are informed by significant psychological and social science work that explains the factors that drive the spread of misinformation. Such factors include confirmation bias [179, 218], perceived credibility of the information source [77, 114, 153], individual media literacy [27, 29, 53, 129], social norms [24, 70, 144, 207], among others.

To identify strategies for interventions that maybe under explored, this paper examines the relationship between the driving factors of misinformation and prior work on countermeasures to address misinformation. To do so, it conducts a two-stage review process, following standard practices in literature review methodology [181]. First, this paper reviews interventions designed to address misinformation. Second, after examining the citations within those papers, our review describes the underlying mechanisms that drive responses to misinformation. We analyze these two sets of papers to examine how the various underlying mechanisms are operationalized and implemented in the reviewed interventions. Details of this literature review process and the analysis are discussed in §2.

The analysis finds that most prior work maintains a primarily individualistic focus, aiming to help individual users identify individual pieces of misinformation (i.e., fact-checkers, signaling credibility of the content, signaling credibility of the source of content, and providing more perspectives on the content) [e.g., 2, 8, 13, 37]. We also found some approaches that focus on addressing how individual users respond to misinformation (e.g., shifting the individuals' attention toward the accuracy of content, blocking users who engage in suspicious behaviors) [e.g., 2, 3, 7, 85, 97, 124, 189, 190]. Finally, despite the significant role of online communities on people's response to misinformation [e.g., 24, 36, 70, 99, 109, 112, 144, 233, 240], our review finds that relatively little work has investigated interventions designed around community factors to address misinformation [e.g., 23, 178].

The analysis further identifies two key issues with this individualistic approach to addressing misinformation. First, such interventions rely on **implicit assumptions** that in practice are unlikely to hold. For example, they tend to assume that individuals are rational actors who are likely to respond to corrective information. However, much of an individual's decision-making stems from sources other than pure rationality, such as self-perceived rationality of the content and its source [145], community narratives [219], and motivated reasoning [131, 147]. Second, individualistic approaches result in **a number of blind spots**. For example, they make it difficult to account for statements that are factually true but misleading [47, 204]. An individualistic focus also draws attention away from the systemic nature and consequences of misinformation.

This paper argues that to overcome the limitations of individualistic approaches to addressing misinformation, future interventions need to go beyond an individualistic approach. One way to do so is by focusing on the role of community factors in addressing the spread of misinformation and its impacts. While a wealth of different group and community level processes could be relevant, this paper illustrates a focus on community factors by describing the role of social norms [148, 207]. Social norms influence people's behavior in a variety of domains [20, 163, 237, 237], including responses to misinformation [70, 101, 144]. In addition, social norms impact the social contexts in which misinformation is encountered. This way, social norms may impact the ways in which misinformation contributes to the broader issues at the community and societal level. While not the only community-oriented factor that contributes to societal issues of misinformation, the interplay between social norms and people's response to misinformation can amplify these issues.

To reiterate, social norms offer just one example. The primary contribution of this paper is to highlight how prior work on mechanisms to address misinformation embodies an individualistic focus, to point out the issues associated with that individualistic focus, and to suggest that future work should explore how community-oriented factors could inform the design of misinformation interventions.

2 Methods

To identify under-explored strategies for misinformation interventions, we reviewed existing interventions, as well as the various drivers of misinformation those interventions leverage. Of the different types of literature reviews, this goal aligns most closely with what is often called a mapping review or systematic map [102]. However, a mapping review usually characterizes studies based on their design and rarely includes any quality assessment [102]. In contrast, our review sought not only to assess the quality of work in this area but also to identify under-explored theories and strategies.

To conduct a systematic literature review, we sought guidelines from prior work [e.g., 102, 181, 185]. For instance, PRISMA guidelines [185] have been used for a variety of reviews [e.g., 30, 51, 55, 105]. However, "The PRISMA 2020 statement has been designed primarily for systematic reviews of studies that evaluate the effects of health intervention"[185, p. 2]. In addition, PRISMA guidelines focus primarily on quantitative reviews that use "pairwise meta-analysis or other statistical synthesis methods"[185, p. 2]. Therefore, a PRISMA approach was poorly suited for the purpose of our review.

Instead, our process draws primarily on the guidelines for a systematic literature review Okoli [181]. According to Okoli, the purposes of a systematic literature review include to "identify gaps in current research, and [to] provide a framework for positioning research endeavors" in the future [181, p. 882]. Such purposes fit the goal of our study more than any other review processes [102]. Thus, this section explains how we adapted and applied the review process from Okoli [181].

2.1 Identify the purpose of this literature review

Following Okoli [181], we first clarified the purpose of our literature review. Specifically, we focused on interventions for addressing misinformation, which we define as *false or misleading information that has the potential to deceive people*, regardless of the intentions of the actor who spreads it. This definition draws on and synthesizes aspects from a variety of prior definitions. Lazer et al. [150, p. 1094] define misinformation as "false or misleading information", and disinformation refers to "false information that is *purposefully* spread to deceive people" (emphasis added). Similarly, Wardle et al. [247] refer to misinformation as false or misleading information that the person spreading it believes is true, and defines disinformation as false information that the person disseminating it knows is false. Both Lazer et al. [150] and Wardle et al. [247] emphasize the deliberateness of disseminating false information to distinguish disinformation from misinformation.

In contrast, Starbird et al. [224] argue that disinformation is a form of information operation, and they emphasize that disinformation aims to undermine the integrity of the information space, and overwhelm individuals to make sense of information. According to Starbird et al. [224], therefore, disinformation is broader than a single piece of information, and is rather a collaborative work of

Rather than make a clear distinction between misinformation and disinformation, our definition of misinformation combines elements from each of these prior definitions. Noting the difficulty in determining intent behind content, our definition makes no strong claim about the poster's intent to deceive (or lack thereof) [cf. 247]. This approach enables us to include both interventions that do and that do not attempt to infer the poster's intent, making this review more inclusive. At the same time, our definition draws on the insight from Starbird et al. [224] that the phenomena surrounding misinformation and its effects often transcends any single piece of content. This stance enables our review to include both interventions that attempt to address larger, coordinated campaigns [e.g., 3, 7, 85], and interventions that are geared toward individual actors or pieces of content[e.g., 37, 124]. While this approach reviews interventions that are designed to address spread of misinformation with different intentions under the same category, it helps analyze the broader landscape of interventions around misinformation.

2.2 Our Review Process

Following Okoli [181], we used the above scope to formulate a set of steps to conduct our review. This subsection briefly overviews our review process, which is then described in greater detail in the remainder of this Methods section.

Based on the goal of mapping existing interventions and identifying under explored approaches, we started by collecting intervention papers (§2.2.1). This corpus of interventions was used to identify various factors and mechanisms that act as drivers of misinformation (§2.2.2). These individual mechanisms and factors were inductively analyzed and thematically categorized (§2.2.3). These categories of drivers provided a means of analyzing the corpus of interventions to address misinformation, determining which drivers are more or less often used to inform the design of these interventions(§2.2.4).

Figure 1 summarizes this review process. It includes the number of papers identified or retained at each step, as well as the relationships among the different steps. The remainder of this Methods section describes each of these four main steps in greater detail.

2.2.1 Build the Intervention Corpus This section describes the steps we took based on Okoli [181]'s guideline to build the intervention corpus, and overviews the intervention corpus obtained from this step.

2.2.1.1 Initial Search. For our initial literature search, we used a search of the ACM Digital Library to collect papers. We searched papers published in PACMHCI and in conferences and journals where SIGCHI is a sponsor or co-sponsor. This library was chosen as our initial database because one of the key elements of our literature scope is investigating how the CSCW community and the broader SIGCHI community approach misinformation and design interventions to address this phenomenon. This initial literature search helps to ensure a comprehensive coverage of the relevant literature in SIGCHI community.

We used the search term "misinformation" to search for any papers that included the word "misinformation" anywhere in the paper. We used a single search term (i.e., misinformation) since we were specifically interested in the phenomena of misinformation (as apposed to "disinformation" and "fake news," both of which prior work distinguishes from "misinformation" [16, 150, 224, 247]). Other prior reviews have similarly conducted their search processes around very constrained search terms [e.g., 35, 46, 52, 79, 139] as this approach helps to focus exactly on the phenomenon of interest and to exclude other related but dissimilar phenomena.

2.2.1.2 Application of Inclusion Criteria. In the resulting papers, we looked at the title and the abstract of each paper to identify the papers that met the following inclusion criteria:

- The paper was a peer-reviewed published work.

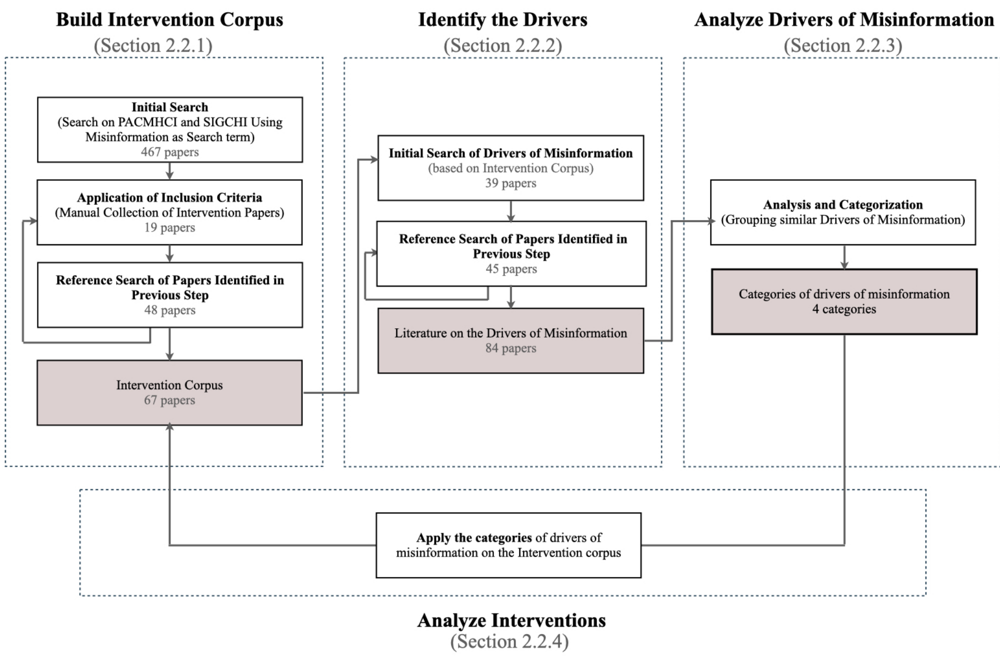

Fig. 1. An overview of our review process. The number of papers examined and retained by search iteration is shown in the boxes.

- The paper presented an intervention to address the spread or impacts of misinformation, or the paper investigated the efficacy of already existing interventions to combat misinformation.
- The intervention in the paper focused primarily on influencing the *response* to misinformation (e.g., when reading the news) rather than influencing the *creation* of misinformation (e.g., by a news reporter).

If the title and abstract were not clear enough about whether or not the paper focused on an intervention, we read the paper in more detail. To do so, we looked at the introduction and methods sections of the paper.

Papers that did not match these criteria were excluded. Specifically, while we included approaches that aimed at helping online users identify misinformation, we excluded the various methods that were aiming at improving classifiers to detect misinformation. In addition, extended abstract and working papers were excluded from our corpus, partly because it would have been resource-intensive to analyze that many papers, and partly because they did not always contain enough information for our analysis. Given the goal of this work (i.e., identifying under explored interventions), our review was not bound within any particular years to ensure that all the explored interventions in this community are covered in our review.

2.2.1.3 Reference search of papers identified in previous step. Following the guidelines from Okoli [181], for each paper we went through the reference list to conduct a second pass of search. That is, we included in the review the intervention approaches that were cited by these resulting papers, but were not included in PACMHCI or SIGCHI proceedings. References that seemed relevant (i.e., the papers which include an intervention to address misinformation or investigated the efficacy of such interventions), were selected based on the title and the abstract of the papers.

Next, we repeated the process of going through the references and seeking the relevant papers until we reached a stage of saturation (i.e., where there were no more new publications being added to the corpus). Each interaction resulted in fewer relevant manuscripts and more cross-reference within the already found publications. Prior research noted such a systematic approach of paper seeking to be very effective at constructing a corpus of publications that are related to the same theme [142].

2.2.1.4 Overview of Intervention Corpus. The first step of literature search on ACM digital library (i.e., §2.2.1) returned a total of 467 manuscripts that include the word "misinformation" anywhere in the paper. After excluding papers that did not meet our aforementioned criteria, this step identified 19 papers around interventions to address misinformation.

According to step two and third in our approach, we repeated the process of going through the references and seeking the relevant papers until we reached a stage of saturation. This process resulted in 48 papers. Therefore, the whole process yielded 67 intervention papers being included in our review of intervention papers.

2.2.2 Identify the Drivers of Misinformation.

2.2.2.1 Initial search of drivers of misinformation. Next, we sought to identify the factors that drive misinformation. Most of the papers in our intervention corpus (generated in §2.2.1) include a subsection, usually within related work, where factors that drive misinformation are discussed. We used the citations that were introduced in those reviews to begin building our review of drivers of misinformation. We collected all the factors and mechanisms that were described as influencing the way people identify misinformation and respond to it. These drivers of misinformation were described based on prior work in journal such as PNAS (Proceedings of the National Academy of Sciences of the United States of America), Science, and Nature, as well as psychology journals, such as Journal of Experimental Psychology: General, Applied Cognitive Psychology, Psychological Bulletin. This step identified 39 papers around drivers of misinformation.

2.2.2.2 Reference search of papers identified in previous step. Similar to our method in building the intervention corpus, we again applied the literature review methodology offered by Okoli [181]. That is, for each paper identified at each step, we went through its reference list to conduct a second round of searching. This process resulted in 45 papers. Therefore, the whole process resulted in 84 papers on drivers of misinformation being included in our review.

2.2.3 Analyze Drivers of Misinformation. After collecting relevant research for our review, following the guideline by Okoli [181], we started reading the papers and extracting the information required for our analysis. Aiming to map the interventions to the drivers of misinformation, we began our analysis by investigating the papers on the drivers of misinformation (from §2.2.2). The goal was to collect all the drivers of misinformation that were identified in prior work and to recognize patterns that might exist within those factors. Therefore, we conducted a thematic, inductive analysis [50], looking for patterns, similarities, and trends in terms of the ways that these papers conceptualized the various mechanisms, entities, and factors that might drive misinformation.

To do so, the authors read the papers and met periodically throughout the reading and analysis process to discuss observations, and emergent themes to reach consensus [50]. This step resulted in the development of a list of drivers of misinformation that prior work introduced as influential

Table 1. List of Papers in Intervention Corpora Categorized Based on The Driver of Misinformation that the Interventions Focus on

The orientation of Intervention	Number of Articles	References
Content	40 papers	[2, 6, 8, 10, 13, 25, 45, 56, 60, 61, 63, 68, 83, 84, 86, 93, 94, 97, 98, 106, 113, 119, 140, 174, 176, 177, 180, 183, 198, 215, 216, 223, 225, 227, 236, 238, 245, 253, 254, 258]
Source	10 papers	[37, 76, 77, 83, 113, 121, 134, 191, 210, 223]
Individual Users	16 papers	[3, 7, 33, 34, 39, 42, 74, 85, 124, 125, 135, 189, 190, 208, 241, 243]
Community	4 papers	[23, 37, 138, 178]

on people's response to misinformation. The authors then discussed this list of drivers of misinformation and identified a pattern based on the entities that each of these drivers and influential factors can be related to. The results of this step are discussed in §3.

2.2.4 Analyze Interventions. We applied the categories of drivers of misinformation (from §2.2.3) on the intervention corpus (from §2.2.1) to analyze the papers in the intervention papers. Specifically, we aimed to map each intervention paper to one or more of the categories of drivers. To provide this mapping, the authors examined each intervention paper to determine what drivers of misinformation are used as the motivation for, or as the focus of, the design of the intervention. If a paper used more than one driver of misinformation to design an intervention, we allow the intervention to sit in multiple categories. To conduct this analysis, the authors read the papers, and discussed the observations to reach consensus on the category (or the categories) that each paper falls under. The results of this step are discussed in §3.

3 Results

Using the methods described above, we reviewed prior work that investigated factors that influence the way people identify misinformation and respond to it, and identified four categories: content factors [e.g. 120, 124, 229], source factors [e.g. 5, 38, 77], individual users factors [e.g. 49, 65, 164, 179], and community factors [e.g. 24, 70, 144, 207].

Based on the analysis described in §2.2.4, we categorized the interventions designed to address misinformation into four groups: *Content-Based strategies* used to address misinformation; *Source-Based strategies* used to address misinformation; *Individual User-Based strategies* used to address misinformation; *Community-Based strategies* used to address misinformation. Table 1 shows a list of papers in each of these categories.

This section reports the different categories of drivers of misinformation identified by prior work, as well as the interventions that are designed to address the spread of misinformation. With the goal to identify strategies that maybe under explored, this section provides a mapping of the interventions to address misinformation to the drivers of misinformation.

3.1 Factors about the Content

3.1.1 Content-Related Mechanisms The characteristics of an individual piece of content (e.g., a news article) can influence how people interact with and respond to it [120, 124, 223, 229, 246]. Such characteristics include the number of quoted sources [228], prior exposure to a news article [188], supporting evidence in the article [124], whether or not a piece of content appears biased or not [124], and the number of related articles written about the same news event by other news organizations [124, 229]. Put differently, each of these content-related factors can work as a heuristic that guides individuals' response to (mis)information.

Such factors can also be manipulated to influence perceptions about the reliability of content, and information processing. For example, examining the role of framing in individuals' assessment of articles that evaluate truthfulness of news, Kreiner and Gamliel [146] show that framing effect influence the perceived reliability of the evaluation articles when the outcomes were favorable but not the outcomes were unfavorable.

3.1.2 Content-Based Strategies Used to Address Misinformation We find three general approaches that focus on specific pieces of content: disputing False information Using Fact-checkers, signaling credibility of content, and reducing the visibility of misleading content. For each, we describe some of the dominant approaches, and summarize findings from studies of how these countermeasures are used in practice.

3.1.2.1 Disputing False information Using Fact-checkers One of the most common approaches to mitigate the spread of misinformation is disputing false information using fact-checkers [8, 13]. Fact-checkers assess the veracity of different claims made by public figures (e.g., politicians, pundits, corporations, etc.) that are likely to be misleading based on several signals. These signals include observing expression of dis-beliefs to these content, the speed by which these content spread, and the output of machine learning models that predict false information [69, 86, 245].

A variety of entities work to provide accurate fact-checkers, from social media companies such as Facebook and Twitter [8, 13, 119], to journalists [130, 239], to academic institutions [61, 176, 198, 254]. There are different forms of fact-checkers, from websites such as PoliticFact and Snopes, which evaluate factual claims of news, to credible news media such as Washington Post and Wall Street Journal, to content moderation techniques (either paid agents or volunteers) used by platforms such as Facebook and Reddit [84, 94], to automatic fact-checkers [110]. Figure 2a shows an example of disputing a piece of news on Facebook using fact-checkers. Prior studies demonstrate the success of fact-checkers in identifying low credible information and mitigating its spread [60, 68, 253].

A large body of research has investigated the ways online users engage with fact-checkers and show individual responses to fact-checkers vary based on different factors [93, 98, 174, 183, 215, 216, 227, 258]. For example, examining people's response to rumors on Twitter before and after their veracity is determined by fact-checkers, Zubiaga et al. [258] show many people share tweets that support a rumor which is still not verified. However, once the rumor has been debunked, they are less likely to make the same effort to communicate to their followers that the content they previously shared was untrue. In another case, Shin et al. [215] investigate the spread of rumors on Twitter during the 2012 U.S. presidential election and show that people choose the outcome of fact-checkers subjectively. That is, partisans selectively shared fact-checkers' messages that were advantageous to their group and denigrated the opposing group. As a result, rumors had been propagated even after fact-checker organization debunked them. This subjective use of fact-checkers occurs because many individuals tend not to question the credibility of information unless it contradicts their view and prompts them to do so [150]. These findings align with the phenomenon of selective exposure, where individuals prefer to read information that is in line with their prior beliefs [95], and avoid the content that contradicts their beliefs [1].

Mosleh et al. [174] investigate downstream consequences of social corrections on the users' future content sharing behavior. Specifically, [174] investigate the behavior of large scale users on Twitter who shared false information and received replies to their false tweet with links to fact-checkers. Examining of the users' subsequent activities shows that this method decreased the quality of content they shared, and increased language toxicity.

(a) (b)

Fig. 2. a) Facebook uses third-party fact-checkers to mitigate the spread of misinformation [123]. b) Facebook suggest related-articles to provide additional information on articles with low credibility [72].

3.1.2.2 Signaling Credibility of Content Relatively less work has explored different ways to give insights about the credibility of content [2, 106, 140]. Facebook, for instance, displayed red flags on articles that were disputed by fact-checkers to signal their lack of credibility. Warning against misinformation has been shown to reduce its perceived accuracy [63, 140, 153]. However, prior research suggests that strong language or visualization in warnings can backfire and strengthen prior beliefs [97, 153, 180, 213].

In addition, warnings such as red flags are likely to prompt people to click on the false content [153, 183]. Clicking on false information in turn increases the visibility of content, the risk of repeating false information, and may result in the increase in accepting misinformation as true [2, 150]. Lastly, a red flag can only signal the credibility of content that is false, and cannot communicate any information on the veracity of partly false and unproven content.

Therefore, Facebook removed the red flag feature in favor of a new feature, named "Related Articles" [183]. Unlike the fact-checkers that inform whether a content is "true" or "false", this approach aims to provide people with additional information (e.g., reporting from a certified third-party fact-checker or the stories published by another publisher) and help them decide by themselves whether the news is misleading or not. Figure 2b shows an example of how related articles are suggested under low credible or unproven articles. Similarly, Twitter uses labels or warning messages to provide additional information or clarifications about tweets that contain harmful and unconfirmed claims [238].

Kirchner and Reuter [140] examine the efficacy of different warning-based approaches (i.e., a simple warning that shows the article is disputed, related articles underneath the post with headlines contradictory to the false claim, and a warning extended by a short explanation). The results of their experiment show that all these warning messages are effective, but adding an explanation to warning messages is the most effective warning-based approach. Specifically, adding an explanation to a warning is shown to be more effective than the related articles approach. In another case, Bode and Vraga [45] show when people have strong prior beliefs about an issue (e.g., anti-vaccine beliefs), providing related stories are less likely to correct initial misconceptions that misinformation creates. Instead, in this case corrective information might even backfire and result in accepting false information as correct more strongly.

3.1.2.3 Reducing visibility of misleading content Reducing the visibility of harmful content, including misleading information, has been used by several social media companies to mitigate the spread of such content [56, 236]

One approach to lower the visibility of misleading content is to use the crowds to rate trustworthiness of content, and use this information in the ranking algorithms of news feeds [6, 10]. Epstein et al. [83] investigate the efficacy of this approach, and examine whether or not laypeople game this crowdsourcing mechanism to promote content of their own interests. Their findings suggest people are less likely to game the system. Indeed, the participants trusted mainstream sources much more than hyper-partisan or fake news sources, regardless of their partisanship. However, many people distrusted unfamiliar outlets. Epstein et al. [83] argue that while using the crowds trust ratings are effective in discerning between high and low quality content, this approach may still result in a rise in polarization.

In another case, YouTube claims it uses a combination of machine learning methods and human evaluation to identify misleading, harmful videos and videos that include borderline content [236]. While YouTube does not remove these videos, it removes them from recommendation systems as a way to lower their exposure. Reddit uses this YouTube approach to lower the visibility of misleading content. Buntain et al. [56] investigate the efficacy of this approach on Twitter and Reddit, and demonstrate that de-recommendation results in a significant decrease in sharing conspiracy labeled content on both platforms. However, this approach resulted in an increase of sharing conspiracy labeled content on conspiracy oriented communities on Reddit. These findings suggest that reducing exposure to harmful, misleading content might be beneficial for some communities, but harmful for some others (e.g., the communities on the borders).

Another approach to reduce visibility of misleading content, employed by Facebook, is reducing the size of low credible headlines and articles. Kirchner and Reuter [140] conducted semi-structured interviews to investigate the efficacy of this approach (in addition to several other approaches of addressing misinformation). While some of the participants found this method favorable as it draws attention away from misleading content, some others believe that it is rather an extreme approach and almost like censoring content. Overall, participants preferred warning based approaches that provide additional information over reducing the size of misleading content.

3.2 Factors about the Source

3.2.1 Source-Related Mechanisms In addition to factors related to individual pieces of content, factors related to the source of information can influence individuals' assessment of that information [5, 38, 76, 77, 114, 153]. Lewandowsky et al. [153] explain that when people lack the motivation and the knowledge to investigate a message in detail, they tend to defer to their assessment of the source's credibility. If the source is perceived as credible, people are more likely to consider the content as credible [81, 194]. In this case, a lack of detailed information about a source may decrease users' perceptions of the credibility of content from that source [91].

Multiple factors can influence the perceived credibility of a source. These factors include the "look and feel" of a website [89, 90], how information on the site is structured [89], the professionalism of the site design [166], and official-looking logos and the domain names [249].

Kahan [132] argues that many people's assessment of news sources can also be influenced by their partisan bias. In this case, as a result of motivated assessment, unreliable content may be perceived as reliable. Pennycook et al. [190], however, demonstrate that failing to assess misinformation often results from a lack of reasoning about, rather than a from motivated assessment of, news sources. In another study, Epstein et al. [83] show that if people are asked to reflect on trustworthiness of news sources, their judgment may not be disproportionately swayed by their partisanship. Pennycook et al. [190] also argue that if prior experience with an outlet and the content the outlet shares is necessary to shape an accurate assessment of its reliability, many people cannot judge most outlets as there are so many outlets with which people may have not experienced. Future research is therefore required to investigate how people assess the legitimacy of news outlets, and to examine the link between their partisanship and their assessment of the news sources.

3.2.2 Source-Based Strategies Used to Address Misinformation Recent work investigates leveraging the impacts of the news source and its perceived credibility on individuals' assessment of a message [37, 77, 191, 223]. For example, Bhuiyan et al. [37] employed a nudge-based intervention based authority of a source, referred to as the *reliable* nudge. They designed a browser extension for Twitter, named NudgeCred to investigate the efficacy of this nudge-based intervention. Through a five-day field experiment, this study demonstrates that NudgeCred improves individuals' recognition of misinformation. That is, signaling authority of source influenced the perceived credibility of the content. In particular, the participants rated posts with *reliable nudge* as more credible. (This study also investigated a community-based nudge, as discussed in §3.4.2).

In another case, Schwarz and Morris [210] present a visualization tool which augments search results with information on the credibility of web pages to help online users with credibility assessment. Conducting a user study, they demonstrate that signaling credibility of web pages can help people to assess information and identify credible content from non-credible content. However, this approach can result in perceiving the content published in new outlets whose credibility is unknown as less credible [210]. In a similar approach, Im et al. [121] design an approach to provide social signals about online accounts and inform others about whether a certain account engages in the spread of misinformation. The accounts that spread misinformation will receive a tag of misinformation where all other users can see. Through a field study, they investigate the utility of the approach and show the participants find these social signals helpful. Future research is required to investigate whether and how such signals might impact the behaviors of those users who receive these misinformation tags on their profile. Specifically, future research should investigate whether this approach prompt the receiver of this tag to engagement more with misinformation spread.

Epstein et al. [83] investigate using crowdsourcing to identify outlets that produce misinformation. Next, they use the crowds' ratings as an input to social media ranking algorithms. The results of their study demonstrate layperson trust ratings are an effective way to distinguish between high and low quality news outlets. The results of their experiment show the participants tend to trust mainstream sources much more than fake news sources. To test whether the participants would use this approach as a way to game the system, the participants were told their trust rating of news outlets will influence ranking algorithms of social media feeds. This information did not influence participants to game the system and promote their content of interest. In addition, Epstein et al. [83] argue that focusing on the source evaluation is more practical than focusing on individual articles for two reasons. First, rating credibility of news sources requires a much lower volume of rating. Thus, source-level rating is more scalable due to fewer number of news sources than news articles. Second, source-level ratings seem less susceptible to variation based on the idiosyncrasies of specific headlines. In another case, Pennycook and Rand [191] also investigate the crowds' rating of news sources trustworthiness, and suggests using these ratings in social media ranking algorithms. They also acknowledge crowd-sourced trust ratings can effectively differentiate more versus less reliable sources (i.e., the participants of their study rated mainstream sources as more trustworthy than hyperpartisan or fake news sources.)

In another example, using survey experiments, Dias et al. [77] investigate whether or not making the publisher information of a post more visible via adding a logo banner can improve assessment of the post. They demonstrate that publisher information had no significant effects on whether participants perceived the headline as accurate. When the headline was accurate but the publisher was a distrusted source, people tend to rate the content as less credible. Put differently, providing publisher information could increase the likelihood of mistakenly perceiving true headlines as false. Future research is required to investigate how providing different types of information about the source (e.g., adding a logo banner vs. detailed information) might have different efficacy in news assessment.

3.3 Factors about the Individual User

3.3.1 Individual User-Related Mechanisms The characteristics of individual users can impact how they interact and respond to different pieces of misinformation [27, 29, 54, 65, 80, 95, 173, 179]. These characteristics include people's prior beliefs [95, 173, 179], their media literacy skills [27, 29, 53, 129], as well as their personalities [54, 54, 65, 80]. The remainder of this subsection describes how each of these attributes play a role in people's response to misinformation.

3.3.1.1 Prior Beliefs. Individuals' prior beliefs about different topics influence news consumption [95, 171, 173, 179], and news assessment [179], but do not necessarily predict the way people share information [189].

Individuals' *news consumption* are influenced by their prior beliefs. That is, many individuals prefer to read and seek out information that is in line with their prior beliefs [124, 168]. This phenomenon is known as selective exposure effect [95, 96], and contributes to the spread of misinformation and its impacts. For example, Del Vicario et al. [75] show that selective exposure plays a role in the spread of conspiracy theories on Facebook. Selective exposure can also result in homogeneous and polarized communities [75, 151, 172], wherein people are more exposed to consonant content and less exposed to incompatible arguments. Such a closed system contributes to the creation of echo chamber [95, 186, 199], wherein people mostly see content that agrees with their preexisting.

Prior beliefs also influence the process of *news assessment* [179]. That is, many people tend to believe information that confirms their preexisting beliefs, a concept known as confirmation bias [179]. Due to confirmation bias, many people do not question the veracity of information with which they agree [149, 173, 179]. Confirmation bias can even influence people not to respond to corrective information objectively [107, 150]. In one case, Sleegers et al. [218] demonstrate many people tend to interpret ambiguous feedback on their incorrect beliefs in favor of their beliefs. In another case, Hameleers and van der Meer [107] illustrate that many people tend to avoid the articles that evaluate truthfulness of news (i.e., fact-checkers) if those outcomes are incompatible with their prior beliefs [107]. Similarly, prior beliefs can prevent some people from accepting corrective information [150, 234]. Taber and Lodge [234] argue that when people's prior beliefs are questioned, it might have a boomerang or backfire effect [59]. This backfire effect occurs because many people tend to actively counterargue incongruent evidence (i.e., dis-confirmation bias) [234]. They engage in "motivated reasoning" [88, 147], bringing arguments to defend their prior beliefs [131]. As a result, their prior beliefs will become stronger and they are more likely to believe in the original false content.

While prior beliefs influence *news consumption*[95, 168], and *news assessment* [107, 124, 173, 179], recent research suggests *sharing information* is not necessarily a reflection of prior beliefs [189]. Individuals' decisions about information sharing are more likely influenced by peripheral cues, such as the source of the information, and are moderated by individuals' self-presentation goals [62]. For example, individuals who want to fit in with a group prefer to share popular information from sources generally well-known and perceived as credible within their group. However, individuals who aim to stand out tend to share a piece of content regardless of the popularity of the content or its source [62].

3.3.1.2 Media-Related Literacy. Researchers have explored the link between different types of media-related literacy (i.e., media literacy, information literacy, news literacy, and digital literacy) and the way people identify and respond to misinformation [27, 29, 53, 129]. For example, investigating the role of digital media literacy on the way people respond to misinformation, Aufderheide [29] shows that people with greater digital media literacy are less susceptible to the misinformation and its impacts. In another case, Jones-Jang et al. [129] show that while both digital media literacy and information literacy are important, the latter is comparatively more influential on the way people respond to misinformation.

Sirlin et al. [217] investigate the role of digital media literacy, and acknowledged that of digital literacy is associated with less ability to tell truth from falsehood. This result does not vary based on the participants' partisanship and the type of the news. However, the skills of digital media literacy does not make people less likely to *share* false information. This finding could be because people are less motivated with the accuracy of content than with factors such as their emotion, and social feedback [116, 162, 192]. Buchanan [53] also argues that while media literacy is an important variable mediating the spread of misinformation and its impacts, people may know that a piece of content is untrue and still spread it anyway. They might be sympathetic to the content's intentions, or aim to signal their social identity or adherence to some political group or movement.

3.3.1.3 Personality. A line of research investigates whether personality traits have any effects on people's engagement with and their response to misinformation [54, 65, 80]. To measure personality traits, the Five Factor Model of personality (FFM), also known as the "Big Five" Costa Jr and McCrae [73], has been widely used [54, 65]. FFM includes extraversion, agreeableness, conscientiousness, neuroticism, and openness (See [73] for more details on these factors.)

Chen [65] shows that *openness* has a positive influence on sharing misinformation, while *neuroticism* has a negative influence on the sharing of misinformation. However, Buchanan and Benson [54] demonstrates that of the five dimension of personality, only *Agreeableness* personality had an impact on people's response to misinformation.

Zhu et al. [257] argue to investigate why people are influenced by misinformation, the role of personalities should be investigated along with other individuals' characterises that can induce false memories. They investigate the interaction effects between personality characteristics and cognitive abilities on individuals' vulnerability to misinformation. The result of their study demonstrate that low fear of negative evaluation, low harm avoidance, high cooperativeness, high reward dependence, and high self-directedness in combination with relatively low cognitive abilities make people more vulnerable to misinformation. In another study, Doughty et al. [80] show that personality characteristics are associated with memory conformity, which can make some people more susceptible to accepting misinformation as true. Specifically, they suggest low openness, extraversion, low neuroticism, and high agreeableness are related to memory conformity, and can make people more likely to accept misinformation as true.

As is apparent in this discussion, the results of prior studies that investigate the link between personality characteristics and people's susceptibility to the spread of misinformation draw different conclusions. Therefore, future research is required to clarify the sources of these mixed findings.

3.3.2 Individual User-Based Strategies Used to Address Misinformation Among the papers that focus on addressing misinformation at the individual user level, we found four general approaches: interventions that work towards improving individuals' media literacy skills [e.g., 33, 208], interventions aiming at shifting an individual's attentions towards credibility of content [124, 134, 189, 190], interventions that aims at identifying inauthentic accounts [e.g., 3, 7, 85], and interventions that work towards identifying and deplatforming dissimulators of misinformation [e.g., 242].

3.3.2.1 Improving Individuals' Media literacy Skills Various approaches have been designed to improving individuals' media literacy skills [34, 208, 241, 243]. In recent years, several game-based psychological interventions have been investigated to improve individuals' skills to identify misinformation [125, 169, 175, 208]. For example, Roozenbeek and van der Linden [208] designed a game, named "the bad news game", wherein players learn various skills to mitigate misinformation spread, such as detecting discredit, polarized arguments, impersonation, etc. Basol et al. [33] also designed another game, named "Go Viral!" to improve abilities of individuals to detect manipulation techniques that are used in COVID-19 misinformation. The game increased the abilities of the participants to identify misinformation, and reduced their willingness to share misinformation with others. In another case, Jeon et al. [125] designed a game, named "ChamberBreaker", which is designed to increase a player's awareness of echo chamber effect and the importance of maintaining diverse perspectives when consuming information. After playing the game, the players showed greater intention to see information from more diverse perspectives and more awareness of the possible echo chambers.

Karduni et al. [135] designed a visual analytic system, named "Verifi2", that help social media users distinguish misinformation. Verifi2 highlights different aspects of a piece of news, such as its linguistics, its networks of spread, and highlights image features related to the news to help online users to learn dimensions that characterize misinformation, and learn how suspicious news are different with true content.

In another case, Tsipursky and Morford [241] designed an intervention named the Pro-Truth Pledge (PTP) where the signees agree to abide by twelve behaviors to mitigate the spread of misinformation, including verify, balance, cite, clarify, acknowledge, reevaluate, defend, align, fix, educate, defer, and celebrate. (See [241] for more details). The participants, including both private citizens and public figures, self-reported the impact of this pledge on their behaviors regarding their response to misinformation.

While valuable, improving people's skills to identify false and misleading content does not always lead to changing people's intention to share misinformation. For example, many people may know that a piece of content is untrue and still spread it anyway for reasons such as signaling their social identity or adherence to some groups [53, 129]. To improve the efficacy of the interventions around media literacy, other factors such as the role of prior beliefs should also be taken into account. In addition, the studies that are reviewed here analyse the efficacy of the intervention based on self-reported data. Future research should investigate how these approaches may help in real-word settings, outside the experiment. Additionally, future investigation is required to assess the long-term efficacy of these prebunking-based and pledge-based interventions.

3.3.2.2 Shifting an Individual's Attention to the Credibility of Content Only recently researchers focused on reducing the spread of misinformation via shifting attention of online users toward the credibility of what they share [39, 124, 189, 190]. Recent work shows that many people are motivated by their emotions and social feedback when sharing a piece of content online [116, 162]. Motivated by these insights, Pennycook et al. [189] examine whether encouraging online users to reflect on the accuracy of content could make them less likely to spread misinformation [189]. Using a field experiment on Twitter, they selected Twitter users who regularly shared misinformation. Next, the researchers send those users private messages, asking them to rate the accuracy of a headline. By inviting these users to reflect on the accuracy aspect of the news, this intervention increased the average accuracy of the news that those users shared in the next 24 hours.

In another example, Jahanbakhsh et al. [124] investigated the effects of two behavioral nudges which request accuracy assessments and rationales, on sharing false news. Specifically, the first nudge asked people to assess the accuracy of the content they were about to share, and the second nudge asked people why they think the content is or is not accurate at the time of sharing it. The results of this study revealed that both accuracy assessment nudge and rational nudge can reduce the sharing both false content and true content. However, these nudges also reduced sharing of true content to a lesser degree compared to the sharing of false content, resulting in an overall decrease in the fraction of shared content that is false [124].

3.3.2.3 Identifying and Removing Disseminators of Misinformation At the first glance, this category of interventions focus on the creators of misinformation (e.g., identifying and removing inauthentic accounts, deplatforming disseminators of misinformation), which does not fall in the scope of this review. However, these interventions, by removing the disseminators of misinformation and signaling the audience about it, may still impact people's perceptions of and their *response* to the content they have previously read from these accounts. Therefore, this review includes and examines these approaches as well.

One of the approach in this category is *identifying and removing inauthentic accounts* [3, 7, 85]. This approach focuses on the activities of individual users and addresses the users who engage in suspicious behaviors and spread misinformation(e.g., artificially boost the popularity of content, or impersonating another person such as politicians or celebrities). Platforms such as Facebook and Twitter, identify and remove accounts that exhibit inauthentic behavior [3, 7, 85]. For example, Facebook introduced a concept named "coordinated inauthentic behavior" (CIB), which is defined as the use of Facebook or Instagram assests (i.e, accounts, pages, groups, or events) to mislead people [85]. The networks of people who engage in coordinated inauthentic behavior focus on two activities. In one case, they create fictitious, independent media entities and personas to engage unwitting individuals to amplify their content and expand their reach. In another case, they drive people to other websites that their networks control [85]. Facebook identities and disables these accounts. Upon disabling an account, all the content within the account becomes inaccessible to other users.

While identifying and removing these networks of accounts might mitigate the spread of misinformation [12], recent research reveals that the success of these accounts heavily lies on the activities of online crowds (as apposed to coordinated networks of accounts) and the ways they engage in the spread of misinformation and its impacts [104, 224, 246]. The engagement of online communities with misleading content, amplifying it, and expanding its reach make it possible for coordinated networks to expand their networks and push their goals. Therefore, policies that focus on addressing misinformation by identifying "coordinated" inauthentic behaviors fall short in addressing the circulation of misleading content that is shared by online crowds.

Another approach in this category is *deplatforming*, a moderation strategy that refers to the permanent ban of conversational influencers for spreading misinformation, conducting harassment, or violating other platform policies [103]. Facebook, Twitter, Instagram, YouTube and other platforms have employed deplatforming to mitigate the spread of misinformation, hate speech, and conspiracy theories in various cases [42, 74, 143]. For example, upon investigating tweets from the @realDonaldTrump account and the way they were being interpreted on and off Twitter, Twitter suspended the account permanently to prevent the risk of further inducement of violence by this account [242]. (See Figure 3).

While deplatforming has been shown to decline toxicity level of supporters [126], the efficacy of this approach to combat misinformation is nuanced by a lot of factors. For example, in the

Fig. 3. The figure shows deplatforming of @realDonaldTrump account, which occurred on January 8th, 2021. Twitter announced the account due to the risk of further incitement of violence [242].

case of combating anti-vaccine misinformation, Armitage [26] argues that deplatfoming of anti-vaccine campaigners is likely to reinforce individuals' strongly held beliefs about vaccination and vaccine conspiracies. In addition, Innes and Innes [122] argue that de-platforming might result in "re-platforming", which refers to behaviors such as developing a network of alternative accounts and signaling their presence, and migrating to one or more other platforms.

3.4 Factors about the Community

3.4.1 Community-Related Mechanisms Prior work has acknowledged the different mechanisms by which online communities contribute to the spread of misinformation and its impacts [36, 99, 109, 112, 212, 233, 240]. These mechanisms can be organized into roughly three groups: individual level mechanisms, such as familiarity bias and trust in community-shared content [36, 99, 109, 112, 117, 212, 233, 240]; network mechanisms, such as social network structure and homophily [78, 99, 112]; and social norms, especially perceived norms [70, 101, 101, 144, 207]. The remainder of this subsection describes how each of these mechanisms works in the spread and the impacts of misinformation.

First, online communities can influence the way individual members respond to misinformation via different mechanisms. For example, the activities of online communities (e.g., sharing, liking, commenting) can increase the likelihood of individual members' exposure to misinformation. For instance, when a contact of ours “likes” a post, we are more likely to see the content [24]. This increased visibility can also increase individuals' exposure to misinformation, which contributes to familiarity bias [36, 109, 212, 233, 240], wherein an individual remembers the content itself, but forgets the contexts and details around it. Familiarity bias increases the likelihood of accepting familiar but false information as true [36, 109, 212, 233, 240], leading to the further spread of misinformation.

Another mechanism by which online communities influence individual members' response to misinformation is via increasing trust in the content that is shared in the community [112]. That is, many people are more likely to trust what their networks share [112] and less likely to question content that is shared by their networks [112]. As a result, the content that is shared by an individual's communities are more likely to be perceived as true [99].

Second, the structure of social networks within and across online communities play a role in how misinformation spread and has its impacts [92, 149, 152]. For example, online communities might form homophilous network structures, where there are more connections among people who share similar views [184, 202]. Such networks make people more likely to be exposed to consonant information, and less likely to observe arguments that might challenge their view [1, 17, 43, 57, 71, 118, 209, 220]. For example, examining two polarized communities on Facebook, science and conspiracy communities, Quattrociocchi et al. [202] show that people of each community tend to be connected only with like-minded people and not to interact with people of the other community. In such networks, people's views are less likely to be challenged, leading many individuals to become more confident in their views [230]. Homophily can also explain why people in the same network are more likely to believe in the same rumor [78]. This way, networks that connect similar individuals make it possible for misinformation to spread more quickly within a community.

Third, social norms are another key mechanism by which online communities play a role in the spread of misinformation and its impacts. Social norms refer to people's perceptions around what others do (i.e., descriptive norms), what others approve of and what they condemn (i.e., injunctive norms), and how an individual thinks they are expected to behave (i.e., subjective norms) [67, 207]. These perceptions, combined with the desire to be liked or to obtain approval from others, can influence people's behaviors [67, 207]. These behaviors include the types of content that people share within their community [23, 28, 182, 187, 251], and the way they identify misinformation and respond to it [70, 101, 144]. For example, descriptive norms around content sharing (i.e., perceptions about what other members of the community share) can influence the perceived popularity of certain information. The perceived popularity of a claim in turn increases its perceived reliability [31, 36], making it more likely to be accepted as true. Injunctive norms can influence people's perceptions about what a community approves or disapproves of sharing. People also form perceptions about how they are expected to behave (i.e., subjective norms), which influence their decision about whether or not to share a piece of content [70]. Perceptions of social norms can also influence the way people assess and respond to misinformation [70, 101, 144]. For example, people who perceive of correcting misinformation as a common practice within a community may be more likely to correct misinformation themselves [144].

At the same time, individuals' may have inaccurate perceptions of social norms [207]. Such misperceptions can contribute to the "majority illusion effect" [137], wherein a few highly connected users with misleading views can skew the perceptions of many others and even trigger a rapid change in the community's view [137, 152]. For example, misperceptions due to majority illusion effect can make anti-vaccine views seem like the majority's opinion, even when they are not [128, 221]. The majority illusion effect induces some people to overestimate the prevalence of their view in the population and assume the majority share their view [159], a concept known as "false consensus" [159, 161]. Additionally, majority illusion can make people incorrectly think the opposite of their view is held by the majority of the population when in fact the majority share their view, a concept known as "pluralistic ignorance" [170, 200]. As a result of the pluralistic ignorance effect, group members may behave contrary to their own preferences in favor of what they think is popular. The pluralistic ignorance effect may be a key factor that contributes to people's hesitation to correct perceptions of others about misinformation [153]. Thus, the perception of social norms can play a key role in how communities respond to and spread misinformation.

3.4.2 Community-Based Strategies Used to Address Misinformation Despite the significant role of online communities in the spread of misinformation and its impacts (Discussed in §3.4), much less attentions has been paid to designing around community factors that contribute to the spread of misinformation. We found only two interventions that are designed based on community factors for this purpose [23, 37]. We found several interventions that leverage the crowds to improve the efficacy of approaches that aim at identifying misinformation [138, 178, 178]. However, these interventions are not designed around the community-oriented factors of misinformation spread.

In one case, Andı and Akesson [23] investigate whether a social norm-based nudge can result in sharing less misinformation. To do so, they use a message that inform participants about the abundance of false information online and warns them that most responsible people think twice before sharing a piece of news with their networks (i.e., descriptive norms intervention). The message is displayed above the articles that the participants see and reads as "NOTICE: There is a lot of misleading and false information online. Most responsible people think twice before sharing content with their friends and followers". The participants who were nudged using this message expressed less willingness to share false information. However, it is not immediately clear whether the result is solely due to priming effect of informative message about the abundance of false information, or due to conformity to social norms.

In another case, Bhuiyan et al. [37] employed a nudge-based intervention based on users' collective opinion on a report. This nudge is designed to highlight the number of question marks in the comment section, and is named as the *questionable* nudge. To investigate the efficacy of this approach, Bhuiyan et al. [37] designed a browser extension for Twitter, named NudgeCred. Through a five-day field experiment, they demonstrate the NudgeCred influenced the perceived credibility of the content. That is, the participants rated posts with *questionable nudge* as less credible.

A few other studies leverage online crowds to improve the efficacy of automatic fact-checkers [138, 178]. For example, Nguyen et al. [178] combine machine learning techniques with the crowd annotations to improve the efficacy of fact-checking approach in terms of predictive performance, its speed, as well as interpretability of the the predictive model. In particular, the crowd helps to provide explanations about the reputation of the news source to improve acceptance of the outcome of the model. The provided explanations about the reputation of the news source improved users' satisfaction and trust in model predictions. While the presented model improves transparency of the outcome, and helps speeding the fact-checking process, the focus of this study is not on leveraging the influence of community-oriented factors to impact response to misinformation. Instead, the model still focuses on identifying individual pieces of misinformation, and helping online users to be informed of the credibility or lack of credibility of different news sources.

4 Discussion: Consequences that Arise from an Individual Focus Approach to Addressing Misinformation

Current approaches to addressing misinformation heavily focus on individual pieces of misinformation and the actions individuals take in response to those individual pieces. Despite the variation in the details of the intervention methods reviewed above, our analysis shows that of 67 intervention papers, 40 of them focus on *individual pieces of content*, 10 examine strategies based on attributes of source, and 16 target *individual users* to address the phenomena of misinformation.

This section argues that neither a focus on individual pieces of misinformation (either based on attributes of content itself or based on attributes of the source of content) nor a focus on individual users will be sufficient in addressing the issues associated with misinformation. To do so, it first highlights a series of two implicit *assumptions* that our analysis reveals in approaches that focus on individual content or users. In practice, we argue, these assumptions rarely hold. Second, it points out *blind spots* that arise from a focus on individual content (either based on the content itself or its source) or on individual users. These blind spots occur in part because this individualistic focus makes crucial aspects of the issues around misinformation either less readily apparent or entirely out of scope.

4.1 Assumptions

4.1.1 Humans are rational actors persuaded by additional or corrective information. Most individual-focused approaches are predicated on the assumption that individuals are rational actors who engage in logical reasoning about the content view. Many such approaches provide additional, corrective information or alternative stories. For example, TwitterTrails [167, p. 71] "does not answer directly the question of a story's validity, [but] it provides information that a critically thinking person can use." Implicitly, a person with more information will make the correct decisions. Many other interventions similarly provide additional information to help the user be more informed [2, 63, 140, 153].

However, much of an individual's decision-making stems from sources other than pure rationality. Self-perceived rationality of the content and its source [145], community narratives [219], motivated reasoning [131, 147], familiarity bias [233] and other factors can have a greater influence than purely providing more information. For example, people may even know a piece of content is untrue and still spread of anyway for reasons such as signaling their social identity [53]. In such cases, providing additional or corrective information is less likely to influence an individual's decision about sharing misinformation.

Furthermore, Lazer et al. [150] discuss how tools that attempt to provide corrective information, such as fact-checkers, can actually reinforce the false information they seek to correct. This downside occurs via a two-step process. First, fact-checkers simply increase the familiarity with claims that are false, contributing to familiarity bias. Second, familiarity bias makes people more likely to accept information that is familiar to them [36, 109, 212, 233, 240]. This way, familiarity bias further contributes to the spread of misinformation and its impacts [150]. However, approaches such as fact-checkers that hinge on identifying individual pieces of untrue content, not only fall short in addressing the broader effects of familiarity bias, but also contribute to the strength of this phenomenon. As a result, these approaches are less likely to influence people's response to the spread of misinformation and not sufficient in mitigating its impacts [98, 149].

4.1.2 Each individual encounter with misinformation occurs in isolation. The individualistic approaches we reviewed make it difficult to account for the role of prior beliefs and the social contexts in which individuals encounter misinformation. However, as discussed throughout §3, an individual's prior beliefs [173, 179], as well as the social context in which misinformation is encountered, can drive an individual's response to misinformation.

For example, if corrective or additional information contradict an individuals' prior beliefs, it is less likely to change the person's opinion about the original misinformation [149]. Indeed, corrective or additional information might even fuel "backfire effects" [21, 98, 180]. This effect occurs because many people do not assess information objectively [173, 179]. Instead, they process information to confirm to their preexisting beliefs, a phenomena known as "motivated reasoning" [88, 133, 147]. For example, if the corrective information is perceived as an identity threat, people can become more defensive and bring different reasons to counter the corrective information [131, 147, 157]. This phenomenon can in turn strengthen their beliefs in the original false information.

Additionally, many people trust information that is shared by their network [112] and often accept such information at face value [99]. In such cases, corrective or additional information is less likely to change people's opinion about the original misinformation [149].

Furthermore, social contexts in which people encounter misinformation influence people's perceptions of what that community accepts and what that community condemns [196]. Such perceptions play a role in the assessment of information and the way people respond to misinformation. However, the individualistic approaches that we reviewed implicitly assume individuals identify and respond to misinformation only based on misinformation itself.

4.2 Blind spots

4.2.1 The social context(s) in which people encounter misinformation. People's response to different content is not only a function of the content itself, or even individuals' knowledge about the content [149, 173, 179]. Rather, as discussed in Sections 3.3 and 3.4, people's response to misinformation is influenced by different factors, including the social context in which they encounter misinformation and the ways others within a community engage with and respond to misinformation [70, 144, 195, 231]. For example, due to a desire to gain social approval, people may consider how others respond to an issue, such as in comments [99], when deciding on how to respond to it themselves [196].

As a result, even if identifying all the false or misleading arguments were possible, and even if it was possible to combat misinformation by influencing individual responses to misinformation, it is not sufficient to address misinformation only based on misinformation itself.

4.2.2 Statements that are factually true but misleading. Approaches at content level only focus on and evaluate the truth value of individual claims. However, not all misleading claims are factually incorrect. In many cases, the actors state their argument based on factually correct pieces of content, but reshape the true statements and fit the pieces together to push a misleading argument [224].

For example, actors with anti-vaccine beliefs might argue vaccines include certain chemical ingredients, e.g.: "Thimerosal is a mercury-based preservative used in vaccines. mercury is a known neurotoxin." Both of these statements are factually correct. However, the actors who spread the content, purposefully omit the fact that medical research suggests that the portion of these chemical ingredients in vaccines is safe for the human body [204]. For another example, anti-immigrant groups often argue that immigrants take away jobs from American workers [47], thus claiming that immigrants are a threat to Americans. However, they do not mention that immigrants create new jobs by forming new businesses, paying taxes, and contributing to the productivity of businesses in the U.S. [214].

Indeed, for these misleading arguments to be influential, they need to be stated based on a core of verifiable information [44]. Thus, approaches that solely focus on the truth or falsity of individual pieces of content simply are unable to recognize such instances of a misleading arguments that are based on factually true statements.

4.2.3 The systemic nature of the impacts of misinformation. The individualistic approaches draw attention away from the way an online community as a whole responds to misinformation. The response of a community to misinformation is influential on the way people perceive a community, and make inferences about what a community as a whole accept of and what the community disapprove of [196, 207]. These perceptions about a community influence how one thinks they are expected to behave within a certain community [66]. These behaviors include responses to misinformation.

However, a community is more than the aggregate of its members, and the response of a community to misinformation goes beyond an aggregation of the responses of its individual members to misinformation. Thus, the overall community response cannot be improved simply via changes in individual responses to specific pieces of misinformation. Rather, the response of a community to misinformation is also influenced by other properties of the community, including the pattern of interactions, norms of content selections and content sharing, and the way the community as a whole views various issues. Together, these properties and the perceptions they create form how a community as a whole responds to misinformation, impact the response of its members to various issues.

The individualistic approaches draw attention away from the broader, systemic issues to which the spread of misinformation contributes. For example, the circulation of misinformation can cause long-term attitudinal and behavioral shifts [197, 256]. As another example, the spread of misinformation allows actors with political goals to influence and reshape social structures and the types of conversations that take place around certain issues, as well as the arguments that are heard more within a community [224]. Manipulating social structure for political goals can contribute to broader impacts of misinformation, including the rise in polarization [32, 150, 224, 226], amplifying political divisions among a society, and undermining faith in authorities and science [108, 154]. These broad, systemic impacts of misinformation can be more harmful than the spread of factually incorrect content [22]. However, focusing primarily on individual pieces of content or users constrains our vision, both in terms of what misinformation can do to online communities and in terms of how we might most effectively respond.

4.3 Summary

As discussed throughout this section, current approaches of combating misinformation usually treat misinformation as individual pieces of content that need to be addressed. Given the definition of misinformation as "false or misleading information" [150, 224, 247], which refers to information as pieces of content, this finding is perhaps unsurprising. However, in addition to the content of misinformation, there are other factors that drive an individual's response to misinformation, including their prior beliefs, as well as the social contexts in which they encounter misinformation.

In addition, while misinformation itself refers to false or misleading information, its impacts go beyond misleading people about individual factually untrue statements. Therefore, rather than completely abandon the existing approaches of combating misinformation that focus on individual pieces of misinformation, we should instead complement those approaches with interventions that address the broader, systemic nature of misinformation and its impacts.

5 Limitations

This work has three primary limitations: First, this review uses a search of the ACM Digital Library, and in our community of interest, i.e., SIGCHI, to collect papers. While this search space is chosen for the aforementioned reasons (§2), future work can move beyond this community, and investigate the under-explored areas to design interventions in other communities. Second, the paper chooses a single search term, i.e., misinformation, to exclude work not specifically related to the misinformation phenomena. Other prior reviews have similarly conducted their search processes around very constrained search terms [e.g., 35, 46, 52, 79, 139]. Such a search process helps to focus exactly on the phenomenon of interest and to exclude other relevant but not similar phenomena. However, there might be work that refers to the same phenomena using alternative keywords (e.g., false news). Searching for and including such potential papers can result in a more comprehensive literature search.

Third, given the rapid movement of social media, and the ever chaining nature of social media scenarios [11, 155, 244], the findings of this paper may not hold in future social media scenarios. Specifically, the gaps found in interventions around misinformation may shift from community oriented factors to any of other drivers of misinformation in the future. In addition, the drivers of misinformation may evolve over time. For example, community oriented aspects of misinformation discussed in this work may change over time. Therefore, it is important for future work to both examine drivers of misinformation based on new social media scenarios, and to examine and reveal under-explored areas to design interventions. Of course, this challenge is not limited to the phenomenon of misinformation, and applies on many other reviews that studies phenomena related to social media. That said, the current review and its findings reveals important gap in research around misinformation interventions at the time of publishing this work, and guides future work to move towards addressing this gap.

6 Future Work: Leveraging the Role of Social Norms as One Way to Account for Community Factors in Addressing Misinformation and its Impacts

To address the consequences of individualistic focus on misinformation described above, it is important for future work to complement these existing interventions with additional approaches that account for the community-oriented factors that contribute to the spread of misinformation. There is a wealth of different group and community level processes that could be relevant. These processes include familiarity bias, social network structures, majority illusion, among many others (discussed in §3.4). Space precludes a full discussion of how each of these mechanisms might be enlisted to inform the design of interventions to address misinformation.

Instead, as just one example, this section describes how social norms could be leveraged to mitigate the spread and impacts of misinformation. Social norms are defined as "rules or standards that are understood by members of a group, and guide/or constrain social behavior without the force of law" [67]. Norms influence people's behaviors in a variety of subjects [e.g., 58, 156, 163, 237], including the way people respond to misinformation [70, 101, 144]. This section offers potential interventions on two different levels (i.e., individual level and community level) to illustrate the use of social norms to mitigate the spread and impacts of misinformation.

At individual level, social norms influence the way people identify misinformation and respond to it [70, 101, 144]. Therefore, future work could make valuable contributions by investigating possible ways to manipulate perceptions of social norms and guide individual responses to misinformation. For example, perceived social norms can be manipulated to promote the use of current approaches of identifying misinformation (e.g., fact-checkers, related articles, reporting misinformation, etc.), and encourage engagement with practices of combating misinformation (e.g., flagging, reporting and correcting misinformation). One possible path to do so could be via providing normative information. For example, normative information can be framed around descriptive norms, emphasizing on how the other members of the community put high importance on a quality of content that is shared in their community, and engage in practices of combating misinformation. Given that people are reactive to others behavior [207, 222], such information can potentially influence people's response to misinformation. Alternatively, normative information can be framed around injunctive norms, emphasizing on how a community does not approve of sharing false and misleading information. Given that people avoid risking interpersonal harmony by going against others' expectations [66], providing such information can make people more wary of the quality of the content they share. Future research is required to investigate how such information should be framed to be influential.

These few examples illustrate how it may be possible to draw on the literature about social norms, and guide the design of sociotechnical interventions to mitigate the spread and impacts of misinformation.

At community level, social norms influence the social contexts in which people encounter misinformation via different mechanisms. Social norms influence content sharing [23, 28, 182, 187, 251], social network structures [23, 28, 182, 187, 251]. Together, these mechanisms contribute to the ways by which misinformation has its impacts on online communities, broader than misleading individual members about single pieces of misinformation. For example, via influencing social network structures and behavior around content sharing, social norms might contribute to issues such as the rise in polarization [160], and the spread of mistrust in science and authorities [15, 40].

These insights suggest opportunities for future research to investigate ways to leverage social norms and mitigate the systemic impacts of misinformation. For example, future research could examine different ways to leverage the impacts of social norms to encourage seeking two-sided arguments and cross cutting discussions as a way to mitigate the impacts of the selective exposure effect. One way to do so would be to communicate social norms and underscore how other members of a community seek and share different perspectives on the same topic (i.e., descriptive norms around content selection). Given that people are reactive to others behavior [207, 222], such information can potentially mitigate selective exposure effects. Alternatively, normative information can be highlighted to show how other members of a community approve of and favor hearing different perspectives on the same topic (i.e., injunctive norms). Due to a desire to be approved by others [153, 211, 248], this information might promote people to engage more in cross-cutting discussion and to hear the other side.

While social norms provides a pathway to address some of the limitations of the individualist approaches to misinformation, designing around social norms has several **challenges and limitations**. First, social norms arise organically and are not directly the result of design and planning [41]. It is hard to ask for certain social norms and to bring the desired social norms into practice [87]. For example, enforcing social norms that value truth and protect integrity of information space can be extremely difficult. One way to address this challenge can be to influence *perceptions* of social norms rather than aiming to influence community's actual social norms (i.e., collective norms). Indeed, individuals' perceptions of social norms are more influential on people's behavior compared to the actual norms [205], and have been successfully employed in enforcing positive behavioral change around various subjects [205, 206, 252]. Second, people are often members of different communities, which can all have different social norms. In such cases, salience of membership [136], different types of anonymity in each community [115], needs to belong [41], and other factors can impact individuals' conformity to those social norms. Future research should investigate and account for such influential factors when designing interventions to manipulate perceptions of social norms. Third, norm-based interventions are less likely to work in communities that do not put high importance on good intentions, such as intellectual honesty and fairness, truthfulness, and impartiality of discussions. For example, in a white supremacist group, where many people hold radical beliefs and the community as a whole actively denies social and political freedom of minorities, social norms interventions are less to be practical. Fourth, even in communities that do support good intentions (e.g., truthfulness, impartiality of discussions), social norms may be less likely to influence those individuals who hold strong beliefs about an issue. For instance, an individual who has strong anti-vaccine beliefs may be less likely to change their beliefs as a result of normative influences.

7 Conclusion

While there are numerous approaches designed to address misinformation, our analysis reveals that most of these approaches heavily focus on individual pieces of misinformation and the response of individual users to misinformation. We argue that these individualistic approaches of misinformation rely on some implicit assumptions that rarely hold in practice. That is, they tend to assume that people are persuaded by corrective or additional information. However, other than the content, there are other factors that influence an individual's response to misinformation, such as self-perceived rationality of the content and its source [145], community narratives [219], motivated reasoning [131, 147], familiarity bias [233]. In addition, the individualistic approaches implicitly assume that people encounter misinformation in a vacuum, and do not account for the role of prior beliefs, and the social contexts in which misinformation occurs. However, prior work has acknowledged such factors influence the way people identify and respond to misinformation [e.g., 70, 101, 171, 173, 179].

In addition, we argue that individualistic approaches overlook some aspects and impacts of misinformation. In particular, they make it difficult to account for statements that are factually true, but misleading arguments that are based on factually true statements [47, 204]. In addition, individualistic approaches draw attention away from the systemic nature and consequences of misinformation. Examples of the systemic issues of misinformation include the rise in polarization, and the spread of distrust among communities. These issues pose varying degrees of harm on online communities, broader than misleading an individual about certain issues [22].

To address misinformation and its broader impacts, future research should expand its scope beyond individual pieces of content to account for the community-oriented factors that contribute to the spread of misinformation and its impacts. Put differently, rather than completely abandon, say, fact checkers, future work should complement such individualistic approaches with community oriented interventions to address the broader, systemic nature of misinformation and its impacts.

To provide an illustration of how community-oriented factors can be used for this purpose, this paper discusses the role of social norms. We argue that social norms can be leveraged to address some of the limitations of individualistic approaches to misinformation without leaving behind the individual user and their individual response to misinformation. We acknowledge that social norms are not the only mechanism for an intervention to address misinformation. However, social norms are among the key community-oriented mechanisms that provide a valuable pathway by which misinformation and its various impacts can be mitigated. Similarly, identifying and leveraging other community-oriented mechanisms that contribute to the spread and impacts of misinformation warrant future research attention. Indeed, such interventions that move beyond an individualistic approach will be necessary if we hope to address the broader, systemic nature of misinformation and its impacts.

8 Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments.

References

[1] 2006. Hearing the other side: Deliberative versus participatory democracy. Cambridge University Press.
[2] 2017. Fact-checking fake news on Facebook works - just too slowly. https://apnews.com/article/north-america-technology-business-fake-news-ap-fact-check-e03283b4169f4d8c8a7e51042d61bcb5.
[3] 2018. How Twitter is fighting spam and malicious automation. https://blog.twitter.com/en_us/topics/company/2018/how-twitter-is-fighting-spam-and-malicious-automation
[4] 2018. How whatsapp helped turn an Indian village into a lynch mob. https://www.bbc.com/news/world-asia-india-44856910
[5] 2019. Designing new ways to give context to news stories. https://newsroom.fb.com/news/2018/04/inside-feed-article-context/
[6] 2019. Helping ensure news on Facebook is from trusted sources. https://about.fb.com/news/2018/01/trusted-sources/
[7] 2019. Information operations directed at Hong Kong. https://blog.twitter.com/en_us/topics/company/2019/information_operations_directed_at_Hong_Kong
[8] 2020. Expanding Facebook's U.S. Fact-Checking Program and Supporting the Fact-Checking Ecosystem. https://www.facebook.com/journalismproject/fact-checking-expansion-and-investment-2020
[9] 2020. Fake news and its impact on the economy. https://priorityconsultants.com/blog/fake-news-and-its-impact-on-the-economy/.
[10] 2020. Helping fact-checkers identify false claims faster. https://about.fb.com/news/2019/12/helping-fact-checkers/
[11] 2021. The evolution of social media: How did it begin and where could it go next? https://online.maryville.edu/blog/evolution-social-media/
[12] 2021. Removing coordinated inauthentic behavior. https://about.fb.com/news/2020/09/removing-coordinated-inauthentic-behavior-russia/
[13] 2021. Twitter launches crowd-sourced fact-checking project. https://apnews.com/article/twitter-launch-crowd-sourced-fact-check-589809d4c9a7eceda1ea8293b0a14af2
[14] 2022. https://sites.google.com/view/design4misinformation/home. (Accessed on 10/05/2022).
[15] 2022. How Americans came to Distrust Science. https://bostonreview.net/articles/andrew-jewett-science-under-fire/
[16] 2022. Research guides. https://guides.temple.edu/fakenews. (Accessed on 10/05/22).

[17] Lada A Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 US election: divided they blog. In Proceedings of the 3rd international workshop on Link discovery. 36–43.

[18] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. Journal of economic perspectives 31, 2 (2017), 211–36.

[19] Jennifer Allen, Cameron Martel, and David G Rand. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In CHI Conference on Human Factors in Computing Systems. 1–19.

[20] Kimberley R Allison, Kay Bussey, and Naomi Sweller. 2019. 'I'm going to hell for laughing at this' Norms, Humour, and the Neutralisation of Aggression in Online Communities. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), 1–25.

[21] Craig A Anderson, Mark R Lepper, and Lee Ross. 1980. Perseverance of social theories: the role of explanation in the persistence of discredited information. Journal of personality and social psychology 39, 6 (1980), 1037.

[22] Janna Anderson and Lee Rainie. 2020. The Future of Truth and Misinformation Online. https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/

[23] Simge Andı and Jesper Akesson. 2020. Nudging Away False News: Evidence from a Social Norms Experiment. Digital Journalism 9, 1 (2020), 106–125.

[24] Nicolas M Anspach. 2017. The new personal influence: How our Facebook friends influence the news we read. Political Communication 34, 4 (2017), 590–606.

[25] Ahmer Arif, John J Robinson, Stephanie A Stanek, Elodie S Fichet, Paul Townsend, Zena Worku, and Kate Starbird. 2017. A closer look at the self-correcting crowd: Examining corrections in online rumors. In Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. 155–168.

[26] R Armitage. 2021. Online 'anti-vax'campaigns and COVID-19: censorship is not the solution. Public Health 190 (2021), e29.

[27] American Library Association et al. 2009. Association of College and Research Libraries. Information literacy competency standards for higher education (2009), 2–3.

[28] Charles K Atkin. 1985. Informational utility and selective exposure to entertainment media. Selective exposure to communication (1985), 63–91.

[29] Patricia Aufderheide. 2018. Media literacy: From a report of the national leadership conference on media literacy. In Media literacy in the information age. Routledge, 79–86.

[30] Cathrine Axfors, Andreas M Schmitt, Perrine Janiaud, Janneke van't Hooft, Sherief Abd-Elsalam, Ehab F Abdo, Benjamin S Abella, Javed Akram, Ravi K Amaravadi, Derek C Angus, et al. 2021. Mortality outcomes with hydroxychloroquine and chloroquine in COVID-19 from an international collaborative meta-analysis of randomized trials. Nature communications 12, 1 (2021), 1–13.

[31] Frederick T Bacon. 1979. Credibility of repeated statements: Memory for trivia. Journal of Experimental Psychology: Human Learning and Memory 5, 3 (1979), 241.

[32] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. Science 348, 6239 (2015), 1130–1132.

[33] Melisa Basol, Jon Roozenbeek, Manon Berriche, Fatih Uenal, William P McClanahan, and Sander van der Linden. 2021. Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. Big Data & Society 8, 1 (2021), 20539517211013868.

[34] Melisa Basol, Jon Roozenbeek, and Sander van der Linden. 2020. Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. Journal of cognition 3, 1 (2020).

[35] Eric PS Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing reflection: on the use of reflection in interactive system design. In Proceedings of the 2014 conference on Designing interactive systems. 93–102.

[36] Ian Maynard Begg, Ann Anas, and Suzanne Farinacci. 1992. Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. Journal of Experimental Psychology: General 121, 4 (1992), 446.

[37] 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 (2021), 1–30.

[38] 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'& Consumers' Perspectives. Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (2021), 1–31.

[39] 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. 205–208.

[40] Cristina Bicchieri, Enrique Fatas, Abraham Aldama, Andrés Casas, Ishwari Deshpande, Mariagiulia Lauro, Cristina Parilli, Max Spohn, Paula Pereira, and Ruiling Wen. 2021. In science we (should) trust: Expectations and compliance across nine countries during the COVID-19 pandemic. PloS one 16, 6 (2021), e0252892.
[41] Cristina Bicchieri and Ryan Muldoon. 2011. Social norms. (2011).
[42] Nick Bilton. 2019. The downfall of Alex Jones shows how the internet can be saved. https://www.vanityfair.com/news/2019/04/the-downfall-of-alex-jones-shows-how-the-internet-can-be-saved
[43] George D Bishop and David G Myers. 1974. Informational influence in group discussion. Organizational behavior and human performance 12, 1 (1974), 92–104.
[44] Ladislav Bittman. 1985. The KGB and Soviet disinformation: an insider's view. Washington: Pergamon-Brassey's.
[45] 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 (2015), 619–638.
[46] Kirsten Boehner, Janet Vertesi, Phoebe Sengers, and Paul Dourish. 2007. How HCI interprets the probes. In Proceedings of the SIGCHI conference on Human factors in computing systems. 1077–1086.
[47] George J. Borjas, Michael Grunwald, Thomas Frank, Jeremy B. White, Sam Sutton Sitrin, Carly, Bill Mahoney Gerstein, and Josh. [n.d.]. Yes, immigration hurts American workers. https://www.politico.com/magazine/story/2016/09/trump-clinton-immigration-economy-unemployment-jobs-214216/
[48] Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. Nature communications 10, 1 (2019), 1–14.
[49] Samantha Bradshaw, Lisa-Maria Neudert, and Philip N Howard. 2018. Government responses to malicious use of social media. NATO StratCom Centre of Excellence, Riga, Working Paper (2018).
[50] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. Qualitative research in psychology 3, 2 (2006), 77–101.
[51] Lukas E Brümmer, Stephan Katzenschlager, Mary Gaeddert, Christian Erdmann, Stephani Schmitz, Marc Bota, Maurizio Grilli, Jan Larmann, Markus A Weigand, Nira R Pollock, et al. 2021. Accuracy of novel antigen rapid diagnostics for SARS-CoV-2: A living systematic review and meta-analysis. PLoS medicine 18, 8 (2021), e1003735.
[52] Hronn Brynjarsdottir, Maria Håkansson, James Pierce, Eric Baumer, Carl DiSalvo, and Phoebe Sengers. 2012. Sustainably unpersuaded: how persuasion narrows our vision of sustainability. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 947–956.
[53] Tom Buchanan. 2020. Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation. Plos one 15, 10 (2020), e0239666.
[54] Tom Buchanan and Vladlena Benson. 2019. Spreading Disinformation on Facebook: Do Trust in Message Source, Risk Propensity, or Personality Affect the Organic Reach of "Fake News"? Social Media+ Society 5, 4 (2019), 2056305119888654.
[55] Diana Buitrago-Garcia, Dianne Egli-Gany, Michel J Counotte, Stefanie Hossmann, Hira Imeri, Aziz Mert Ipekci, Georgia Salanti, and Nicola Low. 2020. Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic review and meta-analysis. PLoS medicine 17, 9 (2020), e1003346.
[56] Cody Buntain, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2021. YouTube recommendations and effects on sharing across online social platforms. Proceedings of the ACM on Human-Computer Interaction 5, CSCW1 (2021), 1–26.
[57] Eugene Burnstein and Amiram Vinokur. 1973. Testing two classes of theories about group induced shifts in individual choice. Journal of Experimental Social Psychology 9, 2 (1973), 123–137.
[58] Leonardo Bursztyn, Georgy Egorov, and Stefano Fiorin. 2020. From extreme to mainstream: The erosion of social norms. American Economic Review 110, 11 (2020), 3522–48.
[59] Sahara Byrne and Philip Solomon Hart. 2009. The Boomerang Effect A Synthesis of Findings and a Preliminary Theoretical Framework. Annals of the International Communication Association 33, 1 (Jan. 2009), 3–37. https://doi.org/10.1080/23808985.2009.11679083
[60] Joseph N Cappella and Kathleen Hall Jamieson. 1994. Broadcast adwatch effects: A field experiment. Communication Research 21, 3 (1994), 342–365.
[61] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In Proceedings of the 20th international conference on World wide web. 675–684.
[62] Gizem Ceylan and Norbert Schwarz. 2020. Look What I Am Re-Sharing: How Self-Presentation Goals Impact What Consumers Spread on Social Networks. ACR North American Advances (2020).
[63] Man-pui Sally Chan, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. Psychological science 28, 11 (2017), 1531–1546.
[64] Emily Chen, Herbert Chang, Ashwin Rao, Kristina Lerman, Geoffrey Cowan, and Emilio Ferrara. 2021. COVID-19 misinformation and the 2020 US presidential election. The Harvard Kennedy School Misinformation Review (2021).

[65] Xinran Chen. 2016. The influences of personality and motivation on the sharing of misinformation on social media. IConference 2016 Proceedings (2016).

[66] Adrienne Chung and Rajiv N Rimal. 2016. Social norms: A review. Review of Communication Research 4 (2016), 1–28.

[67] Robert B Cialdini and Melanie R Trost. 1998. Social influence: Social norms, conformity and compliance. (1998).

[68] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 2020. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. Political Behavior 42, 4 (2020), 1073–1095.

[69] Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. 2011. Computational Journalism: A Call to Arms to Database Researchers. In 5th Biennial Conference on Innovative Data Systems Research (CIDR'll), ACM.

[70] Jonas Colliander. 2019. "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. Computers in Human Behavior 97 (2019), 202–215.

[71] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In Fifth international AAAI conference on weblogs and social media.

[72] Josh Constine. 2017. Facebook fights fake news with links to other angles. https://techcrunch.com/2017/08/03/facebook-related-articles/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAGkjbvnLbViPDaKtfsQZuRdpcmsVLSFLP9egNExbvgObqXa8_QOM_jXOWpCX6vtZXEnVS5tzwdqWZSpnzL6_M6w67dQj8U5bheaA93Ch_pqUia_EAcEIqfJcrGNJgS48eIAGdIoam0-nPoWIRfVQxIln5vHobK3wGAXy0xQ3AByx

[73] Paul T Costa Jr and Robert R McCrae. 2008. The Revised Neo Personality Inventory (neo-pi-r). Sage Publications, Inc.

[74] Joseph Cox and Jason Koebler. 2019. Facebook Bans White Nationalism and White Separatism. https://www.vice.com/en_us/article/nexpbx/facebook-bans-white-nationalism-and-white-separatism

[75] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. Proceedings of the National Academy of Sciences 113, 3 (2016), 554–559.

[76] Nicholas Dias, Gordon Pennycook, and David G Rand. 2020. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. Harvard Kennedy School Misinformation Review 1, 1 (2020).

[77] Nicholas Dias, Gordon Pennycook, and David G. Rand. 2021. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media: HKS Misinformation Review. https://misinforeview.hks.harvard.edu/article/emphasizing-publishers-does-not-reduce-misinformation/

[78] Nicholas DiFonzo, Martin J Bourgeois, Jerry Suls, Christopher Homan, Noah Stupak, Bernard P Brooks, David S Ross, and Prashant Bordia. 2013. Rumor clustering, consensus, and polarization: Dynamic social impact and self-organization of hearsay. Journal of Experimental Social Psychology 49, 3 (2013), 378–399.

[79] Carl DiSalvo, Phoebe Sengers, and Hrönn Brynjarsdóttir. 2010. Mapping the landscape of sustainable HCI. In Proceedings of the SIGCHI conference on human factors in computing systems. 1975–1984.

[80] Nicole Doughty, Helen M Paterson, Carolyn MacCann, and Lauren A Monds. 2017. Personality and memory conformity. Journal of Individual Differences (2017).

[81] Alice H Eagly and Shelly Chaiken. 1993. The psychology of attitudes. Harcourt brace Jovanovich college publishers.

[82] Adam M Enders, Joseph E Uscinski, Casey Klofstad, and Justin Stoler. 2020. The different forms of COVID-19 misinformation and their consequences. The Harvard Kennedy School Misinformation Review (2020).

[83] Ziv Epstein, Gordon Pennycook, and David 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. 1–11.

[84] Facebook. 2013. About Fact-Checking on Facebook. https://www.facebook.com/business/help/2593586717571940?id=673052479947730. Online; accessed 13 January 2022.

[85] Facebook. 2018. Removing Coordinated Inauthentic Behavior From China. https://about.fb.com/news/2019/08/removing-cib-china/. Online; accessed 13 January 2022.

[86] Facebook. 2021. How fact-checking works. https://transparency.fb.com/features/how-fact-checking-works/. Online; accessed 13 January 2022.

[87] Yuval Feldman and Janice Nadler. 2006. The law and norms of file sharing. San Diego L. Rev. 43 (2006), 577.

[88] DJ Flynn, Brendan Nyhan, and Jason Reifler. 2017. The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. Political Psychology 38 (2017), 127–150.

[89] Brian J Fogg, Jonathan Marshall, Othman Laraki, Alex Osipovich, Chris Varma, Nicholas Fang, Jyoti Paul, Akshay Rangnekar, John Shon, Preeti Swani, et al. 2001. What makes web sites credible? A report on a large quantitative

study. In Proceedings of the SIGCHI conference on Human factors in computing systems. 61–68.

[90] Brian J Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. 2003. How do users evaluate the credibility of Web sites? A study with over 2,500 participants. In Proceedings of the 2003 conference on Designing for user experiences. 1–15.

[91] Krisandra S Freeman and Jan H Spyridakis. 2004. An examination of factors that affect the credibility of online health information. Technical communication 51, 2 (2004), 239–263.

[92] Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor cascades. In Proceedings of the International AAAI Conference on Web and Social Media, Vol. 8.

[93] 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 (2018), 1–16.

[94] Megan Garber. 2012. The newest factchecker: Reddit. https://www.theatlantic.com/technology/archive/2012/10/the-newest-factchecker-reddit/263238/

[95] R Kelly Garrett. 2009. Echo chambers online?: Politically motivated selective exposure among Internet news users. Journal of computer-mediated communication 14, 2 (2009), 265–285.

[96] R. Kelly Garrett. 2009. Politically Motivated Reinforcement Seeking: Reframing the Selective Exposure Debate. Journal of Communication 59, 4 (2009), 676–699. https://doi.org/10.1111/j.1460-2466.2009.01452.x

[97] R Kelly Garrett and Shannon Poulsen. 2019. Flagging Facebook falsehoods: Self-identified humor warnings outperform fact checker and peer warnings. Journal of Computer-Mediated Communication 24, 5 (2019), 240–258.

[98] R Kelly Garrett and Brian E Weeks. 2013. The promise and peril of real-time corrections to political misperceptions. In Proceedings of the 2013 conference on Computer supported cooperative work. 1047–1058.

[99] Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate. In Proceedings of the 2020 CHI conference on human factors in computing systems. 1–14.

[100] Amira Ghenai and Yelena Mejova. 2018. Fake cures: user-centric modeling of health misinformation in social media. Proceedings of the ACM on human-computer interaction 2, CSCW (2018), 1–20.

[101] Henner Gimpel, Sebastian Heger, Christian Olenberger, and Lena Utz. 2021. The effectiveness of social norms in fighting fake news on social media. Journal of Management Information Systems 38, 1 (2021), 196–221.

[102] Maria J Grant and Andrew Booth. 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. Health information & libraries journal 26, 2 (2009), 91–108.

[103] James Grimmelmann. 2015. The virtues of moderation. Yale JL & Tech. 17 (2015), 42.

[104] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 US presidential election. Science 363, 6425 (2019), 374–378.

[105] Destin Groff, Ashley Sun, Anna E Ssentongo, Djibril M Ba, Nicholas Parsons, Govinda R Poudel, Alain Lekoubou, John S Oh, Jessica E Ericson, Paddy Ssentongo, et al. 2021. Short-term and long-term rates of postacute sequelae of SARS-CoV-2 infection: a systematic review. JAMA network open 4, 10 (2021), e2128568–e2128568.

[106] Felix Hamborg, Norman Meuschke, and Bela Gipp. 2017. Matrix-based news aggregation: exploring different news perspectives. In 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE, 1–10.

[107] Michael Hameleers and Toni GLA van der Meer. 2020. Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? Communication Research 47, 2 (2020), 227–250.

[108] Lawrence C Hamilton, Joel Hartter, and Kei Saito. 2015. Trust in scientists on climate change and vaccines. Sage Open 5, 3 (2015), 2158244015602752.

[109] Lynn Hasher, David Goldstein, and Thomas Toppino. 1977. Frequency and the conference of referential validity. Journal of verbal learning and verbal behavior 16, 1 (1977), 107–112.

[110] Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The quest to automate fact-checking. In Proceedings of the 2015 computation+ journalism symposium.

[111] Lucien Heitz, Krisztina Rozgonyi, Bojana Kostic, Deniz Wagner, and Julia Haas. 2021. AI in Content Curation and Media Pluralism. (2021).

[112] Alfred Hermida, Fred Fletcher, Darryl Korell, and Donna Logan. 2012. Share, like, recommend: Decoding the social media news consumer. Journalism studies 13, 5-6 (2012), 815–824.

[113] Benjamin D Horne, Dorit Nevo, Sibel Adali, Lydia Manikonda, and Clare Arrington. 2020. Tailoring heuristics and timing AI interventions for supporting news veracity assessments. Computers in Human Behavior Reports 2 (2020), 100043.

[114] Carl I Hovland and Walter Weiss. 1951. The influence of source credibility on communication effectiveness. Public opinion quarterly 15, 4 (1951), 635–650.

[115] Guanxiong Huang and Kang Li. 2016. The effect of anonymity on conformity to group norms in online contexts: A meta-analysis. International journal of communication 10 (2016), 398–415.

[116] Y Linlin Huang, Kate Starbird, Mania Orand, Stephanie A Stanek, and Heather T Pedersen. 2015. Connected through crisis: Emotional proximity and the spread of misinformation online. In Proceedings of the 18th ACM conference on computer supported cooperative work & social computing. 969–980.
[117] Robert Huckfeldt, Paul Allen Beck, Russell J Dalton, and Jeffrey Levine. 1995. Political environments, cohesive social groups, and the communication of public opinion. American Journal of Political Science (1995), 1025–1054.
[118] Robert Huckfeldt, Jeanette Morehouse Mendez, and Tracy Osborn. 2004. Disagreement, ambivalence, and engagement: The political consequences of heterogeneous networks. Political Psychology 25, 1 (2004), 65–95.
[119] Elle Hunt. 2017. Disputed by multiple fact-checkers': Facebook rolls out new alert to combat fake news. The Guardian 21 (2017).
[120] Juan-José Igartua and Lifen Cheng. 2009. Moderating effect of group cue while processing news on immigration: Is the framing effect a heuristic process? Journal of Communication 59, 4 (2009), 726–749.
[121] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized social signals: Computationally-derived social signals from account histories. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 1–12.
[122] H Innes and M Innes. 2021. De-platforming disinformation: conspiracy theories and their control. Information, Communication & Society (2021), 1–19.
[123] Mike Isaac. 2016. How facebook's fact-checking partnership will work. https://www.nytimes.com/2016/12/15/technology/facebook-fact-checking-fake-news.html
[124] 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 (2021), 1–42.
[125] Youngseung Jeon, Bogoan Kim, Aiping Xiong, Dongwon Lee, and Kyungsik Han. 2021. ChamberBreaker: Mitigating the Echo Chamber Effect and Supporting Information Hygiene through a Gamified Inoculation System. Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (2021), 1–26.
[126] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the effectiveness of deplatforming as a moderation strategy on twitter. Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (2021), 1–30.
[127] Shan Jiang and Christo Wilson. 2018. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. Proceedings of the ACM on Human-Computer Interaction 2, CSCW (2018), 1–23.
[128] Neil F Johnson, Nicolas Velásquez, Nicholas Johnson Restrepo, Rhys Leahy, Nicholas Gabriel, Sara El Oud, Minzhang Zheng, Pedro Manrique, Stefan Wuchty, and Yonatan Lupu. 2020. The online competition between pro-and anti-vaccination views. Nature 582, 7811 (2020), 230–233.
[129] S Mo Jones-Jang, Tara Mortensen, and Jingjing Liu. 2021. Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. American Behavioral Scientist 65, 2 (2021), 371–388.
[130] The Wall Street Journal. [n.d.]. Breaking news, business, Financial Economic News, World News and Video. https://www.wsj.com/
[131] Dan M Kahan. 2012. Ideology, motivated reasoning, and cognitive reflection: An experimental study. Judgment and Decision making 8 (2012), 407–24.
[132] Dan M Kahan. 2017. Misconceptions, misinformation, and the logic of identity-protective cognition. (2017).
[133] Joseph Kahne and Benjamin Bowyer. 2017. Educating for democracy in a partisan age: Confronting the challenges of motivated reasoning and misinformation. American Educational Research Journal 54, 1 (2017), 3–34.
[134] Aimée A Kane, Sara Kiesler, and Ruogu Kang. 2018. Inaccuracy Blindness in Collaboration Persists, even with an Evaluation Prompt. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 1–9.
[135] 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. 312–323.
[136] Harold H Kelley. 1955. Salience of membership and resistance to change of group-anchored attitudes. Human Relations 8, 3 (1955), 275–289.
[137] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 137–146.
[138] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In Proceedings of the eleventh ACM international conference on web search and data mining. 324–332.
[139] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J Wisniewski, and Munmun De Choudhury. 2021. A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms. Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (2021), 1–34.

[140] 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 (2020), 1–27.
[141] Baris Kirdemir, Joseph Kready, Esther Mead, Muhammad Nihal Hussain, Nitin Agarwal, and Donald Adjeroh. 2021. Assessing bias in YouTube's video recommendation algorithm in a cross-lingual and cross-topical context. In International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer, 71–80.
[142] Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. (2007).
[143] CJason Koebler. 2018. Deplatforming Works. vice.com/en/article/bjbp9d/do-social-media-bans-work
[144] Alex Zhi-Xiong Koo, Min-Hsin Su, SangWon Lee, So-Yun Ahn, and Hernando Rojas. 2021. What Motivates People to Correct Misinformation? Examining the Effects of Third-person Perceptions and Perceived Norms. Journal of Broadcasting & Electronic Media (2021), 1–24.
[145] Yong Ming Kow, Yubo Kou, Xitong Zhu, and Wang Hin Sy. 2019. "Just My Intuition": Awareness of Versus Acting on Political News Misinformation. In International Conference on Information. Springer, 469–480.
[146] Hamutal Kreiner and Eyal Gamliel. 2021. Framing fake news: Asymmetric attribute-framing bias for favorable and unfavorable outcomes. Journal of Experimental Psychology: Learning, Memory, and Cognition (2021).
[147] Ziva Kunda. 1990. The case for motivated reasoning. Psychological bulletin 108, 3 (1990), 480.
[148] Maria Knight Lapinski and Rajiv N Rimal. 2005. An explication of social norms. Communication theory 15, 2 (2005), 127–147.
[149] David Lazer, Matthew Baum, Nir Grinberg, Lisa Friedland, Kenneth Joseph, Will Hobbs, and Carolina Mattsson. 2017. Combating fake news: An agenda for research and action. (2017).
[150] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. Science 359, 6380 (2018), 1094–1096.
[151] Sune Lehmann and Yong-Yeol Ahn. 2018. Complex spreading phenomena in social systems. Springer.
[152] Kristina Lerman, Xiaoran Yan, and Xin-Zeng Wu. 2016. The" majority illusion" in social networks. PloS one 11, 2 (2016), e0147617.
[153] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. Psychological science in the public interest 13, 3 (2012), 106–131.
[154] Stephan Lewandowsky and Klaus Oberauer. 2016. Motivated rejection of science. Current Directions in Psychological Science 25, 4 (2016), 217–222.
[155] Stephan Lewandowsky and Sander Van Der Linden. 2021. Countering misinformation and fake news through inoculation and prebunking. European Review of Social Psychology 32, 2 (2021), 348–384.
[156] Björn Lindström, Simon Jangard, Ida Selbing, and Andreas Olsson. 2018. The role of a "common is moral" heuristic in the stability and change of moral norms. Journal of Experimental Psychology: General 147, 2 (2018), 228.
[157] Milton Lodge and Charles S Taber. 2013. The rationalizing voter. Cambridge University Press.
[158] Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nature human behaviour 5, 3 (2021), 337–348.
[159] Robert Luzsa and Susanne Mayr. 2021. False consensus in the echo chamber: Exposure to favorably biased social media news feeds leads to increased perception of public support for own opinions. Cyberpsychology: Journal of Psychosocial Research on Cyberspace 15, 1 (2021).
[160] Diane M Mackie. 1986. Social identification effects in group polarization. Journal of Personality and Social Psychology 50, 4 (1986), 720.
[161] Gary Marks and Norman Miller. 1987. Ten years of research on the false-consensus effect: An empirical and theoretical review. Psychological bulletin 102, 1 (1987), 72.
[162] Cameron Martel, Gordon Pennycook, and David G Rand. 2020. Reliance on emotion promotes belief in fake news. Cognitive research: principles and implications 5, 1 (2020), 1–20.
[163] Philipp K Masur, Dominic James DiFranzo, and Natalya Natalie Bazarova. 2021. Behavioral Contagion on Social Media: Effects of Social Norms, Design Interventions, and Critical Media Literacy on Self-Disclosure. (2021).
[164] Sandra C Matz, Michal Kosinski, Gideon Nave, and David J Stillwell. 2017. Psychological targeting as an effective approach to digital mass persuasion. Proceedings of the national academy of sciences 114, 48 (2017), 12714–12719.
[165] Melinda McClure Haughey, Martina Povolo, and Kate Starbird. 2022. Bridging Contextual and Methodological Gaps on the "Misinformation Beat": Insights from Journalist-Researcher Collaborations at Speed. In CHI Conference on Human Factors in Computing Systems. 1–15.

[166] D Harrison McKnight and Charles J Kacmar. 2007. Factors and effects of information credibility. In Proceedings of the ninth international conference on Electronic commerce. 423–432.
[167] Panagiotis Takas Metaxas, Samantha Finn, and Eni Mustafaraj. 2015. Using twittertrails. com to investigate rumor propagation. In Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing. 69–72.
[168] Miriam J Metzger and Andrew J Flanagin. 2013. Credibility and trust of information in online environments: The use of cognitive heuristics. Journal of pragmatics 59 (2013), 210–220.
[169] Nicholas Micallef, Mihai Avram, Filippo Menczer, and Sameer Patil. 2021. Fakey: A Game Intervention to Improve News Literacy on Social Media. Proceedings of the ACM on Human-Computer Interaction 5, CSCW1 (2021), 1–27.
[170] Dale T Miller and Cathy McFarland. 1991. When social comparison goes awry: The case of pluralistic ignorance. (1991).
[171] Randall K Minas, Robert F Potter, Alan R Dennis, Valerie Bartelt, and Soyoung Bae. 2014. Putting on the thinking cap: using NeuroIS to understand information processing biases in virtual teams. Journal of Management Information Systems 30, 4 (2014), 49–82.
[172] Martin Moore and Damian Tambini. 2018. Digital dominance: the power of Google, Amazon, Facebook, and Apple. Oxford University Press.
[173] Patricia Moravec, Randall Minas, and Alan R Dennis. 2018. Fake news on social media: People believe what they want to believe when it makes no sense at all. Kelley School of Business Research Paper 18-87 (2018).
[174] Mohsen Mosleh, Cameron Martel, Dean Eckles, and David Rand. 2021. Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment. In proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 1–13.
[175] Alexander Muscat and Jonathan Duckworth. 2018. WORLD4: Designing ambiguity for first-person exploration games. In Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play. 341–351.
[176] Ryosuke Nagura, Yohei Seki, Noriko Kando, and Masaki Aono. 2006. A method of rating the credibility of news documents on the web. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 683–684.
[177] An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. 2018. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology. 189–199.
[178] An T Nguyen, Aditya Kharosekar, Matthew Lease, and Byron Wallace. 2018. An interpretable joint graphical model for fact-checking from crowds. In Thirty-Second AAAI Conference on Artificial Intelligence.
[179] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. Review of general psychology 2, 2 (1998), 175–220.
[180] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. Political Behavior 32, 2 (2010), 303–330.
[181] Chitu Okoli. 2015. A guide to conducting a standalone systematic literature review. Communications of the Association for Information Systems 37, 1 (2015), 43.
[182] James M Olson and Mark P Zanna. 1979. A new look at selective exposure. Journal of Experimental Social Psychology 15, 1 (1979), 1–15.
[183] B. Ortutay. [n.d.]. Replacing Disputed Flags With Related Articles.
[184] Mathias Osmundsen, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Michael Bang Petersen. 2020. Partisan polarization is the primary psychological motivation behind “fake news” sharing on Twitter. (2020).
[185] 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, et al. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Systematic reviews 10, 1 (2021), 1–11.
[186] Eli Pariser. 2011. The filter bubble: How the new personalized web is changing what we read and how we think. Penguin.
[187] Namkee Park, Hyun Sook Oh, and Naewon Kang. 2012. Factors influencing intention to upload content on Wikipedia in South Korea: The effects of social norms and individual differences. Computers in Human Behavior 28, 3 (2012), 898–905.
[188] Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. Journal of experimental psychology: general 147, 12 (2018), 1865.
[189] 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 (2021), 590–595.
[190] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G 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 (2020), 770–780.
[191] Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. Proceedings of the National Academy of Sciences 116, 7 (2019), 2521–2526.
[192] Gordon Pennycook and David G Rand. 2020. Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. Journal of personality 88, 2 (2020), 185–200.
[193] Gordon Pennycook and David G Rand. 2021. Examining false beliefs about voter fraud in the wake of the 2020 Presidential Election. The Harvard Kennedy School Misinformation Review (2021).
[194] Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In Communication and persuasion. Springer, 1–24.
[195] Shruti Phadke, Mattia Samory, and Tanushree Mitra. 2021. What Makes People Join Conspiracy Communities? Role of Social Factors in Conspiracy Engagement. Proceedings of the ACM on Human-Computer Interaction 4, CSCW3 (2021), 1–30.
[196] Lara SG Piccolo, Alisson Puska, Roberto Pereira, and Tracie Farrell. 2020. Pathway to a Human-Values Based Approach to Tackle Misinformation Online. In International Conference on Human-Computer Interaction. Springer, 510–522.
[197] Sara Pluviano, Caroline Watt, and Sergio Della Sala. 2017. Misinformation lingers in memory: failure of three pro-vaccination strategies. PloS one 12, 7 (2017), e0181640.
[198] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In European Conference on Information Retrieval. Springer, 810–817.
[199] Hafizh A Prasetya and Tsuyoshi Murata. 2020. A model of opinion and propagation structure polarization in social media. Computational Social Networks 7, 1 (2020), 1–35.
[200] Deborah A Prentice and Dale T Miller. 1996. Pluralistic ignorance and the perpetuation of social norms by unwitting actors. In Advances in experimental social psychology. Vol. 28. Elsevier, 161–209.
[201] Neha Puri, Eric A Coomes, Hourmazd Haghbayan, and Keith Gunaratne. 2020. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. Human Vaccines & Immunotherapeutics (2020), 1–8.
[202] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. 2016. Echo chambers on Facebook. Available at SSRN 2795110 (2016).
[203] Ankita Rao. 2021. Guardian News and Media. (2021, January 5). US pharmacist who tried to ruin Covid vaccine doses is a conspiracy THEORIST, police say. https://www.theguardian.com/us-news/2021/jan/04/wisconsin-pharmacist-covid-19-vaccine-doses-steven-brandenburg.
[204] Thomson Reuters. [n.d.]. Partly false Claim: Vaccines contain toxic levels of ALUMINUM, polysorbate 80, yeast and other substances.
[205] Rajiv N Rimal and Maria K Lapinski. 2015. A re-explication of social norms, ten years later. Communication Theory 25, 4 (2015), 393–409.
[206] Rajiv N Rimal, Maria K Lapinski, Rachel J Cook, and Kevin Real. 2005. Moving toward a theory of normative influences: How perceived benefits and similarity moderate the impact of descriptive norms on behaviors. Journal of health communication 10, 5 (2005), 433–450.
[207] Rajiv N Rimal and Kevin Real. 2003. Understanding the influence of perceived norms on behaviors. Communication Theory 13, 2 (2003), 184–203.
[208] Jon Roozenbeek and Sander van der Linden. 2019. Fake news game confers psychological resistance against online misinformation. Palgrave Communications 5, 1 (2019), 1–10.
[209] Dietram A Scheufele and Nicole M Krause. 2019. Science audiences, misinformation, and fake news. Proceedings of the National Academy of Sciences 116, 16 (2019), 7662–7669.
[210] Julia Schwarz and Meredith Morris. 2011. Augmenting web pages and search results to support credibility assessment. In Proceedings of the SIGCHI conference on human factors in computing systems. 1245–1254.
[211] Norbert Schwarz. 2004. Metacognitive experiences in consumer judgment and decision making. Journal of Consumer Psychology 14, 4 (2004), 332–348.
[212] Norbert Schwarz, Lawrence J Sanna, Ian Skurnik, and Carolyn Yoon. 2007. Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. Advances in experimental social psychology 39 (2007), 127–161.
[213] Colleen M Seifert. 2002. The continued influence of misinformation in memory: What makes a correction effective? In Psychology of learning and motivation. Vol. 41. Elsevier, 265–292.
[214] Arloc Sherman, Danilo Trisi, Chad Stone, Shelby Gonzales, and Sharon Parrott. 2019. Immigrants Contribute Greatly to US Economy, Despite Administration's ZPublic Chargey Rule Rationale. JSTOR.
[215] Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. 2017. Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction. new media & society 19, 8 (2017), 1214–1235.

[216] Jieun Shin and Kjerstin Thorson. 2017. Partisan selective sharing: The biased diffusion of fact-checking messages on social media. Journal of Communication 67, 2 (2017), 233–255.

[217] Nathaniel Sirlin, Ziv Epstein, Antonio A Arechar, and David G Rand. 2021. Digital literacy is associated with more discerning accuracy judgments but not sharing intentions. Harvard Kennedy School Misinformation Review (2021).

[218] Willem WA Sleegers, Travis Proulx, and Ilja van Beest. 2019. Confirmation bias and misconceptions: Pupillometric evidence for a confirmation bias in misconceptions feedback. Biological psychology 145 (2019), 76–83.

[219] Steven Sloman, Steven A Sloman, and Philip Fernbach. 2018. The knowledge illusion: Why we never think alone. Penguin.

[220] Laura M Smith, Linhong Zhu, Kristina Lerman, and Zornitsa Kozareva. 2013. The role of social media in the discussion of controversial topics. In 2013 International Conference on Social Computing. IEEE, 236–243.

[221] Melodie Yun-Ju Song and Anatoliy Gruzd. 2017. Examining sentiments and popularity of pro-and anti-vaccination videos on YouTube. In Proceedings of the 8th international conference on social media & society. 1–8.

[222] Gregg Sparkman and Gregory M Walton. 2017. Dynamic norms promote sustainable behavior, even if it is counter-normative. Psychological science 28, 11 (2017), 1663–1674.

[223] Francesca Spezzano, Anu Shrestha, Jerry Alan Fails, and Brian W Stone. 2021. That's Fake News! Reliability of News When Provided Title, Image, Source Bias & Full Article. Proceedings of the ACM on Human-Computer Interaction 5, CSCW1 (2021), 1–19.

[224] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), 1–26.

[225] Jonathan Stray. 2017. Defense against the dark arts: networked propaganda and counter-propaganda. Tow Center for Digital Journalism. https://medium.com/tow-center/defense-against-the-darkartsnetworked-propaganda-and-counter-propaganda-deb7145aa76a (2017).

[226] Natalie Jomini Stroud. 2010. Polarization and partisan selective exposure. Journal of communication 60, 3 (2010), 556–576.

[227] Sharifa Sultana and Susan R Fussell. 2021. Dissemination, Situated Fact-checking, and Social Effects of Misinformation among Rural Bangladeshi Villagers During the COVID-19 Pandemic. Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (2021), 1–34.

[228] S Shyam Sundar. 1998. Effect of source attribution on perception of online news stories. Journalism & Mass Communication Quarterly 75, 1 (1998), 55–68.

[229] S Shyam Sundar, Silvia Knobloch-Westerwick, and Matthias R Hastall. 2007. News cues: Information scent and cognitive heuristics. Journal of the American society for information science and technology 58, 3 (2007), 366–378.

[230] Cass R Sunstein. 1999. The law of group polarization. University of Chicago Law School, John M. Olin Law & Economics Working Paper 91 (1999).

[231] Cass R Sunstein and Adrian Vermeule. 2009. Conspiracy theories: Causes and cures. Journal of Political Philosophy 17, 2 (2009), 202–227.

[232] Briony Swire, Adam J Berinsky, Stephan Lewandowsky, and Ullrich KH Ecker. 2017. Processing political misinformation: comprehending the Trump phenomenon. Royal Society open science 4, 3 (2017), 160802.

[233] Briony Swire, Ullrich KH Ecker, and Stephan Lewandowsky. 2017. The role of familiarity in correcting inaccurate information. Journal of experimental psychology: learning, memory, and cognition 43, 12 (2017), 1948.

[234] Charles S Taber and Milton Lodge. 2006. Motivated skepticism in the evaluation of political beliefs. American journal of political science 50, 3 (2006), 755–769.

[235] Samia Tasnim, Md Mahbub Hossain, and Hoimonty Mazumder. 2020. Impact of rumors and misinformation on COVID-19 in social media. Journal of preventive medicine and public health 53, 3 (2020), 171–174.

[236] The YouTube Team. 2019. Continuing our work to improve recommendations on YouTube. https://blog.youtube/news-and-events/continuing-our-work-to-improve/

[237] Hanneke A Teunissen, Renske Spijkerman, Mitchell J Prinstein, Geoffrey L Cohen, Rutger CME Engels, and Ron HJ Scholte. 2012. Adolescents' conformity to their peers' pro-alcohol and anti-alcohol norms: The power of popularity. Alcoholism: Clinical and experimental research 36, 7 (2012), 1257–1267.

[238] The Economic Time. 2020. Twitter: No more fake news: Twitter will label tweets that contain harmful, misleading content on coronavirus. https://economictimes.indiatimes.com/magazines/panache/no-more-fake-news-twitter-will-label-tweets-that-contain-harmful-misleading-content-on-coronavirus/articleshow/75688104.cms. Online; accessed 13 January 20201.

[239] The Washington Post. 2013. Fact Checker. https://www.washingtonpost.com/news/fact-checker/. Online; accessed 13 January 2022.

[240] Emily Thorson. 2015. Identifying and correcting policy misperceptions. Unpublished Paper, George Washington University. Available at http://www.americanpressinstitute.org/wp-content/uploads/2015/04/Project-2-Thorson-2015-Identifying-Political-Misperceptions-UPDATED-4-24.pdf (2015).
[241] Gleb Tsipursky and Zachary Morford. 2018. Addressing behaviors that lead to sharing fake news. Behavior and Social Issues 27 (2018), AA6–AA10.
[242] Twitter. 2021. Permanent suspension of @realDonaldTrump. https://blog.twitter.com/en_us/topics/company/2020/suspension.
[243] Sander van Der Linden, Jon Roozenbeek, and Josh Compton. 2020. Inoculating against fake news about COVID-19. Frontiers in psychology 11 (2020), 2928.
[244] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. 2009. On the evolution of user interaction in facebook. In Proceedings of the 2nd ACM workshop on Online social networks. 37–42.
[245] Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In Proceedings of the ACL 2014 workshop on language technologies and computational social science. 18–22.
[246] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. Science 359, 6380 (2018), 1146–1151.
[247] Claire Wardle, Hossein Derakhshan, et al. 2018. Thinking about 'information disorder': formats of misinformation, disinformation, and mal-information. Ireton, Cherilyn; Posetti, Julie. Journalism,'fake news'& disinformation. Paris: Unesco (2018), 43–54.
[248] Kimberlee Weaver, Stephen M Garcia, Norbert Schwarz, and Dale T Miller. 2007. Inferring the popularity of an opinion from its familiarity: A repetitive voice can sound like a chorus. Journal of personality and social psychology 92, 5 (2007), 821.
[249] Sam Wineburg and Sarah McGrew. 2017. Lateral reading: Reading less and learning more when evaluating digital information. (2017).
[250] Donghee Yvette Wohn, Casey Fiesler, Libby Hemphill, Munmun De Choudhury, and J Nathan Matias. 2017. How to handle online risks? Discussing content curation and moderation in social media. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. 1271–1276.
[251] Magdalena Wojcieszak, Stephan Winter, and Xudong Yu. 2020. Social norms and selectivity: Effects of norms of open-mindedness on content selection and affective polarization. Mass Communication and Society 23, 4 (2020), 455–483.
[252] Jules Woolf, Rajiv N Rimal, and Pooja Sripad. 2014. Understanding the influence of proximal networks on high school athletes' intentions to use androgenic anabolic steroids. Journal of Sport Management 28, 1 (2014), 8–20.
[253] Waheeb Yaqub, Otari Kakhidze, Morgan L Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of credibility indicators on social media news sharing intent. In Proceedings of the 2020 chi conference on human factors in computing systems. 1–14.
[254] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, et al. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In Companion Proceedings of the The Web Conference 2018. 603–612.
[255] Xichen Zhang and Ali A Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management 57, 2 (2020), 102025.
[256] Bi Zhu, Chuansheng Chen, Elizabeth F Loftus, Qinghua He, Chunhui Chen, Xuemei Lei, Chongde Lin, and Qi Dong. 2012. Brief exposure to misinformation can lead to long-term false memories. Applied Cognitive Psychology 26, 2 (2012), 301–307.
[257] Bi Zhu, Chuansheng Chen, Elizabeth F Loftus, Chongde Lin, Qinghua He, Chunhui Chen, He Li, Robert K Moyzis, Jared Lessard, and Qi Dong. 2010. Individual differences in false memory from misinformation: Personality characteristics and their interactions with cognitive abilities. Personality and Individual differences 48, 8 (2010), 889–894.
[258] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. PloS one 11, 3 (2016), e0150989.

Received January 2022; revised July 2022; accepted November 2022