

Mapping the Scientific Literature on Misinformation Interventions: A Bibliometric Review

Catherine King, Peter Carragher, Kathleen M. Carley

Carnegie Mellon University
Pittsburgh, PA 15213
cking2, pcarragh, carley@andrew.cmu.edu

Abstract

In recent years, social media misinformation has become a growing problem worldwide. Researchers in various fields have been investigating the most effective and acceptable ways to counter fake news online. This scoping review paper conducts a bibliometric analysis of relevant papers to provide policymakers and academics with a contextual background on the state of the literature in this field. The analysis reveals that several types of interventions are studied much more frequently than others, and the literature focuses almost exclusively on the effectiveness of interventions without considering the critical metric of user acceptance. Furthermore, publication venues in this field have primarily remained clustered by discipline, although several collaborative and multi-disciplinary publications have emerged in recent years. These findings highlight the need for increased collaboration and a greater focus on user-centered approaches in future research.

Zotero Repository — https://www.zotero.org/groups/5961522/misinformation_interventions

Social media platforms have allowed people and organizations to access and spread information faster than ever before (Vosoughi, Roy, and Aral 2018). They have also increased the speed and reach of misinformation, which has been shown to have pressing societal impacts ranging from undermining democracy (Tucker et al. 2018), increasing extremism (Warner and Neville-Shepard 2014), and lowering the uptake of public health measures (Oleksy et al. 2021).

A growing number of researchers have been investigating ways to counter misinformation. This research can be challenging, as many researchers need more direct access to social media data from social media platforms (Courchesne, Ilhardt, and Shapiro 2021). Even if data is available, there are ethical challenges associated with sharing social media data with other researchers (Bishop and Gray 2017).

A recent review article from Courchesne and colleagues found that certain types of platform interventions are over-studied relative to others and relative to the frequency with which platforms implement them (Courchesne, Ilhardt, and Shapiro 2021). Specifically, fact-checking and debunking are the most common interventions studied, but little re-

search investigates countermeasures that target creators and require direct data access from social media platforms.

There have been several other review articles in this field. Some, like Helmus and Kepe from the Rand Corporation, focus on related policy papers (Helmus and Kepe 2021). Others examine specific intervention categories, such as content moderation (Jiang et al. 2023) or media literacy (Jeong, Cho, and Hwang 2012). However, none have analyzed the broader picture, including both platform interventions and possible government policies. Additionally, most platform review articles focus on testing interventions and analyzing their effectiveness but fail to discuss the equally important metric of their practicality and acceptability to users. We seek to fill this gap by considering user-level, platform-level, and policy interventions. Through a bibliometric analysis of the citation network of relevant papers, we aim to gain insights into the misinformation intervention space. Our research questions are:

1. What journals and academic disciplines have published research on misinformation interventions over the last 20 years, and how has this evolved over time?
2. Which types or categories of interventions have been studied the most, and how has this changed over time?
3. What set of impacts have been researched? The primary impacts of an intervention include its effectiveness and level of user acceptance.

Intervention Categorization

Before initializing our review of the literature, we developed categories of interventions. There are several ways misinformation interventions have been categorized in the literature (Courchesne, Ilhardt, and Shapiro 2021; Gwiazdziński et al. 2023; Helmus and Kepe 2021; Yadav 2021).

We selected four of the most comprehensive review articles to analyze when developing this typology (Courchesne, Ilhardt, and Shapiro 2021; Aghajari, Baumer, and DiFranzo 2023; Blair et al. 2024; Kozyreva et al. 2024). These four articles were selected for their recency, the breadth of interventions covered, and their diverse disciplines. The Courchesne article primarily discusses platform interventions (Courchesne, Ilhardt, and Shapiro 2021) and was published in the *Harvard Misinformation Review*, an interdisciplinary journal mainly associated with social sciences. The Aghajari ar-

ticle was published in the proceedings of an HCI conference (Aghajari, Baumer, and DiFranzo 2023). The Blair article highlights research from both the Global North and the Global South, and it was published in a psychology venue (Blair et al. 2024). Finally, the Kozyreva article was published most recently with numerous high-profile scholars in this research area and appeared in the prestigious *Nature Human Behavior* journal (Kozyreva et al. 2024). Synthesizing the categorizations used by the four review articles and drawing from prior work (King 2025), Table 1 shows our proposed eight categories of interventions.

| Category | Example Interventions |
|------------------------|--|
| Content Distribution | accuracy prompts, friction, redirection |
| Content Moderation | algorithmic downranking, fact-checking, content removal |
| Account Moderation | account removal, shadow banning, demonetization |
| Content Labeling | warning labels, source credibility labels, context labels |
| User-based Measures | reporting users or posts, social corrections, social norms |
| Media Literacy / Edu. | lateral reading strategies, training games, inoculation |
| Institutional Measures | media support, data sharing, government regulation |
| Other | combining interventions, new interventions, generative AI |

Table 1: General misinformation intervention categories

Method: Literature Review

This section describes how papers were selected, the inclusion criteria, and the analysis plan. We conducted a scoping literature review, a type of systematic review that is broader in nature, to find insights into the misinformation intervention field and answer these research questions. We followed the methods used in two previous computer science review articles focusing on interventions (Aghajari, Baumer, and DiFranzo 2023; Zainudin et al. 2024). More specifically, we adhered to the modified PRISMA guidelines for scoping review (Tricco et al. 2018), as well as the more specific guidelines for systematic reviews in the information systems field developed by Okoli (Okoli 2015).

Review Protocol

Our literature search was conducted in two stages (Figure 1). First, we selected the four prominent review papers that were used to develop our typology of interventions as “seed papers” (Courchesne, Ilhardt, and Shapiro 2021; Aghajari, Baumer, and DiFranzo 2023; Blair et al. 2024; Kozyreva et al. 2024). We gathered all the studies analyzed by the seed papers, removed duplicates, applied the exclusion criteria, and labeled the papers according to the definitions provided in the appendix, resulting in 365 labeled papers.

After labeling the initial set of papers, we proceeded to the second stage of the literature search. Using Scopus, we

searched for any intervention papers published after these review papers (2024-2025) using keyword searches. Paper titles needed to include either the words “misinformation” or “disinformation” and either “intervention” or “counter*”. Additionally, we searched for any specific intervention that had been labeled in fewer than 10 of the papers in the initial set. Table 2 displays the under-studied interventions and the associated keywords used in our Scopus keyword search. All these keywords were used alongside the words “misinformation” or “disinformation” in the title and “intervention” or “counter*” in the title or abstract, limited to papers published after 2004. This search did not retrieve any papers for shadow banning and data sharing, prompting us to redo the same keyword search without requiring the words “intervention” or “counter*” to be included. After applying the exclusion criteria and labeling the newly added papers, this process resulted in 86 labeled papers by the end of Stage 2, bringing the total to 451 labeled papers in our dataset.

| Intervention Label | Scopus Keywords Used |
|------------------------|--|
| Context Labels (9) | (“context” AND “label*”) OR (“community” AND “note”) |
| Alg. Moderation (8) | “downranking” |
| Advertising policy (5) | “advertising” |
| Redirection (5) | “redirect*” |
| Media Support (3) | “local news*” OR “media support” |
| Account Removal (2) | “deplatform*” |
| Reporting (2) | “user” AND “reporting” |
| Shadow Banning (0) | “shadow*” AND “ban” |
| Blocking (0) | “blocking” |
| Data Sharing (0) | “data sharing” |
| Generative AI (0) | “gen* AI” OR “chatbot” |
| Gov Regulation (0) | (“government” AND “regulation”) OR (“government policy”) |

Table 2: The intervention labels used the least after Stage 1, sorted from highest to lowest according to the number of papers assigned to those labels.

Inclusion Criteria

We aimed to include as wide range of interventions and factors studied as possible. For a paper to be included in our analysis, it must have met the following criteria:

- Content:** One of the article’s main focuses should be interventions or countermeasures to misinformation. The paper does not need to address social media specifically but must primarily concentrate on interventions. The included articles could directly test the efficacy of one or more interventions through experimental studies or could be a review or discussion-based article.
- Article Type:** The article is a research article. It is not an opinion piece, research proposal, or simply an abstract.
- Venue:** The paper comes from a reputable venue or institution, but inclusion is not restricted to peer-reviewed publications only. To ensure quality, we enumerated the types of venues that could be included:

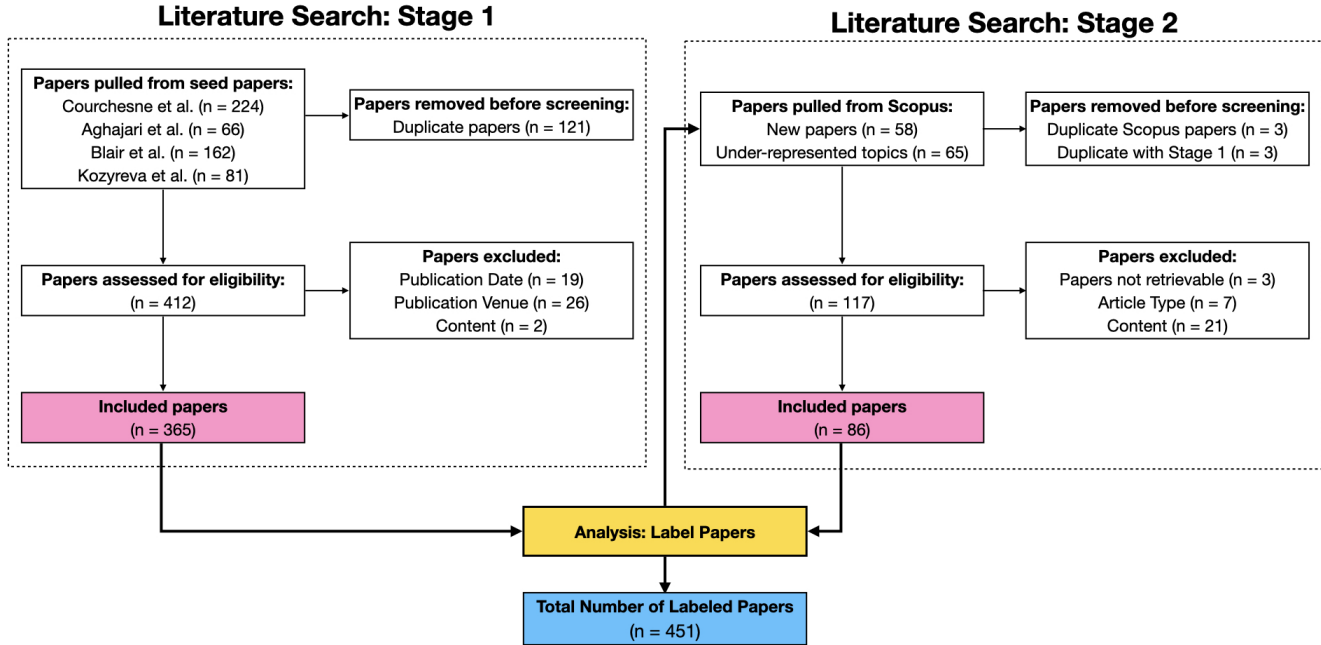


Figure 1: Literature review process.

- *Peer-reviewed Articles* - Papers from peer-reviewed indexed journals and conference proceedings, including workshop papers.
- *Technical Reports* - Reports from reputable and high-quality institutions like think tanks, non-profits, research centers, or governmental organizations.
- *Preprints* - Preprints posted in 2020 or later were included. Preprints are especially common in the fast-moving field of computer science. Preprints from before 2020 were excluded unless they had later been published in a peer-reviewed venue.

Websites, newspaper articles, blog posts, preprints older than 2020, or undergraduate theses were not included.

4. **Publication Date:** The paper was published in 2004 or later since we primarily want to focus on the last twenty years of research in the social media era.
5. **Language:** The paper was written in English or translated into English. This criteria was due to a limitation of our labelers.

Paper Labeling

We derived a comprehensive list of 35 possible countermeasures from these eight categories. In addition to the intervention descriptions, we assigned labels to review articles, meta-analyses, and papers examining intervention effectiveness or user acceptance. In the appendix, we describe all topic labels used in this article. It is important to note that

these labels are not mutually exclusive, as some papers can cover multiple interventions.

Approximately 25% of the included Stage 1 papers were randomly assigned to have two non-expert but trained human labelers (101 papers). The remaining papers (265) were assigned to one person and ChatGPT 4 as the second labeler. As experts in the field, two of the authors resolved disagreements between raters to determine the final assigned labels.

Given the varying number of labelers and potential labels per paper, Jaccard similarity was calculated for both inter-rater reliability metrics and agreement with the final labels. This calculation corresponds to a straightforward percentage agreement for effectiveness and acceptance labeling. Inter-rater similarity was 0.956 for effectiveness labels, 0.855 for acceptance labels, and 0.545 for intervention labels. Table 3 displays Jaccard similarity agreement with the final label categorized by labeler type (ChatGPT, Expert, Rater).

| Rater | Papers | Effectiveness | Acceptance | Topics |
|-----------|--------|---------------|------------|--------|
| Rater (3) | 390 | 0.972 | 0.921 | 0.667 |
| ChatGPT | 264 | 0.962 | 0.864 | 0.746 |
| Expert | 76 | 0.987 | 0.974 | 0.938 |

Table 3: Jaccard Similarity agreement with the final labels broken up by rater type. Rater types are sorted by total number of papers labeled.

Network Analysis

We analyzed descriptive statistics of the data and also analyzed the network using four node sets: Article, Author, Venue, and Topic (Label). We examined the Co-Topics network (Topic x Topic), the Co-Authorship Publication Network (Publication Venue x Publication Venue), and the Co-Authorship Network (Author x Author).

Results and Analysis

In this section, we use descriptive statistics and network analysis to report on the growth of the field over time, as well as to examine leading venues, topics, and authors.

Number of Articles Per Year

First, we analyze the number of papers in this review published per year. There is an exponential growth of articles in this field. The spike in 2021 and the subsequent slight drop-off can likely be attributed to the fact that our seed paper that reviewed the most articles originated from a 2021 review article (Courchesne, Ilhardt, and Shapiro 2021).

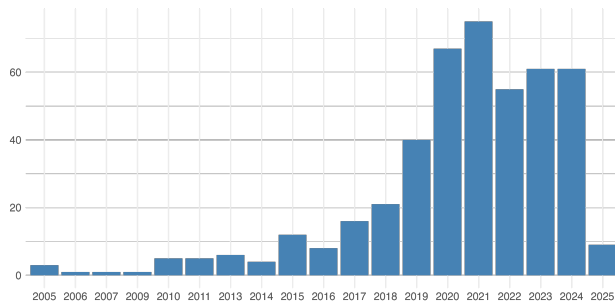


Figure 2: Number of reviewed papers published per year.

Publication Venue Analysis

In this section, we analyze the top publication venues in the dataset and visualize the Co-Publication Venue network.

Top Venues Table 4 shows the top 10 publication venues in the data set by two metrics. The total degree centrality for each venue indicates how many other venues it is connected to in the co-publication venue network. It represents venues that can be considered “central locations,” with the most connections to other venues, via authors who have published at both venues (Carley 2017). The second ranks venues by how many publications they have in this review.

Table 4 demonstrates that having many publications does not necessarily ensure a top ranking for that venue based on centrality metrics. Several venues, including *Scientific Reports* and the *CHI Conference*, do not rank highly in terms of degree centrality within the Co-Publication Venue network.

Co-Publication Venue Network In the Co-Publication Venue network, a link exists between two venues if an author in the dataset has published in both venues, and the links are weighted. Figure 3 illustrates the largest component. Nodes are sized by total degree centrality and colored by discipline.

Discipline information for all 209 venues was sourced from Scimago’s journal rank data from 2023¹.

The density of this network is low, at 0.035. Of the 209 publication venues in our dataset, only about half (114) belong to the largest component. This finding indicates a degree of disjointedness in the literature in this area. *Nature Human Behavior*, *Science Advances*, and *Harvard Misinformation Review* appear highly central in the network, underscoring their high rankings in total degree centrality. The centrality of these venues suggests that they are relatively interdisciplinary journals connecting various fields and authors who typically publish in other journal disciplines.

Additionally, the top-left side of the network mainly consists of psychology journals, highlighted in yellow. We find the social science journals on the right side of the network. The top-right comprises communication and journalism venues, while the bottom-right features many political science venues. Finally, on the bottom-left, we have the computer science venues. Although these fields are connected through several interdisciplinary journals, the venues within each discipline are clustered together and primarily linked to one another.

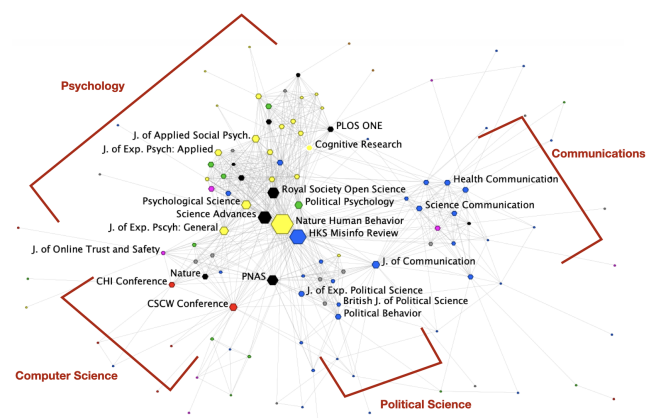


Figure 3: The Publication Venue x Publication Venue network. Nodes are sized by total degree centrality and colored by discipline (red for CS, yellow for psychology, blue for social sciences, black for multidisciplinary, and grey for other.)

Discipline Figure 4 shows the number of papers included in this review, categorized by discipline. The publication venue determines disciplines, and since venues can be affiliated with multiple disciplines, papers may also belong to more than one discipline as well. We observe that the social sciences initiate the literature in this area, with computer science and psychology competing for second. The “other” discipline, which represents all remaining fields from medicine to environmental science, shows a notable spike near the end of the timeline, indicating a growing interest in researching misinformation interventions across various domains.

¹<https://www.scimagojr.com/journalrank.php>

| Rank | Centrality | # Publications |
|------|---|---|
| 1. | Nature Human Behaviour | CSCW Conference (18) |
| 2. | Harvard Misinformation Review | Harvard Misinfo Review (15) |
| 3. | Science Advances | PLOS ONE (11) |
| 4. | Royal Society Open Science | CHI Conference (9) |
| 5. | PNAS | Science Communication (9) |
| 6. | Psychological Science | Scientific Reports (8) |
| 7. | J. of Experimental Psych: General | Journal of Communication (7) |
| 8. | Journal of Communication | Nature Human Behaviour (7) |
| 9. | CSCW; J. of Applied Social Psych.; Political Psych. | Cognitive Research: Principles and Implications (6) |
| 10. | Cognitive Research: Principles and Implications | Health Communication (6) |

Table 4: The top venues by degree centrality and total number of publications in the data set.

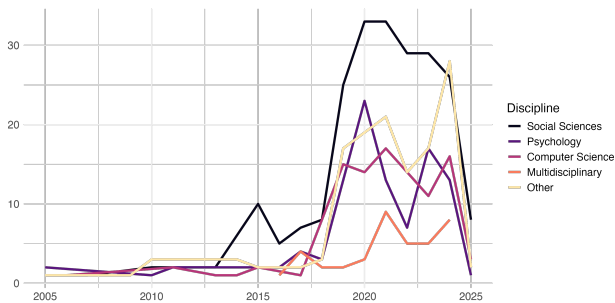


Figure 4: Misinformation is an interdisciplinary field, with research across the computational and social sciences.

Topic Analysis

In this section, we analyze the top topics, investigating whether some interventions are studied more than others, and finding topics that are often studied together. This analysis addresses both the first and second research questions.

Papers per Topic First, we calculate the descriptive statistics for the number of papers assigned to each label. Table 5 in the appendix displays the number of papers and unique authors for each intervention label. This table highlights how certain interventions are studied significantly more frequently than others. Fact-checking appears in 127 papers, while 15 of the 35 interventions are in fewer than 10 papers.

Fact-checking, debunking, and media literacy are outliers in terms of the number of papers examining those topics based on the calculated interquartile range. Similarly, those three intervention types, along with inoculation, are outliers regarding the total number of authors researching those topics. Furthermore, 404 papers (89.6%) analyzed an intervention’s effectiveness, whereas only 40 papers analyzed an intervention’s level of user acceptance (8.9%).

We analyze whether the types of interventions studied have changed over time. Figure 5 shows the number of papers included in this review, categorized by general intervention category studied. Papers could be assigned to more than one intervention label. Content moderation interventions were among the first and most prominently studied in the literature. Fact-checking and debunking are interven-

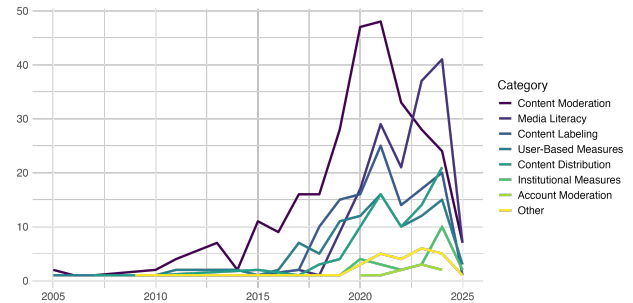


Figure 5: While research historically focused on content moderation, other intervention approaches are catching up.

tions that can be examined without access to social media data and would not be affected by a lack of platform transparency. Media literacy has especially taken off in the last few years, perhaps being studied in a broader range of journal types such as education, medicine, and others. In addition to media literacy, content distribution and institutional measures have recently reached their peak.

Co-Topics Network We next analyzed the *Topic x Topic* (*Co-Topic*) network, which shows the intervention types that are frequently studied together. A link exists between two topics in the network if a paper discusses both topics and the links are weighted. The network density was 0.40, indicating that countermeasures are frequently studied jointly with other countermeasures. Figure 6 shows the Co-Topic network, with nodes sized by total degree centrality and colored based on paper counts. The dark blue nodes represent topics with low paper counts; the lighter the blue, the more papers study that topic.

As shown in Figure 6, many topics, such as fact-checking and media literacy, are highly central to the network. Not only are they among the most studied interventions, but they are also often studied in conjunction with other interventions. Many countermeasures frequently employed by social media platforms, such as redirection, user-based countermeasures, and intervention combinations, remain relatively understudied (Courchesne, Ilhardt, and Shapiro 2021).

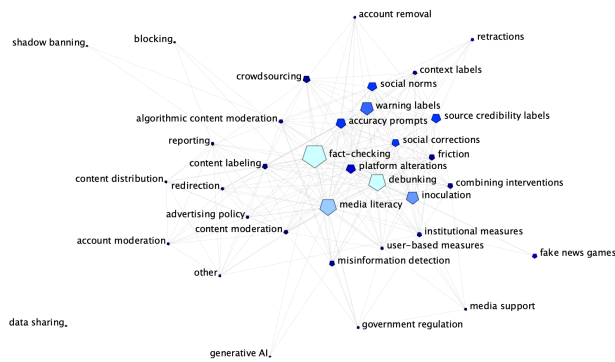


Figure 6: Co-Topic network. Nodes are sized by total degree centrality and colored by paper count. The lighter the blue, the more papers that study that topic.

Author Analysis

In this section, we analyze the top authors in the data and visualize the Co-Authorship network.

Co-Authorship Network Next, we visualize the *Author x Author (Co-Authorship)* network in Figure 7. In the Co-Authorship (Author x Author) network, a link exists between the two authors if those authors have worked together at least once, and the links are weighted by the number of times those authors have published together.

The density of this network is low at 0.005, which is expected considering the large number of authors in the dataset (1,147). However, only 326 authors are part of the largest component. In total, there were 218 components, with all other authors in either isolates, dyads, triads, or other relatively small groups. Eight components ranged in size from 10 to 21 authors, likely indicating research groups and authors who have not published outside their own group. This result shows that 821 authors not in the largest component (71.6% of all authors in our dataset) have primarily published on interventions within their own research groups and have not collaborated with others.

Of the 451 papers, 15 have ten or more authors, including one with 30 authors. This article with 30 authors was one of the seed papers (Kozyreva et al. 2024), indicating it was likely a review article written through the consensus of many leading authors in the field. It features six of the top ten authors by total number of citations. Removing this single paper from the analysis of the Co-Authorship network causes the largest component to split into two: one of size 230 and another of size 96, as illustrated in Figure 8 in the appendix. This finding suggests that many top authors are only connected in the dataset through this one recent review paper. This result suggests that the disjointedness in the literature may have decreased in recent years.

Discussion

We conducted a descriptive and bibliometric analysis of the citation network of prominent papers in the countermeasures literature. The number of articles published in the misinformation intervention space has increased dramatically, indi-

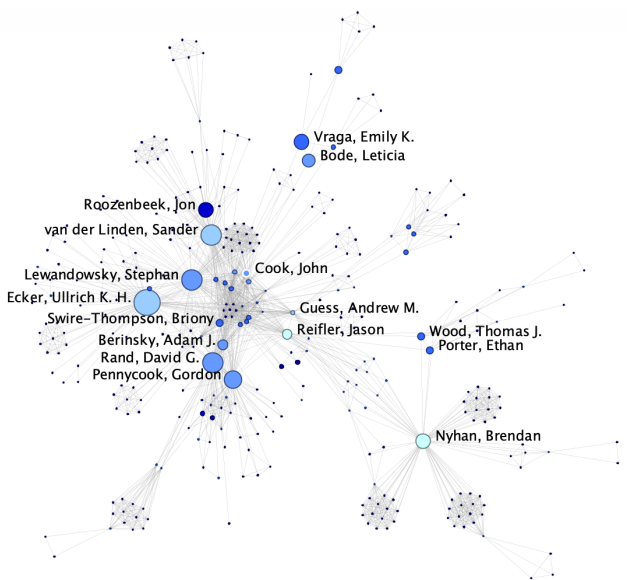


Figure 7: The largest component of the Author x Author network. Nodes are sized by total degree centrality and colored by betweenness (the lighter the blue, the higher the betweenness score). Authors with both high total degree centrality and high betweenness are labeled.

cating that this is a growing field of literature. First, our analysis of the Co-Publication Venue network reveals disjointedness in the literature, with most venues clustered near others within the same discipline. While journals such as the *Harvard Misinformation Review* and *Nature Human Behavior* bridge the gaps between related fields, nearly half of the venues were not part of the largest component. Multiple fields are conducting research in this area and are visible in their own clusters on the network, including Psychology, Political Science, Communications, and Computer Science. This research area is highly interdisciplinary.

Next, we analyzed the top topics studied in the literature. User acceptance is largely overlooked, with only about 9% of papers exploring this aspect. Acceptance is as important a measure as effectiveness. Without acceptance, platforms may hesitate to implement changes for fear of losing users, and governments might struggle to enact beneficial policies (Liu, Yildirim, and Zhang 2022). It is also notable that many studies on intervention effectiveness employ survey instruments which could be easily extended to measure user acceptance (Badrinathan 2021).

Furthermore, similar to a previous review article (Courchesne, Ilhard, and Shapiro 2021), we find that several critical, frequently used, or highly impactful interventions are underexplored in the literature. These include redirection, user-based countermeasures like user reporting and blocking, institutional measures like media support and data sharing, and emerging interventions involving generative AI. A 2021 review of platform policies indicates that redirection is the most prevalent intervention employed by platforms (Yadav 2021). However, there were only five papers related to

redirection in this list of 451 articles. Additionally, institutional measures overall, including analyses of potential government regulations or actions that civic society can take, are underrepresented in the literature compared to individual or platform-based interventions. Although there is a prominent RAND article that reviews countermeasures based on policy reports (Helmus and Kepe 2021), it was not used as a seed paper because this article exclusively evaluated policy reports published by think tanks, non-profits, and government entities and did not include any articles from traditional, peer-reviewed journals.

Underrepresented interventions most relevant to platform policies include demonetization strategies. While platforms do actively demonetize accounts that spread misinformation², a lack of empirical research on this issue is worrying given the financial incentives behind misinformation such as running ads and selling merchandise (Papadogiannakis et al. 2023; Martínez-López, Li, and Young 2022). Only one study addressed interventions that target the viability of, or ‘cost’ of, propagating misinformation on social media (Im et al. 2020). As monetization involves advertising and payment systems, interventions targeting these systems will require a greater level of platform access than is currently available to researchers. For this to happen, closer collaboration between academic institutions and social media platforms is needed.

The practicality of studies on misinformation interventions is undermined when there is a lack of consensus on intervention effectiveness, and several disagreements on the effectiveness of certain interventions were encountered while analyzing the literature. For example, among the most studied countermeasures, we find several sources of contention; a body of work claims the effectiveness of the “Bad News” game for inoculation (Roozenbeek, Linden, and Nygren 2020; Basol, Roozenbeek, and van der Linden 2020), while a meta-review finds their results to be insignificant using pre & post treatment classification accuracy (Modirrousta-Galian and Higham 2023). The effectiveness of source credibility interventions has also been disputed (Bruns et al. 2024). Finally, multiple studies find that media literacy is effective in some countries but not others, and that different types of media literacy are effective in different locations (Badrinathan 2021; Guess et al. 2020). The lack of consensus on these interventions underscores the need for comprehensive evaluation metrics in the field (van der Linden et al. 2021) and highlights the significance of meta-reviews. In order for these interventions to be successfully operationalized on online platforms, these disagreements should be resolved.

Finally, there is a growing divide between the types of interventions that garner popular support from either platforms or their users, and the interventions that appear to be most effective in the literature. For example, suspending or deplatforming user accounts has been shown to be effective in reducing the spread of harmful misinformation

(Thomas and Wahedi 2023; McCabe et al. 2024). However, these interventions are among the least popular with users (Kozyreva et al. 2023; King 2025). Additionally, some platforms, including Meta (Kaplan 2025), have discontinued fact-checking programs, despite their proven effectiveness (Walter and Murphy 2018; Blair et al. 2024). These platforms are instead endorsing Community Notes systems, a variant of contextual labeling, despite recent studies indicating that such systems have mixed efficacy (Blair et al. 2024) or no effect on misinformation engagement (Chuai et al. 2024). This dichotomy between social media platforms and academic research highlights a tradeoff between effectiveness and user acceptance. Therefore, our key finding that user acceptance is largely understudied, poses a serious issue for the practicality of research on misinformation interventions and their implementation in online social media. Future research on misinformation interventions should aim to further investigate and, if possible, reconcile the apparent divide between user acceptance and effectiveness.

Limitations

Our method for selecting papers may have missed articles in specific sub-areas. For example, while the literature tends to focus on social media, recent research has found that search engines play a significant role in the perceived veracity of misinformation (Aslett et al. 2024), motivating the development of interventions that target source credibility in search rankings (Carragher, Williams, and Carley 2025). While out of scope of our review, the intersection of search engines and misinformation is a promising direction for future work.

Additionally, we primarily focused on research conducted in academic peer-reviewed venues. Due to limited data-sharing and access, there may be some discrepancies between what is done in academia and what is done in industry or other institutions. Also only articles written in English were included in our analysis. However, the categories of misinformation we identified as over-studied and under-studied are similar to what other related review articles found, mitigating the risk of missing articles (Courchesne, Ilhardt, and Shapiro 2021).

Finally, we did not have a baseline to compare this bibliometric analysis against. For example, is it common for co-authorship networks to fragment as much as they did in this work when removing one article with a high number of authors (see Figure 7 and 8)? Is the clustering of journals by discipline common in other multidisciplinary fields? These questions pose possible future directions for this work.

Conclusion

A bibliometric analysis was conducted on the literature surrounding user-, platform-, and policy-level misinformation interventions. This analysis found many under- and over-studied interventions in the literature. Additionally, we found that the academic literature primarily focuses on the effectiveness of countermeasures without addressing the critical metric of user acceptance. The intervention space should investigate the popularity of these interventions. If user acceptance is low, implementing that intervention is unlikely, making its relative effectiveness less relevant.

²<https://www.youtube.com/howyoutubeworks/our-commitments/fighting-misinformation/>

³<https://www.tiktok.com/transparency/en-us/combating-misinformation/>

The analysis also uncovered structural issues in the research ecosystem. Publication venues in this field have primarily remained siloed by discipline, although several collaborative and multidisciplinary publications have emerged in recent years. Despite this, the field is fragmented: there is limited integration across intervention types, inconsistent findings on effectiveness, and a lack of cohesion in the authorship network. These gaps point to a need for both conceptual and methodological consolidation. Future work must rigorously evaluate interventions not only in terms of efficacy but also acceptance and scalability. Only by bridging disciplinary divides and focusing on the human factors that mediate intervention success can the field mature. This review is intended as a step in that direction.

Acknowledgements

This work was supported by the Knight Foundation, the Office of Naval Research's MURI: Persuasion, Identity, & Morality in Social-Cyber Environments grant N00014-21-12749, the Center for Computational Analysis of Social and Organizational Systems (CASOS), and the Center for Informed Democracy and Social-cybersecurity (IDeas). The views and conclusions contained in this document are those of the authors alone. The funders have no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors would like to thank Carter Thompson, Joelle Cheeseman, and Kaitlyn Zhang for their assistance in labeling the papers.

References

Aghajari, Z.; Baumer, E. P. S.; and DiFranzo, D. 2023. Reviewing Interventions to Address Misinformation: The Need to Expand Our Vision Beyond an Individualistic Focus. In *Proceedings of the ACM on Human-Computer Interaction*, volume 7 of CSCW, 87:1–87:34. ACM.

Allen, J.; Martel, C.; and Rand, D. G. 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. New Orleans LA: ACM.

Arif, A.; Robinson, J. J.; Stanek, S. A.; Fichet, E. S.; Townsend, P.; Worku, Z.; and Starbird, K. 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*, CSCW, 155–168. New York, NY, USA: ACM.

Aslett, K.; Sanderson, Z.; Godel, W.; Persily, N.; Nagler, J.; and Tucker, J. A. 2024. Online searches to evaluate misinformation can increase its perceived veracity. *Nature*, 625(7995): 548–556.

Assenmacher, D.; Weber, D.; Preuss, M.; Calero Valdez, A.; Bradshaw, A.; Ross, B.; Cresci, S.; Trautmann, H.; Neumann, F.; and Grimme, C. 2022. Benchmarking Crisis in Social Media Analytics: A Solution for the Data-Sharing Problem. *Social Science Computer Review*, 40(6): 1496–1522.

Badrinathan, S. 2021. Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India. *American Political Science Review*, 115(4): 1325–1341.

Badrinathan, S.; and Chauchard, S. 2023. "I Don't Think That's True, Bro!" Social Corrections of Misinformation in India. *The International Journal of Press/Politics*, 394–416.

Bago, B.; Rand, D. G.; and Pennycook, G. 2020. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, 149(8): 1608–1613.

Bak-Coleman, J. B.; Kennedy, I.; Wack, M.; Beers, A.; Schafer, J. S.; Spiro, E. S.; Starbird, K.; and West, J. D. 2022. Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, 1–9.

Basol, M.; Roozenbeek, J.; and van der Linden, S. 2020. Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News. *Journal of Cognition*, 3(1): 1–9.

Bishop, L.; and Gray, D. 2017. Ethical Challenges of Publishing and Sharing Social Media Research Data. In *The Ethics of Online Research*, volume 2, 159–187. Emerald Publishing Limited.

Blair, R. A.; Gottlieb, J.; Nyhan, B.; Paler, L.; Argote, P.; and Stainfield, C. J. 2024. Interventions to counter misinformation: Lessons from the Global North and applications to the Global South. *Current Opinion in Psychology*, 55: 101732.

Bode, L.; and Vraga, E. K. 2018. See Something, Say Something: Correction of Global Health Misinformation on Social Media. *Health Communication*, 33(9): 1131–1140.

Bradshaw, S.; and Neudert, L.-M. 2021. The Road Ahead: Mapping Civil Society Responses to Disinformation. Technical report, National Endowment for Democracy.

Bruns, H.; Dessart, F. J.; Krawczyk, M.; Lewandowsky, S.; Pantazi, M.; Pennycook, G.; Schmid, P.; and Smillie, L. 2024. Investigating the role of source and source trust in prebunks and debunks of misinformation in online experiments across four EU countries. *Scientific Reports*, 14(1): 20723.

Carley, K. M. 2017. ORA: A Toolkit for Dynamic Network Analysis and Visualization. In Alhajj, R.; and Rokne, J., eds., *Encyclopedia of Social Network Analysis and Mining*, 1–10. New York, NY: Springer. ISBN 978-1-4614-7163-9.

Carragher, P.; Williams, E. M.; and Carley, K. M. 2025. Misinformation resilient search rankings with webgraph-based interventions. *ACM Transactions on Intelligent Systems and Technology*, 16(1): 1–27.

Chuai, Y.; Tian, H.; Pröllochs, N.; and Lenzini, G. 2024. Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter? *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2): 1–52.

Costello, T. H.; Pennycook, G.; and Rand, D. G. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714): eadq1814.

Courchesne, L.; Ilhardt, J.; and Shapiro, J. N. 2021. Review of social science research on the impact of countermeasures against influence operations. *Harvard Kennedy School Misinformation Review*, 2(5).

- Ecker, U. K. H.; and Antonio, L. M. 2021. Can you believe it? An investigation into the impact of retraction source credibility on the continued influence effect. *Memory & Cognition*, 49(4): 631–644.
- Ecker, U. K. H.; Sanderson, J. A.; McIlhiney, P.; Rowsell, J. J.; Quekett, H. L.; Brown, G. D.; and Lewandowsky, S. 2023. Combining refutations and social norms increases belief change. *Quarterly Journal of Experimental Psychology (2006)*, 76(6): 1275–1297.
- Epstein, Z.; Berinsky, A. J.; Cole, R.; Gully, A.; Pennycook, G.; and Rand, D. G. 2021. Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School Misinformation Review*.
- Gao, M.; Xiao, Z.; Karahalios, K.; and Fu, W.-T. 2018. To Label or Not to Label: The Effect of Stance and Credibility Labels on Readers' Selection and Perception of News Articles. In *Proceedings of the ACM on Human-Computer Interaction*, volume 2 of CSCW, 55:1–55:16.
- Gillespie, T. 2022. Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, 8(3): 20563051221117552.
- Gimpel, H.; Heger, S.; Olenberger, C.; and Utz, L. 2021. The Effectiveness of Social Norms in Fighting Fake News on Social Media. *Journal of Management Information Systems*, 38(1): 196–221.
- Gorwa, R.; Binns, R.; and Katzenbach, C. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1).
- Guess, A. M.; Lerner, M.; Lyons, B.; Montgomery, J. M.; Nyhan, B.; Reifler, J.; and Sircar, N. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, 117(27): 15536–15545.
- Gwiazdźński, P.; Gundersen, A. B.; Piksa, M.; Krysińska, I.; Kunst, J. R.; Noworyta, K.; Olejniuk, A.; Morzy, M.; Rygula, R.; Wójtowicz, T.; and Piasecki, J. 2023. Psychological interventions countering misinformation in social media: A scoping review. *Frontiers in Psychiatry*, 13.
- Helmus, T. C.; and Chandra, B. 2024. Generative Artificial Intelligence Threats to Information Integrity and Potential Policy Responses. Technical report, RAND Corp.
- Helmus, T. C.; and Kepe, M. 2021. A Compendium of Recommendations for Countering Russian and Other State-Sponsored Propaganda. Technical report, RAND Corp.
- Im, J.; Tandon, S.; Chandrasekharan, E.; Denby, T.; and Gilbert, E. 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.
- Jeong, S.-H.; Cho, H.; and Hwang, Y. 2012. Media Literacy Interventions: A Meta-Analytic Review. *The Journal of Communication*, 62(3): 454–472.
- Jhaver, S.; and Zhang, A. X. 2023. Do users want platform moderation or individual control? Examining the role of third-person effects and free speech support in shaping moderation preferences. *New Media & Society*.
- Jiang, J. A.; Nie, P.; Brubaker, J. R.; and Fiesler, C. 2023. A Trade-off-centered Framework of Content Moderation. *ACM Transactions on Computer-Human Interaction*, 30(1): 3:1–3:34.
- Johns, A.; Bailo, F.; Booth, E.; and Rizoiu, M.-A. 2024. Labelling, shadow bans and community resistance: did Meta's strategy to suppress rather than remove COVID misinformation and conspiracy theory on Facebook slow the spread? *Media International Australia*, 1329878X241236984.
- Jones-Jang, S. M.; Mortensen, T.; and Liu, J. 2021. Does Media Literacy Help Identification of Fake News? Information Literacy Helps, but Other Literacies Don't. *American Behavioral Scientist*, 65(2): 371–388.
- Kaplan, J. 2025. More Speech and Fewer Mistakes. <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>. Accessed: 2025-03-04.
- Katsaros, M.; Yang, K.; and Fratamico, L. 2022. Reconsidering Tweets: Intervening during Tweet Creation Decreases Offensive Content. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16 of ICWSM, 477–487. AAAI.
- Khan, T.; Michalas, A.; and Akhunzada, A. 2021. Fake news outbreak 2021: Can we stop the viral spread? *Journal of Network and Computer Applications*, 190: 103112.
- King, C. 2025. *Effective and Practical Strategies for Combatting Misinformation*. Ph.D. diss., Software and Societal Systems Dept., Carnegie Mellon Univ., Pittsburgh, PA.
- Kirchner, J.; and Reuter, C. 2020. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–27.
- Kozyreva, A.; Herzog, S. M.; Lewandowsky, S.; Hertwig, R.; Lorenz-Spreen, P.; Leiser, M.; and Reifler, J. 2023. Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, 120(7): e2210666120.
- Kozyreva, A.; Lorenz-Spreen, P.; Herzog, S. M.; Ecker, U. K. H.; Lewandowsky, S.; Hertwig, R.; Ali, A.; Bak-Coleman, J.; Barzilai, S.; Basol, M.; Berinsky, A. J.; Betsch, C.; Cook, J.; Fazio, L. K.; Geers, M.; Guess, A. M.; Huang, H.; Larreguy, H.; Maertens, R.; Panizza, F.; Pennycook, G.; Rand, D. G.; Rathje, S.; Reifler, J.; Schmid, P.; Smith, M. D.; Swire-Thompson, B.; Szewach, P.; van der Linden, S.; and Wineburg, S. 2024. Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, 1–9.
- Lewandowsky, S.; Cook, J.; Ecker, U.; Albarraçín, D.; Amazeen, M. A.; Kendeou, P.; Lombardi, D.; Newman, E. J.; Pennycook, G.; Porter, E.; Rand, D. G.; Rapp, D. N.; Reifler, J.; Roozenbeek, J.; Schmid, P.; Seifert, C. M.; Sinatra, G. M.; Swire-Thompson, B.; van der Linden, S.; Vraga, E. K.; Wood, T. J.; and Zaragoza, M. S. 2020. Debunking Handbook 2020.
- Lewandowsky, S.; and van der Linden, S. 2021. Countering Misinformation and Fake News Through Inoculation and

- Prebunking. *European Review of Social Psychology*, 32(2): 348–384.
- Liu, Y.; Yildirim, P.; and Zhang, Z. J. 2022. Implications of revenue models and technology for content moderation strategies. *Marketing Science*, 41(4): 831–847.
- Maertens, R.; Roozenbeek, J.; Basol, M.; and van der Linden, S. 2021. Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, 27(1): 1–16.
- Martínez-López, F. J.; Li, Y.; and Young, S. M. 2022. Social Media Monetization and Demonetization: Risks, Challenges, and Potential Solutions. In Martínez-López, F. J.; Li, Y.; and Young, S. M., eds., *Social Media Monetization: Platforms, Strategic Models and Critical Success Factors*, 185–214. Cham: Springer International Publishing.
- McCabe, S. D.; Ferrari, D.; Green, J.; Lazer, D. M. J.; and Esterling, K. M. 2024. Post-January 6th deplatforming reduced the reach of misinformation on Twitter. *Nature*, 630(8015): 132–140.
- Modirrousta-Galian, A.; and Higham, P. A. 2023. Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General*, 152(9): 2411–2437.
- Niklewicz, K. 2017. Weeding Out Fake News: An Approach to Social Media Regulation. *European View*, 16(2): 335–335.
- Okoli, C. 2015. A Guide to Conducting a Standalone Systematic Literature Review. *Communications of the Association for Information Systems*, 37.
- Oleksy, T.; Wnuk, A.; Maison, D.; and Łyś, A. 2021. Content matters. Different predictors and social consequences of general and government-related conspiracy theories on COVID-19. *Personality and Individual Differences*, 168.
- Papadogiannakis, E.; Papadopoulos, P.; P. Markatos, E.; and Kourtellis, N. 2023. Who funds misinformation? A systematic analysis of the ad-related profit routines of fake news sites. In *Proceedings of the ACM Web Conference 2023*, 2765–2776.
- Papakyriakopoulos, O.; and Goodman, E. 2022. The Impact of Twitter Labels on Misinformation Spread and User Engagement: Lessons from Trump’s Election Tweets. In *Proceedings of the ACM Web Conference 2022*, WWW ’22, 2541–2551. New York, NY, USA: ACM.
- Pennycook, G.; Epstein, Z.; Mosleh, M.; Arechar, A. A.; Eckles, D.; and Rand, D. G. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592: 590–595.
- Porter, J. 2020. WhatsApp says its forwarding limits have cut the spread of viral messages by 70 percent. <https://www.theverge.com/2020/4/27/21238082/whatsapp-forward-message-limits-viral-misinformation-decline>. Accessed: 2025-03-05.
- Rochefort, A. 2020. Regulating Social Media Platforms: A Comparative Policy Analysis. *Communication Law and Policy*, 25(2): 225–260.
- Roozenbeek, J.; Linden, S. v. d.; and Nygren, T. 2020. Pre-bunking interventions based on “inoculation” theory can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review*, 1(2).
- Sharevski, F.; Alsaadi, R.; Jachim, P.; and Pieroni, E. 2022. Misinformation warnings: Twitter’s soft moderation effects on COVID-19 vaccine belief echoes. *Computers & Security*, 114.
- Tandoc, E. C.; Lim, D.; and Ling, R. 2020. Diffusion of disinformation: How social media users respond to fake news and why. *Journalism*, 21(3): 381–398.
- Thomas, D. R.; and Wahedi, L. A. 2023. Disrupting hate: The effect of deplatforming hate organizations on their online audience. *Proceedings of the National Academy of Sciences*, 120(24): e2214080120.
- Toff, B.; and Mathews, N. 2024. Is social media killing local news? An examination of engagement and ownership patterns in US Community news on Facebook. *Digital Journalism*, 12(9): 1397–1416.
- Tricco, A. C.; Lillie, E.; Zarin, W.; O’Brien, K. K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M. D. J.; Horsley, T.; Weeks, L.; Hempel, S.; Akl, E. A.; Chang, C.; McGowan, J.; Stewart, L.; Hartling, L.; Aldcroft, A.; Wilson, M. G.; Garritty, C.; Lewin, S.; Godfrey, C. M.; Macdonald, M. T.; Langlois, E. V.; Soares-Weiser, K.; Moriarty, J.; Clifford, T.; Tunçalp, O.; and Straus, S. E. 2018. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine*, 169(7): 467–473.
- Tucker, J.; Guess, A.; Barbera, P.; Vaccari, C.; Siegel, A.; Sanovich, S.; Stukal, D.; and Nyhan, B. 2018. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. SSRN:3144139.
- van der Linden, S.; Roozenbeek, J.; Maertens, R.; Basol, M.; Kácha, O.; Rathje, S.; and Traberg, C. S. 2021. How Can Psychological Science Help Counter the Spread of Fake News? *The Spanish Journal of Psychology*, 24: e25.
- Ventura, T.; Majumdar, R.; Nagler, J.; and Tucker, J. A. 2023. Misinformation Exposure Beyond Traditional Feeds: Evidence from a WhatsApp Deactivation Experiment in Brazil. SSRN:4457400.
- Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science*, 359(6380): 1146–1151.
- Walter, N.; and Murphy, S. T. 2018. How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, 85(3): 423–441.
- Warner, B. R.; and Neville-Shepard, R. 2014. Echoes of a Conspiracy: Birthers, Truthers, and the Cultivation of Extremism. *Communication Quarterly*, 62(1): 1–17.
- West, S. M. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11): 4366–4383.
- Yadav, K. 2020. Countering Influence Operations: A Review of Policy Proposals Since 2016. Technical report, Carnegie Endowment for International Peace.
- Yadav, K. 2021. Platform Interventions: How Social Media Counters Influence Operations. Technical report, Carnegie Endowment for International Peace.

Zainudin, J.; Mohamad Ali, N.; Smeaton, A. F.; and Taha Ijab, M. 2024. Intervention Strategies for Misinformation Sharing on Social Media: A Bibliometric Analysis. *IEEE Access*, 12: 140359–140379.

Zhou, H.; Lu, Y.; Zhao, L.; Wang, B.; and Li, T. 2024. Effective reporting system to encourage users' reporting behavior in social media platforms: an empirical study based on structural empowerment theory. *Behaviour & Information Technology*, 43(14): 3490–3509.

Paper Checklist

1. For most authors...

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. Privacy concerns are not violated as the citation data is available to the public already, and it is data that authors expect others to analyze or cite.
- (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes the claims in the abstract reflect the contributions listed in the conclusions.
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we describe in the introduction why a bibliometric approach is useful in understanding the limitations and gaps in the current literature.
- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? NA
- (e) Did you describe the limitations of your work? Yes, limitations are described in their own subsection right before the conclusion.
- (f) Did you discuss any potential negative societal impacts of your work? No, we do not discuss negative societal impacts in the paper. Because this data is public and authors intend for others to know they have published their work, we believe the work can only have neutral or positive impacts by describing gaps in the literature.
- (g) Did you discuss any potential misuse of your work? No we do not, for the same reasons as listed above.
- (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. Most of the items described in this question do not apply, but we have shared the list of papers and labels used so that others may reproduce the results.
- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

- (a) Did you clearly state the assumptions underlying all theoretical results? NA

- (b) Have you provided justifications for all theoretical results? NA
- (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
- (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
- (e) Did you address potential biases or limitations in your theoretical framework? NA
- (f) Have you related your theoretical results to the existing literature in social science? NA
- (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

- (a) Did you state the full set of assumptions of all theoretical results? NA
- (b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
- (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
- (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
- (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? NA

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

- (a) If your work uses existing assets, did you cite the creators? Yes, we cite all papers used in the analysis. We also use the proprietary software ORA, which is cited in the paper.
- (b) Did you mention the license of the assets? NA
- (c) Did you include any new assets in the supplemental material or as a URL? Yes. The URL for the curated list of papers analyzed for this article and their topic labels is included at the beginning of the paper.
- (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No. The data does include

author names, however the authors intend for their work to be cited by others. The data does not include any offensive content.

- (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes. The paper names and the labels we gave them are at a permanent and public location on Zotero. For each paper, the title and the reference information is listed, and the labels we gave each paper are listed under the tags. There is additionally a link to the publisher's version of the paper, to avoid any possible copyright issues. This public Zotero folder includes all the information needed to replicate the results of this paper.
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? No, we did not create a datasheet to go along with the Zotero repository. If requested, we can add a detailed notes document to the Zotero folder detailing why the dataset was collected, how the data was collected, the composition of the dataset, and any other possible uses of the data. Most of this information is detailed already in the Methods section of the paper.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
- (a) Did you include the full text of instructions given to participants and screenshots? NA
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
 - (d) Did you discuss how data is stored, shared, and de-identified? NA

Appendix A: Topic Labels

This appendix describes the labels given to the papers.

Content Distribution Refers to a broad category of interventions concerning how content is distributed on social media.

- *Redirection* - A form of content distribution where users are directed to alternative content (such as official resources) or no content at all when searching for something that could be problematic or harmful. For example, a user searching for COVID-19 leading to a CDC information box (Yadav 2021).
- *Accuracy Prompts* - Sometimes referred to as “nudging”, accuracy prompts are designed to encourage individuals to consider accuracy before posting or sharing content (Pennycook et al. 2021; Epstein et al. 2021).
- *Friction* - Friction encourages individuals to pause and reflect before engaging with content (Bago, Rand, and Pennycook 2020; Katsaros, Yang, and Fratamico 2022).
- *Platform Alterations* - Any modifications to the design or architecture of social media platforms that influence

how content is distributed or displayed to users or how they are encouraged to engage with the platform, such as reducing the size or visibility of a post (Kirchner and Reuter 2020; Gillespie 2022).

- *Advertising Policy* - Refers to how platforms (or governments) regulate, correct, or display advertisements to the public, such as by banning or fact-checking ads (Courchesne, Ilhardt, and Shapiro 2021; Helmus and Kepe 2021).
- *Content Distribution (Other)* - Other types involve limiting users' forwarding or resharing capabilities, which caps the number of recipients to whom a given message can be forwarded (Porter 2020).

Content Moderation Refers to a broad range of interventions regarding how content is displayed on social media, including fact-checking, narrative counterspeech like debunking, and the use of algorithms to assist in moderation and misinformation detection (Jiang et al. 2023).

- *Fact-Checking* - The process of verifying information, typically performed by experts. This verification can be done by experts, journalists, and platforms, and includes multi-modal fact-checking, such as fact-checking videos (Walter and Murphy 2018).
- *Debunking* - Debunking is a stronger form of fact-checking, where context and coherence are typically provided in addition to verifying or correcting content. It can also be described as a “narrative intervention” (Lewandowsky et al. 2020; Bruns et al. 2024).
- *Algorithmic Content Moderation* - Automated content moderation, including automated fact-checking, upranking or downranking of content, removing of content, or labeling of content (Bode and Vraga 2018; Gillespie 2022; Gorwa, Binns, and Katzenbach 2020).
- *Misinformation Detection* - The algorithmic detection of misinformation, typically for content moderation purposes (Khan, Michalas, and Akhunzada 2021)
- *Content Moderation (Other)* - Other forms of content moderation could include user control, which would involve transferring some moderation responsibilities currently done by platforms to users. This approach would give users greater control over the content displayed in their own news feeds (Jhaver and Zhang 2023).

Account Moderation Account moderation involves moderating user accounts by implementing actions such as account suspensions, removals, shadow banning users, or demonetizing accounts (Courchesne, Ilhardt, and Shapiro 2021; West 2018).

- *Account Removal* - Refers to the permanent or temporary banning of users who share misinformation or violate other platform policies a certain number of times. A specific type of account removal, where platforms coordinate their efforts to remove particularly problematic or dangerous user accounts, is typically referred to as deplatforming (McCabe et al. 2024; Thomas and Wahedi 2023).

- *Shadow Banning* - The practice of limiting the reach of posts from certain policy-violating accounts without explicitly banning or suspending them, typically conducted in a concealed or opaque manner (Johns et al. 2024; West 2018).
- *Account Moderation (Other)* - Another form of account moderation is the demonetization of user accounts, which refers to removing or restricting monetization features for a user account that is found to violate a platform's policies repeatedly (Martínez-López, Li, and Young 2022).

Content Labeling Content labeling includes all general types of misinformation disclosure through labeling. Labels are commonly used to present fact-checks, source information or credibility, or to provide additional context on a post (Courchesne, Ilhardt, and Shapiro 2021; Yadav 2021).

- *Crowdsourcing* - Crowdsourcing generally involves utilizing regular individuals to verify information and label content instead of relying on journalists or expert fact-checkers (Allen, Martel, and Rand 2022)
- *Warning Labels* - Warning labels refer to general warnings about misinformation and can address the source, content, or context (Papakyriakopoulos and Goodman 2022). One way to implement this type of intervention is by using click-through warning labels or interstitials (Sharevski et al. 2022).
- *Source Credibility Labels* - This type of intervention involves disclosing or labeling the credibility of a post's source (Gao et al. 2018).
- *Context Labels* - Labels that specifically add context or additional information to a post, such as Community Notes programs (Allen, Martel, and Rand 2022).
- *Content Labeling (Other)* - Another type of labeling involves notifying users when they have posted or interacted with content verified to contain misinformation or originating from a state-run source, through the use of disclosure (Courchesne, Ilhardt, and Shapiro 2021).

User-based Countermeasures User-based measures are interventions that focus on individuals' responses to encountering misinformation (Tandoc, Lim, and Ling 2020).

- *Reporting* - Users can report other users or their posts (Niklewicz 2017; Zhou et al. 2024).
- *Blocking* - Users have the ability to block other users or specific topics (Tandoc, Lim, and Ling 2020).
- *Social Corrections* - Users who fact-check or debunk other users directly (Badrinathan and Chauchard 2023; Bode and Vraga 2018). This intervention may involve publicly commenting on a post or privately messaging the misinformation poster.
- *Social Norms* - The use of social or community influence to change behavior and promote social and self-corrections (Ecker et al. 2023; Gimpel et al. 2021).
- *Retractions* - When users or organizations retract misinformation they have posted and how that impacts individuals who have already encountered the misinformation (Arif et al. 2017; Ecker and Antonio 2021).

- *User-based Measures (Other)* - Other forms of user-based measures might include encouraging users to deactivate their social media accounts (Ventura et al. 2023), along with other user-driven behaviors.

Media Literacy and Education This general intervention category involves any educational or training effort aimed at enhancing the public's civic reasoning, digital literacy, and critical thinking skills when interacting with media messages (Guess et al. 2020; Jeong, Cho, and Hwang 2012).

- *Fake News Games* - Games designed to help players detect misinformation and improve their critical thinking skills (Maertens et al. 2021; Modirrousta-Galian and Higham 2023; Roozenbeek, Linden, and Nygren 2020).
- *Inoculation* - Commonly referred to as "pre-bunking," inoculation consists of warning messages or information about misleading rhetorical techniques meant to prevent people from later believing misinformation (Lewandowsky and van der Linden 2021).
- *Media Literacy (Other)* - Other types of educational initiatives and relevant research studies, such as providing individuals with tips on recognizing fake news (Guess et al. 2020) or evaluating people's information, digital, or news literacy (Jones-Jang, Mortensen, and Liu 2021).

Institutional Measures Institutional measures are those implemented by civic society, governments, the media, or other organizations (Bradshaw and Neudert 2021).

- *Media Support* - Investing in local news or promoting trustworthy local news on social media platforms (Toff and Mathews 2024). Supporting and training the next generation of journalists to engage in high-quality, independent reporting (Bradshaw and Neudert 2021)
- *Data Sharing* - Social media companies should regularly release data and internal research reports on the prevalence, spread, and mitigation of misinformation to the public and outside researchers (Assenmacher et al. 2022; Bishop and Gray 2017).
- *Government Regulation* - This category includes any laws, rules, or regulations at local, state, or federal levels (Niklewicz 2017; Rochefort 2020; Yadav 2020).
- *Other Institutional Measures* - Other measures may involve researching and developing tools to support civil society with these issues, as well as enhancing collaboration among various types of institutions (Bradshaw and Neudert 2021).

Other Interventions This category includes any interventions that do not fit into the previous categories or are newly introduced.

- *Generative AI* - Using generative AI to combat or detect misinformation. An example could include employing AI chatbots to reduce belief in conspiracy theories or misinformation (Costello, Pennycook, and Rand 2024). Policy responses may involve prohibiting, labeling, or disclosing the use of AI or manipulated content to produce deepfakes or ads (Helmus and Chandra 2024).

- *Combining Interventions* - Studies that compare the effects of using multiple interventions simultaneously with using one intervention (Bak-Coleman et al. 2022).
- *Other* - Any other intervention not previously described.

Other Labels

- *Review Article*: A paper that reviews other papers in a specific field.
- *Meta-Analysis*: A review paper that quantitatively analyzes previous results.
- *Effectiveness*: A paper that studies and analyzes the effectiveness of one or more interventions at reducing the creation, spread, or belief in misinformation.
- *Acceptance*: A paper that studies user acceptance or incorporates user feedback in designing and analyzing misinformation interventions.

Appendix B: Results

Table 5: The number of papers and unique authors who have studied each intervention type.

| Intervention | Category | Papers | Authors |
|---------------------------|------------------------|--------|---------|
| fact-checking | Content Moderation | 127 | 336 |
| debunking | Content Moderation | 124 | 359 |
| media literacy | Media Literacy | 76 | 256 |
| inoculation | Media Literacy | 58 | 214 |
| warning labels | Content Labeling | 45 | 170 |
| source credibility labels | Content Labeling | 38 | 155 |
| social corrections | User-Based Measures | 35 | 88 |
| accuracy prompts | Content Distribution | 31 | 116 |
| social norms | User-Based Measures | 30 | 163 |
| fake news games | Media Literacy | 26 | 66 |
| retractions | User-Based Measures | 25 | 45 |
| platform alterations | Content Distribution | 23 | 75 |
| crowdsourcing | Content Labeling | 19 | 67 |
| misinformation detection | Content Moderation | 18 | 72 |
| combining interventions | Other | 14 | 48 |
| institutional measures | Institutional Measures | 13 | 41 |
| friction | Content Distribution | 12 | 63 |
| context labels | Content Labeling | 12 | 67 |
| algorithmic moderation | Content Moderation | 12 | 41 |
| content labeling | Content Labeling | 11 | 34 |
| other | Other | 9 | 37 |
| government regulation | Institutional Measures | 8 | 14 |
| advertising policy | Content Distribution | 7 | 29 |
| media support | Institutional Measures | 6 | 16 |
| content distribution | Content Distribution | 5 | 13 |
| redirection | Content Distribution | 5 | 12 |
| user-based measures | User-Based Measures | 5 | 14 |
| account removal | Account Moderation | 4 | 12 |
| reporting | User-Based Measures | 4 | 13 |
| content moderation | Content Moderation | 3 | 7 |
| generative AI | Other | 3 | 8 |
| account moderation | Account Moderation | 2 | 8 |
| shadow banning | Account Moderation | 1 | 4 |
| blocking | User-Based Measures | 1 | 4 |
| data sharing | Institutional Measures | 1 | 3 |

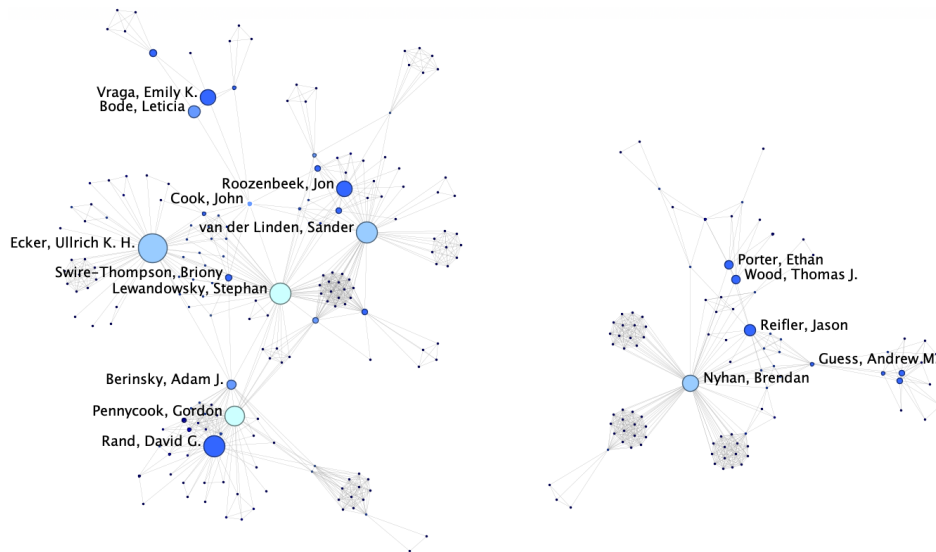


Figure 8: The largest component of the Author x Author network excluding one paper with 30 authors.