Thesis Proposal
Effective and Practical Strategies for Combatting Misinformation

Catherine King

February 20th, 2024

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Prof. Kathleen M. Carley, Chair, Carnegie Mellon University
Prof. Hong Shen, Carnegie Mellon University
Prof. Chris Labash, Carnegie Mellon University
Prof. Pablo Barberá, University of Southern California and Meta

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Abstract

Social media platforms, which are becoming a primary news source for many individuals, can quickly spread mis/disinformation online faster than ever before. These information disorders contribute to increased polarization and extremism, threatening to undermine democracy and trust in public institutions worldwide. Because of this growing problem, researchers have begun investigating the effectiveness of possible interventions to counter this misinformation. This research is critical given the many societal challenges we face that are associated with the spread of false or misleading information.

Most research in the countermeasures space focuses on the effectiveness of some more easily studied interventions. Some interventions, like fact-checking, are studied more than others because they can be analyzed without complete access to comprehensive social media data. Most researchers also focus on determining the effectiveness of an intervention without considering if the public would support the countermeasure. Platforms and governments will likely only implement changes that have public support.

In this thesis, I develop a framework for designing and evaluating misinformation interventions that integrates current research on effectiveness with user acceptance to enable more effective implementation strategies. To accomplish this task, I provide a detailed categorization of interventions. Then, a citation network analysis is run on the literature in this field to find research gaps, areas of disagreement, and possible next steps. I will run a comprehensive survey asking the American public about their social media behavior and their opinions on various interventions. The survey also investigates how possible factors may affect user acceptance and belief in effectiveness. These factors include transparency, fairness, and intrusiveness. Next, an effectiveness study on a training game will be run to add to the currently contentious literature on the effectiveness, or potential lack thereof, of training games. Finally, I will combine this research to assess countermeasures across a comprehensive list of features, aiming to identify the shared characteristics that make countermeasures effective and practical.

Contents

Chapter 1

Introduction

1.1 Thesis Goal

In recent years, there has been an increased research focus on the spread, impact, and mitigation of misinformation online. Social media is aiding in the dissemination of misinformation [1], and researchers are growing more concerned about how social media may be contributing to political polarization and distrust in institutions and the media. Information disorders like misinformation and disinformation have been shown to have pressing societal impacts ranging from undermining democracy [85], increasing extremism [90], and lowering the uptake of various public health measures during a pandemic like COVID-19 [67].

Countering misinformation is a challenging problem, as there are many possible solutions and aspects to consider. Researchers also often only have limited access to social media data, especially data that could be used to evaluate the effectiveness of various countermeasures [28]. Even if data is available, in some cases there are ethical challenges associated with sharing social media data with other researchers [14]. This lack of access contributes to why some countermeasures, like fact-checking, are studied significantly more than others.

According to a review of 223 countermeasures studies since 1972 by Courchesne et al. (2021), there has been a disproportionate amount of research on the effect of fact-checking [4, 6], debunking [23, 31], and prebunking [51, 77]. However, many countermeasures, including those that could target creators of disinformation, have not been studied at all [28]. Finally, most intervention papers focus on the effectiveness of the intervention without considering crucial aspects like user acceptance, political feasibility, and cost.

The goal of this thesis is to better understand the efficacy and practicality of misinformation countermeasures in order to provide analysis-driven recommendations. I propose an approach to developing misinformation interventions that (1) integrates current social science theory about effectiveness with (2) user opinions and acceptability while considering other relevant factors, such as transparency, cost, and fairness. My main research questions are as follows:

1. **How can we assess how practical and effective countermeasures are?**
2. **What do successful countermeasures have in common?**
3. **Can we develop a framework for providing analysis-driven recommendations on what to implement and why?**

This thesis is limited in scope to user-based countermeasures, social media platform countermeasures, and possible government regulation. The selected interventions are described in more detail in the literature review section below.

1.2 Literature Review

1.2.1 Social Cybersecurity

Social media misinformation has become a growing problem around the world. Researchers in various fields have been investigating the most effective and acceptable ways to counter fake news online. The research is in the emerging scientific area known as *social cybersecurity*, and it is defined as investigating the impact of the online information space on society, culture, and politics [21, 27]. This research area analyzes information and network maneuvers and their possible effects on human behavior and opinion.

1.2.2 Existing Reviews of Misinformation Countermeasures

Several review papers and meta-analyses have been written in the misinformation intervention space. Some reviews, like Helmus and Keppe (2021) from the Rand Corporation, focus on related policy papers [38]. Others examine specific intervention categories, such as content moderation [44] or media literacy [41]. The most comprehensive review found so far has been the article from Courchesne and colleagues in 2021 [28]. These researchers found that certain types of platform interventions are overstudied relative to others [28]. Specifically, fact-checking and debunking are by far the most common interventions studied. Still, little to no research has been conducted on other countermeasures that directly target creators, such as redirection.

However, none of these reviews have analyzed the broader picture and included platform interventions and possible government policies. Additionally, most platform review articles focus on testing countermeasures and analyzing their effectiveness but fail to discuss the equally important metric of their practicality and acceptability to users. This thesis seeks to fill this gap by considering both platform and policy interventions.

1.2.3 Types of Countermeasures

User-based Countermeasures

User-based measures are an often overlooked aspect of countering misinformation. If misinformation is successfully posted on social media, other users are the first line of defense, as they can report or debunk the misinformation. Individual-level debunking, especially from trusted messengers, has been found to be effective in a variety of contexts [10, 17, 53, 83, 88]. User-based countermeasures are addressed in more detail in Chapters 2 and 3.

Platform and Government Countermeasures

Many of the review articles used a similar categorization of countermeasures; however there is no common typology [28, 37, 38, 93]. After reviewing the literature and these previous categorizations, I developed eight general categories of countermeasures [45], as shown in Table 1.1. The first six categories can apply to both platforms or governments; platforms could implement these changes willingly, or governments could require these changes.

Category	**Example Interventions**
Content distribution	Delay posting unviewed, de-emphasize/ downrank content
Content / account moderation	Ban or suspend certain users, remove certain posts
Content labeling	Tell users if they have posted misinformation, label posts
Advertising policy	Require fact-checking ads, ban political ads
Media support	Promote and invest in local news
Media literacy and awareness	Invest in and promote educational content, regularly release social media data to 3rd party researchers
User-based countermeasures	Reporting users or posts, social corrections
Other	Government regulation, combining interventions

Table 1.1: Misinformation intervention categories

1.3 Citation Network Analysis

To supplement this literature review, I conducted a citation network analysis of relevant papers to generate a more profound contextual background on the state of the literature in this field.

I derived a comprehensive list of specific interventions from the eight categories described in Table 1.1. Each paper was labeled with which countermeasures they discuss. Papers are also assigned labels if they are review articles, meta-analyses, and papers examining intervention effectiveness or acceptance. Section 1.3.1 describes all labels used in this article.

This work was presented as a poster at SBP-BRIMS in September 2023. It has since expanded, and once additional analyses are complete it will be submitted as a full paper to another conference in the spring. Request draft to see more detailed information on how the papers were selected, the inclusion criteria, and the inter-rater reliability.

1.3.1 Paper Labels

Table 1.2 shows all 31 labels used in this citation network analysis. See the Appendix for more detailed definitions and citations. It is important to note that these labels are not mutually exclusive, as some papers can cover multiple interventions.

Category	Label	Definition
Content Distribution	Content Distribution	The distribution of content on social media
	Redirection	Redirecting users to other content when searching
	Nudging	Nudging users to guide them to better decisions
Content / Account Mod.	Content Moderation	How content is shown or removed on social media
	Fact-Checking	Verification of information
	Debunking	Fact-checking with context, narrative coherence
	Misinformation Detection	Algorithmic detection of misinformation
	Algo. Content Moderation	Automated content moderation
	Continued-Influence Effect	Related to the effectiveness of moderation/corrections
	Account Moderation	Moderating user accounts through suspensions, bans
	Deplatforming	The removal of a user from one or more platforms
Content Labeling	Content Labeling	A type of misinformation disclosure through labels
	Crowdsourcing	Using regular people to verify and label information
	Source Credibility	Disclosing or labeling a post's source
Advertising Policy	Advertising Policy	What ads are shown to which users
Media Support	Media Support	Investing in or promoting local and/or reliable news
Media Literacy and Awareness	Media Literacy	Efforts meant to improve the public's civic reasoning
	Fake News Games	Games designed to help people detect misinformation
	Inoculation	Pre-bunking misinformation
	Proactive Warning	Warnings about misinfo before or while viewing it
	Data Sharing	Sharing high-quality data with researchers
User-based	User-based	How people respond to seeing misinformation
	Reporting	Users can report users or their posts
	Social Corrections	Users that fact-check/debunk other users
	Retraction	When accounts retract misinformation they posted
Other	Government Regulation	Any relevant laws, rules, or regulations
	Combining Interventions	Using multiple interventions at once
Other Qualitative Labels	Review Article	A paper that reviews other papers in a specific field
	Acceptance	A focus on user acceptance, intervention popularity
Other Quantitative Labels	Meta-Analysis	A review paper that analyzes previous results
	Effectiveness	Measuring effectiveness of one or more interventions

Table 1.2: Citation network analysis labels

1.3.2 Topic Analysis

Descriptive Statistics

I used the ORA software [20] to conduct network analysis and create visualizations from this set of papers. First, I analyzed the Topic x Article network, calculating the descriptive statistics on the number of papers assigned to each label. The statistics show there is a minimum of 2 to a maximum of 48 documents per label, with other relevant values: 1st Quartile: 7, Median: 9, Mean: 12.9, and 3rd Quartile: 17.5. Furthermore, I found that 85 papers (60%) analyzed an intervention's effectiveness, 11 papers (8%) examined user acceptance and only two papers concentrated on both.

Over- and Under-Studied Interventions

The Co-Topic network was analyzed next, excluding the qualitative and quantitative metric labels (*Review Article, Acceptance, Meta-Analysis*, and *Effectiveness*). Figure 1.1 shows the Co-Topic network, with nodes sized by Total Degree Centrality and colored based on whether they are relatively under or over-studied. Labels in red represent the bottom quartile, assigned to 7 or fewer papers. Labels in green represent the top quartile, assigned to 17.5 or more papers. Finally, nodes in blue are in the middle 50%.

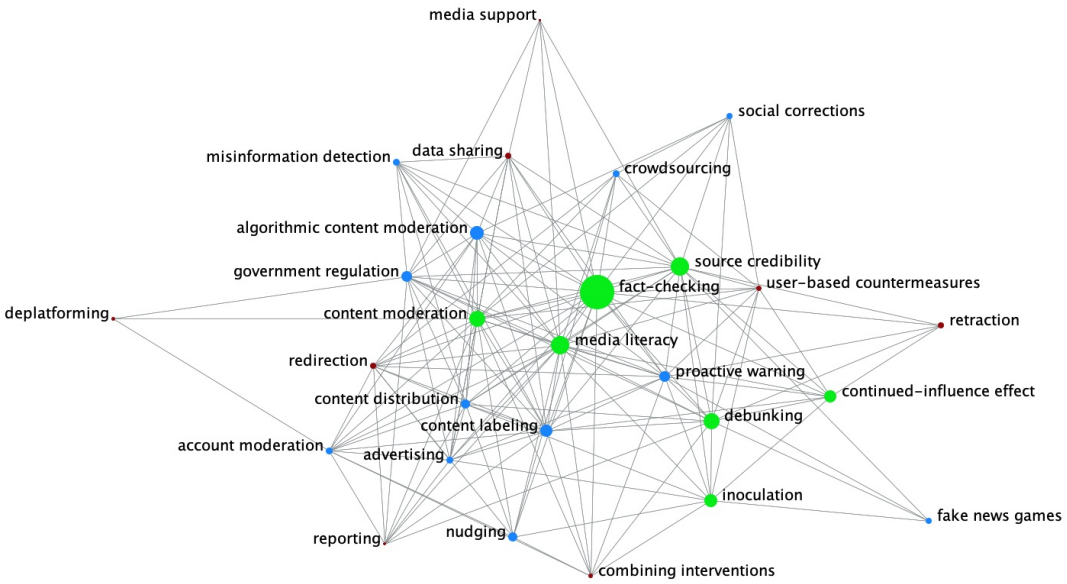

Figure 1.1: Visualization of Topic x Topic Network.

Lack of Consensus

Amongst the most studied countermeasures, there are several sources of disagreement; a body of work claims the effectiveness of the "Bad News" game for inoculation [12, 77], while a meta-review finds their results to be insignificant using ROC curves to compare pre & post treatment

classification accuracy [12]. The effectiveness of debunking [49, 63, 66] and nudging [56] interventions is also disputed despite a wealth of empirical research. This underscores a lack of comprehensive evaluation metrics [54] in the field and highlights the significance of meta-reviews. I plan to dive deeper into why there appears to be a lack of consensus on several countermeasures, possibly by investigating the various methodologies and datasets used.

1.3.3 Disciplines

Few Cross-Disciplinary Journals

I examined the Co-Publication Venue network and found a low density of 0.035. Of the 83 publication venues in our dataset, there are 41 isolates and three dyads, leaving only 39 outlets in the main component. This indicates how disjointed the literature is on this topic. Figure 1.2 shows the large component. Nodes are sized by total degree centrality and colored by betweenness (red indicating higher betweenness and blue indicating lower). A selection of venues are highlighted. The Harvard Misinformation Review has relatively high betweenness, suggesting it is an interdisciplinary journal bridging many fields. Additionally, the left side of the network predominantly consists of Psychology journals. The top-right comprises mainly of Communication and Journalism journals, and the bottom-right represents a mix of fields.

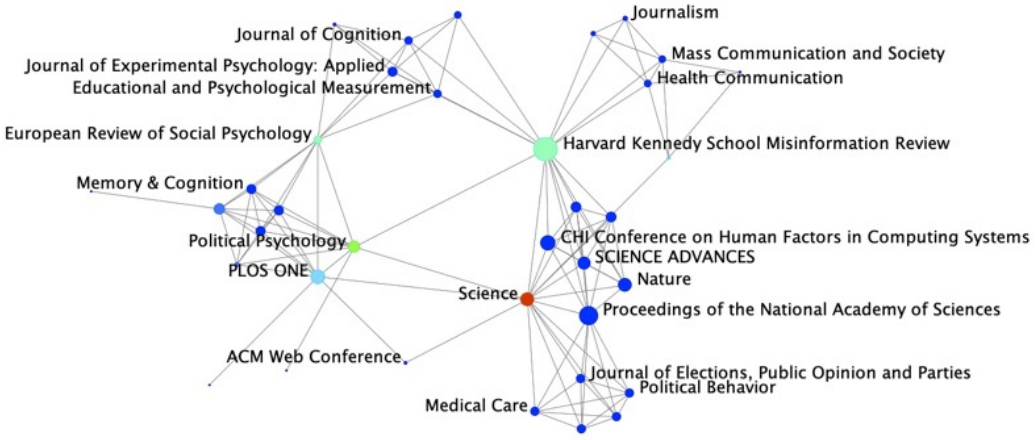

Figure 1.2: Visualization of Publication Venue x Publication Venue Network.

1.3.4 Planned Analyses

The following additional analyses will be conducted on the literature in this field:

1. **Authors and Affiliations** - I will investigate the Co-Authorship network to highlight the potential disjointedness in the research area. Author affiliations, including their primary institution and location will be used to determine where most of the research in this field is conducted.
2. **Venue / Topic Interaction** - I plan to analyze which venues discuss which topics. If different venues primarily look at certain interventions or topics, this would provide further evidence of disjointedness in the literature.
3. **Misinformation Category** - Another possible analysis is to label the papers and venues with the type of misinformation they work on countering. The disjointedness in the Co-Publication Venue network may be explained by researchers primarily citing intervention research from within their field (for example, political or health misinformation researchers only citing from political science or health journals).
4. **Other** - Reviewers of this work may suggest additions to the list of papers or additional analyses.

Chapter 2

Data

This thesis will use five primary datasets. These datasets will be used in characterizing countermeasures while considering several relevant factors.

Countermeasures Survey Data

I will be collecting data from approximately 1000 respondents. This data will include standard demographic questions, behavior and opinions concerning user-based, platform-based, and government-level countermeasures. This dataset will be used in Chapters 2 and 4.

OMEN Training Quiz Data

I will be collecting data from approximately 15-40 OMEN participants. The respondents will take a pre-quiz to measure their knowledge of misinformation and countermeasures detection. They will then undergo relevant training. Finally, they will take a post-training quiz to see if there was any improvement in their detection abilities. This dataset will be primarily used for Chapter 3.

Social Media Posts

For Chapter 3's pre and post-test quiz, we will show them various misinformation, conspiracies, pink slime, and accurate news posts. These posts will mostly be taken from Twitter and Facebook from COVID-19 and climate change datasets. The research team will generate other posts as necessary.

Curated Countermeasures Dataset

For Chapter 5, I will use all previously listed datasets and curate a new one. This dataset will include all countermeasures used in the thesis and their characteristics. These characteristics include effort level, cost, political feasibility, effectiveness, acceptance, policy changes, and information changes.

Curated Policy Dataset

For Chapter 5, I will create a curated policy data set. This will include information on existing or planned laws and regulations and policy analysis.

Chapter 3

Research Plan

3.1 Chapter 2: Characterizing User-based Countermeasures

3.1.1 Introduction

This thesis seeks to characterize all types of misinformation interventions, including user-led countermeasures. Studying individual behavior in response to seeing misinformation is critical because previous research has shown that debunking myths is more effective when it comes from a trusted source, like a friend or family member [53]. Additionally, individual-level debunking has shown to be highly effective [10, 17, 88]. This suggests that individuals responding directly to misinformation in real-time can help slow or stop the spread of misinformation. Social media companies, such as X (formerly known as Twitter), are now piloting programs like Birdwatch where users report posts containing misinformation and add context [4].

In this study, I investigate the behavior and opinions of American social media users when they see or post misinformation. I will survey approximately 1,000 American social media users using at least one platform weekly. This survey will cover the social media platforms where they encounter misinformation, if they have posted misinformation (intentionally or unintentionally), their reactions to seeing or posting misinformation, and their opinions on how they think others should act.

NOTE: *This study's survey pre-registration is currently under revision at* ***Scientific Reports****. It is possible that some suggested changes to the research plan may not be able to be implemented if the peer-reviewed pre-registration is accepted. Please request my draft pre-registration if you wish to see more details. The remaining work for this chapter is updating the pre-registration based on reviewer comments, running the survey, analyzing the results, and writing a complete paper. Complete draft survey available: Google Docs Link.*

3.1.2 Research Questions

RQ1.1 How do people respond and think others should respond when they see misinformation?

There are many ways a social media user could respond to a post containing misinformation, including reporting the post or blocking the misinformation poster. All possible responses are described in more detail in Table 3.1. Users may believe that people should do more to respond to misinformation than what they actually do themselves for a variety of reasons. First, responding to misinformation accurately and effectively can be time-consuming [80]. Additionally, people may feel overwhelmed or like they are having little impact due to the amount of misinformation they see online [82]. If people are hypocritical and hold others to a higher standard than themselves, this gap could be leveraged to induce prosocial behavioral changes (e.g., directly addressing content they believe contains misinformation)[5, 33, 81]. Possible platform policies could encourage individuals to take more agency in countering the misinformation they see.

RQ1.2. How do people respond and think others should respond behave after realizing they have posted misinformation?

A social media user could behave in many ways when they realize they have posted misinformation, including deleting their post or updating it with the correct information. Scholars now believe most people who spread misinformation do so by accident due to a lack of analytical thinking [69]. If this is the case, encouraging people to pause and think analytically before posting could be effective. Therefore, response to oneself posting misinformation can also have important implications for subsequent sharing. Possible responses are described in more detail in Table 3.2.

RQ2. Do responses differ based on who posted the misinformation and where it was posted?

There may be a difference in how people respond to seeing misinformation based on whether it is posted by a close friend or family member, an acquaintance, or a person they have never met offline. Previous work shows users are more likely to correct a close contact because it is perceived as more worthwhile[82]. If users are going to take the time to engage with misinformative content directly, they want to feel like it will have an impact. Based on previous research, I expect that people respond with more effort when the sender of misinformation is a close contact than a somewhat close contact and a somewhat close contact than a not close contact. It will be interesting to see what differences exist between platforms and if any differences may be due to that platform's misinformation policies.

Additionally, I expect that people will expend a different level of effort to respond to misinformation online posted by others than to misinformation they later realize they posted. People may feel embarrassed by misinformation they post and choose to delete it rather than draw attention to it and retract the content. If this is the case, reducing stigma and improving media literacy education may help individuals correct their mistakes without feeling embarrassed. Or, people may be more inclined to correct themselves than other people because they care about their credibility. One study in Singapore found this to be the case [64].

RQ3. What factors affect behavior and opinions on this topic?

Finally, we investigate how behavior and beliefs about responses to misinformation on social media vary by partisanship and other demographic factors (RQ4). Previous research suggests that Democrats tend to be more supportive of more aggressive platform interventions [59, 78]. Does this translate to more support for individual behavioral interventions?

3.1.3 Proposed Work

Ethics Information

The Institutional Review Board of Carnegie Mellon University approved this survey, numbered "STUDY2022_00000143". They approved this study as exempt from a full review because it is a survey that does not collect personally identifiable information. Informed consent will be obtained from all participants. We expect the study to take 18 minutes based on pre-tests. We will pay the equivalent of $10/hour, so for an 18-minute survey, participants will be paid $3 each.

Survey Design

The survey was designed to answer this document's research questions and hypotheses. There are additional related questions on this survey that are not used in this chapter (they are used in Chapter 4).

Sampling Plan

There will be approximately 1,067 participants in our survey. Our survey was implemented using Qualtrics and will be administered through Cloud Research, an online recruiting platform using Mechanical Turk survey participants. Only those respondents who are United States citizens, adults, and use social media at least once a week will be given the entire survey. We will employ several methods to recruit relevant participants and maintain high data quality, including bot and duplicate detection.

Measures

For RQ1, participants are asked about their responses to seeing misinformation (Measure 1a) and their opinions on how others should respond (Measure 1b). Table 3.1 shows a list of these possible responses to seeing misinformation, generalized to apply to various social media platforms, and rated as *no effort*, *minimal effort*, or *most effort*. Participants may select all options that apply. The only *no effort* response is ignoring the post. A *minimal effort* response means an action was taken, but there was no interaction with the content directly. A *most effort* response indicates that the user likely took more time to respond and interacted with the content directly.

For RQ2, participants are asked how they respond after posting misinformation (Measure 2a) and their opinion on how others should respond (Measure 2b). Table 3.2 shows a list of possible actions someone could take after realizing they posted misinformation. Again, they may select all options that apply. They are rated in the same fashion as the efforts described in Table 3.1.

Response	Effort Level
Ignore the post	No Effort
Report the post	Minimal Effort
Report the user	Minimal Effort
Block the user	Minimal Effort
Unfollow or unfriend the user	Minimal Effort
Privately message the user	Most Effort
Comment a correction on the post	Most Effort
Create a separate post with the correct information	Most Effort

Table 3.1: Actions social media users can take when they see misinformation online.

Response	Effort Level
Leave post as is	No Effort
Delete the post	Minimal Effort
Comment a correction on the post	Most Effort
Update the main post with a correction	Most Effort
Create a new post with the correct information	Most Effort

Table 3.2: Actions social media users can take when they realize they have posted misinformation.

Anything labeled in Tables 3.1 and 3.2 as no effort receives a score of 0, as minimal effort will receive a score of 1, and as most effort will receive a score of 2. Coding these as numerical, ordinal values allows for statistical analysis while differentiating between the three levels of effort. To get a participant's total effort expended for these measures, we will sum the total numerical effort level selected by the user. The total possible effort level for responding to misinformation posted by others is 10. The total possible effort level for responding to misinformation posted by oneself is 7.

3.1.4 Summary

Individuals can help defend against the spread of misinformation. Using trusted messengers can be one of the more effective ways to counter misinformation. This study will show how individuals behave when seeing misinformation across platforms, and this information can help inform future design choices. Many of the design choices that could boost user actions can be unobtrusive and transparent, which may lead to more acceptance among users compared with opaque platform suspension policies. Additionally, user-based countermeasures have the least possible conflict with the 1st Amendment in the United States when compared to some platform and government-level countermeasures.

3.2 Chapter 3: Improving User-Based Countermeasures

3.2.1 Introduction

Several studies have shown the possible effectiveness of training people to detect misinformation. Some of these games include the Bad News Game [12], Go Viral! [61], Troll Spotter [50], and Harmony Square [76]. If these games are effective, platforms could require or offer training games that engage with users' critical thinking or misinformation detection skills.

I have helped to design the OMEN game (Operational Mastery of the Information Environment) [46]. The goal of this project is to design and develop a training game to teach analysts and decision-makers how to detect and counter misinformation on social media. OMEN is designed to be a "train-as-you-fight" game, where the storyline is based on real events, and the data is realistic in volume and speed. The game accommodates real tools and workflow, including ORA and NetMapper[1]. It is a multi-day event and, in general, matches what the analysts would encounter on their day job. See our tech report [46] for more details on the OMEN game design, storyline creation, data curation, and learning objectives, and lessons learned.

3.2.2 Research Questions

While the previous chapter analyzed the current status of user-based countermeasures, this chapter tackles how to improve user detection of misinformation. There are two primary research questions:

RQ1. Does targeted training improve misinformation detection?

RQ2. Does targeted training improve the ability to counter misinformation?

Taken together, the results of this work will help inform how to improve user-based countermeasures. I am in the process of designing a project that investigates both of these research questions. This project is described in the next section.

3.2.3 Proposed Work

I propose that this project is done in the context of the OMEN "train-as-you-fight" game because it will combine both a lab and field analysis setting, so it has the advantages of both. Before players participate in the OMEN game, they are currently given several training sessions. These sessions include general social cybersecurity information, discussion of the BEND maneuvers [16], and how to use the ORA software. After the training presentations, the participants practice what they learned on a training data set, which is typically simpler than the data they will face in the actual OMEN exercise.

I propose adding additional training on misinformation detection and countering misinformation, and giving participants pre- and post-training quizzes. About 15-40 participants will be involved in the project. This study will be submitted to the IRB before commencing.

[1] https://netanomics.com/netmapper/

Pre-test

Participants will be given a pre-test that lasts approximately 30 minutes. There will be two parts to the pre-test. First, participants will be shown a series of approximately 20 randomly selected posts from a larger pool of posts. They will be shown at least five pink slime posts, five real local news posts, five misinformation/conspiracy posts, and five real news posts. For each post, they will be asked the following series of questions:

1. What do you believe is the accuracy of the content in this post? *(True, somewhat true, neither true nor false, somewhat false, false)*
2. How trustworthy do you consider the poster of this message to be? *(Trustworthy to untrustworthy on a Likert 1-5 scale)*
3. How confident are you in your answer to question 1? *(Slider from very unsure to very confident, 1-10 scale)*
4. How confident are you in your answer to question 2? *(Slider from very unsure to very confident, 1-10 scale)*
5. Would you consider sharing this post online (for example, on Facebook, Twitter, or Instagram) *(Definitely yes to definitely no on a Likert 1-5 scale)*
6. Do you believe the poster of this message is trying to influence you? *(Definitely yes to definitely no on a Likert 1-5 scale)*
7. Please elaborate on the reasons for your answer to the previous question. *[Write-in]*

The second part of the pre-test concerns countering misinformation. The participants will be presented with a series of five false social media posts. The survey will tell the participants that these are false stories and ask what, if anything, they would do if they came across this post on their news feed. For each post, the participants will be asked the following questions:

1. How would you respond to this post if it appeared on your social media feed? *(These are the same responses as in Chapter 2 survey; they can select more than one option)*
    - Ignore the post
    - Report the post
    - Report the user
    - Block the user
    - Unfollow or unfriend the user
    - Privately message the user
    - Comment a correction on the post
    - Create a separate post with the correct information
    - Other [write-in]
2. Do you think your answer would change depending on how well you knew the person or organization posting it? *(Definitely yes to definitely no on a Likert 1-5 scale*
3. Please elaborate on why or why not your response would change depending on the person or organization posting it. *[Write-in])*

4. Would your response change based on the social media platform you saw this post on? *(Definitely yes to definitely no on a Likert 1-5 scale*
5. Please elaborate on how you would respond differently depending on the platform. *[Write-in]*

Training

The standard OMEN training will be run, plus additional training on misinformation detection and countermeasures will be run. This additional training will add no more than 1 hour to the training materials. After the training presentation, the participants then play with the training data. Possible changes will be made to the training data to include more misinformation and pink slime so that the participants will gain experience detecting misinformation on their own.

Post-test

The post-test will be administered after all the training is complete. It will be identical in structure to the pre-test, except it will have different randomly selected posts.

3.2.4 Summary

Digital media literacy is a critical tool in the fight against misinformation. Several previous studies have shown that some misinformation games can be effective. We developed OMEN, a misinformation training game designed to help train analysts to analyze social media data. I will test the lessons taught in OMEN to see if they helped improve misinformation detection and countering in the participants.

3.3 Chapter 4: Characterizing Platform and Government Countermeasures

3.3.1 Introduction

In addition to research on user-based countermeasures [10], there has been an increased focus on platform [93] and government misinformation mitigation measures [75, 92]. Reviewing this literature shows that public support is critical for countermeasure effectiveness [30, 47, 48]. It has been well-documented in the literature that public opinion impacts public policy implementation and effectiveness [19]. Therefore, knowing why people support or do not support certain countermeasures is important.

NOTE: *The survey questions associated with this Chapter are included in the same survey referenced in Chapter 2. See Chapter 2 for more details.*

3.3.2 Research Questions

In this work, I will consider multiple features that have been previously identified as relevant for climate change policy support. These features include fairness, intrusiveness, and effectiveness [39]. These features are applicable to misinformation interventions, as there are concerns over censorship and fairness among groups[26, 74]. We assume Americans would want fairness to be high and intrusiveness low. Therefore, we aim to develop fair, effective countermeasures and minimize intrusiveness.

RQ1. To what extent does a misinformation intervention's perceived fairness, intrusiveness, and effectiveness predict support?

Next, we consider if support for interventions depends on whether social media platforms or governments implement those interventions. In the United States, there is more concern about the government infringing on free speech than tech companies [59].

In this work, I use the top six categories from Table 1.1, which are content distribution, content and account moderation, content labeling, advertising policy, media support, and media literacy and awareness. These six categories were chosen because they can apply to both platforms and governments. Some of these categories of interventions are less transparent (content distribution, moderation, advertising) than other categories (content labeling, media literacy, and media support). Transparency may interact with fairness, intrusiveness, and effectiveness. This motivates the following two research questions:

RQ2. How do the attributes people consider when forming preferences change due to the implementer of the intervention?

RQ3. How do the attributes people consider when forming preferences change due to the transparency afforded by the intervention type?

3.3.3 Proposed Work

The ethics information, survey design, and sampling plan are identical to Chapter 2 (see Section 3.1.3. After collecting demographic data and data for Chapter 2, participants proceed to the second half of the survey. Each respondent sees every type of countermeasure but is randomly assigned to either the government or the platforms as the implementers.

Platform and Government Interventions

Ten interventions that could be implemented by both a social media platform or a government entity were chosen (Table 3.3). These interventions were selected to span all possible categories in the countermeasure categorization described in Section 1.2.3 (see Table 1.1). One or two representative interventions were chosen for each of the six relevant categories. The participants will be told that the entity classifying information as misinformation is up to the source that they were randomly assigned to. For example, if the participant is randomly assigned the government as the implementer, the government could decide to either set up an agency to determine the truth or outsource misinformation verification to an external, independent third party.

Category	Example Interventions
Content distribution	Temporarily delay users posting content the user did not open or spent less than a certain amount of time viewing, nudging them to think about the accuracy of what they posting
Content distribution	De-emphasize posts that are verified to contain misinformation to curb the spread
Content/account moderation	Permanently ban users who post misinformation a certain number of times
Content / account moderation	Remove posts verified to contain misinformation
Content labeling	Notify users if they posted content verified to contain misinformation
Content labeling	Label posts verified to contain misinformation with information about and from verified sources
Advertising policy	Require all advertising goes through a fact-checking process
Media support	Promote and invest in local media, which is thought to be most in tune with local norms, culture, and context
Media literacy and awareness	Invest in digital media literacy and promote educational content about detecting misinformation on and offline
Media literacy and awareness	Regularly release data and/or internal research reports to about misinformation prevalence, spread, and mitigation to the public and researchers not in industry

Table 3.3: Misinformation intervention categories

Measures

Support for intervention(s): We ask respondents to rate each intervention as {strongly support, somewhat support, neither support nor oppose, somewhat oppose, strongly oppose}. These responses are coded from 1 to 5 (least to most support).
Perceived fairness of intervention(s): We ask respondents to rate each intervention as {very fair, somewhat fair, neither fair nor unfair, somewhat unfair, very unfair}. These responses are coded from 1 to 5 (least to most fair).
Perceived intrusiveness of intervention(s): We ask respondents to rate each intervention as {very intrusive, somewhat intrusive, neither intrusive nor unintrusive, somewhat unintrusive, very unintrusive}. These responses are coded from 1 to 5 (least to most intrusive).
Perceived effectiveness of intervention(s): We ask respondents to rate each intervention as {very effective, somewhat effective, neither effective nor ineffective, somewhat ineffective, very ineffective}. These responses are coded from 1 to 5 (least to most effective).

3.3.4 Summary

Platforms and government entities have a variety of methods they can employ to fight misinformation. Little work has been done to measure public support for various mitigation measures. This work will impact which interventions are chosen as well as public messaging strategies to gather public support and participation.

3.4 Chapter 5: Recommendations for Effective and Practical Countermeasures

3.4.1 Introduction

While previous chapters have characterized a wide range of countermeasures and worked to improve those interventions, this chapter pulls that previous work together to build a framework for developing effective and practical countermeasures. This chapter aims to provide a comprehensive set of recommendations across the intervention space that researchers, companies, and policymakers can use.

3.4.2 Research Questions

To develop this framework, I will combine the research from previous chapters with information on various intervention features, including efficacy, acceptability, cost, and political feasibility. Each countermeasure will also be analyzed to find how they may target different types or aspects of misinformation. The guiding question for this chapter is:

RQ1: What features do effective and practical countermeasures have in common?

To address this larger research question, I must first tackle the following three sub-questions:

RQ1.1 What are the characteristics of misinformation? This sub-question addresses the misinformation suppliers, the types of misinformation, as well as their context, purpose, emotionality, and audience.

RQ1.2 What are the characteristics of various countermeasures? This sub-question addresses features like effectiveness, acceptance, cost, political feasibility, and effort level.

RQ1.3 Which countermeasures target which aspects of misinformation and why? Combining the results from the two previous sub-questions will allow the creation of analysis-driven recommendations.

3.4.3 Proposed Work

The proposed work will be divided into three main parts, roughing following the three sub-research questions in this chapter.

Characteristics of Misinformation

This section will provide an overview and comparison of how different researchers categorize misinformation, and using this prior work, I will describe a proposed characterization.

Table 3.4 shows the misinformation categories described in several prominent review papers. There are several categories that three or all four papers agree on (satire/parody, fabricated content). However, there are other categories (like propaganda and clickbait) where agreement diverges.

Type	Wardle et al. (2017)	Zann. and Siri. (2019)	Brennan et al. (2020)	Wang et al. (2022)
Satire and Parody	×	×	×	×
Fabricated Content	×	×	×	×
Manipulated Content	×		×	×
False Connection	×			×
False Context	×		×	×
Imposter Content	×		×	×
Error (False Content)				×
Propaganda		×		×
Conspiracies/hoaxes		×		
Rumors		×		×
Clickbait/ads		×		×
Other (biased,photo)		×		×

Table 3.4: Misinformation intervention categories

Analyzing this table, we see that many researchers include categories that either somewhat or strongly overlap with one or more other categories. For example, propaganda can be done via fabricated content, manipulated content, false content/error, or other types of misinformation. I noticed a general pattern that some of these categorizations, like propaganda and clickbait, refer more to the purpose and context of the misinformation than the style.

For a deeper analysis, I propose three categorizations of misinformation messages:

1. **Purpose**: The purpose of a misinformation message is its intention and primary goal. While some of these purposes can co-exist (for example, one could be promoting misinformation for both a political agenda and monetary reasons), this categorization helps define motive.
    - *Political/Propaganda* - This misinformation intends to influence political attitudes and opinions. This can include intentionally increasing polarization, inciting violence, and feeding into extremism.
    - *Monetary* - This misinformation is intended to generate clicks and increase revenue.
    - *Distraction* - A message meant to distract the public with a different story, confuse, or cause panic.
    - *Conspiratorial* - A conspiratorial message intended to promote conspiracy theories, hoaxes, and rumors.
    - *Accidental/No Purpose* - False information or context that spreads with no malicious intent.
2. **Context**: The context indicates the circumstances surrounding the message. Context may be related to how difficult it is to debunk misinformation, with some previous studies showing political misinformation is among the most difficult to debunk [87]. Five general news categories are listed below.

- News/Political
- Health/Science
- Business/Consumer
- Entertainment/Sports
- Other

3. **Type**: The type refers to how the message presents misinformation.
   (a) *Satire and Parody* - Humorous content that typically does not intend to cause harm.
   (b) *False Connection* - Content with headlines or captions that don't support the content.
   (c) *False Context* - Correct information shared with false context.
   (d) *Imposter Content* - Information posted while impersonating a genuine source or brand to gain credibility.
   (e) *Manipulated Content* - Text, image or video distortion; or a sensational or "clickbaity" title.
   (f) *Misleading Content* - Misleading information or opinions presented as facts.
   (g) *Fabricated Content* - A false story, completely made-up.
   (h) *Error (False Content)* - Generally a mistake by a reputable news organization or honest person.

In addition, misinformation comes with a supplier and an audience with differing priorities and goals. There are four primary categories of misinformation suppliers:

1. Government/Politicians
2. Vested interests: corporations, NGOs
3. The media
4. Regular people (rumors/hoaxes)

These suppliers will spread messages that align with their moral values and vary in impactfulness depending on the emotionality of their language.

Misinformation suppliers may target a general or a specific audience. Audiences are typically targeted on various demographic characteristics including nationality, age, gender, race, sexuality, religion, income, etc. People may also be targeted based on their membership in a group, like a consumer group, non-profit, company, etc. In my thesis, I will expand on this categorization and include more research and details on misinformation suppliers' emotionality and moral values.

Characteristics of Countermeasures

Each defined intervention used in the literature review in the Introduction (also see Appendix) will be compared and rated across various features. These features will include:

1. **Policy changes:** An intervention that may need a new organizational policy or new law/regulation to be implemented.
2. **Information changes:** Some interventions may require information changes including

altering, adding, or removing information, limiting access or available actions, or tagging certain information.

3. **Effort level:** The effort level for the user, platform, and/or government entity may vary substantially for different intervention types.
4. **Cost:** The cost for the user, platform, and/or government entity may vary substantially for different intervention types.
5. **Political feasibility:** Political feasibility refers to the likelihood that an intervention needing government approval would be implemented.
6. **Effectiveness:** Interventions may be effective in some circumstances, cross-platform, or cross-culturally. Effectiveness is impacted by acceptance level.
7. **Acceptance:** Interventions may be accepted in some circumstances, cross-platform, or cross-culturally. Acceptance is likely impacted by transparency, intrusiveness, privacy, fairness, and the implementer of the intervention.

Some of the results from previous chapters will go into this comparison, including effectiveness and acceptance scores from the survey described in Chapters 2 and 4, and effectiveness found in Chapter 3 for training games. An extended literature review will be conducted to fill out a detailed categorization and rating for each intervention on various characteristics, and add any characteristics as needed.

Analysis-Driven Recommendations

In addition to using previous results and an extended literature review, I will also use other external datasets.

- **Labeled misinformation and countering tweets** - Over 30,000 misinformation tweets and over 30,000 refuting tweets are labeled in the dataset used for the paper "The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic" [58]. One of the authors, Bing He, has shared this data. These tweets will be analyzed to determine additional characteristics and patterns in misinformation and countering posts.
- **Information on existing or proposed laws and regulations** - I will find a reliable database to gather this information.

To validate ratings, I will contact experts in the field or conduct a focus group.

3.4.4 Summary

Recommendations on the most effective and practical countermeasures are needed due to conflicting effectiveness research and lack of acceptance research. This chapter will discover which features are important in both misinformation and possible interventions. These features can then be used to develop and implement future countermeasures.

Chapter 4

Contributions and Limitations

4.1 Contributions

In this thesis, I have proposed an approach to developing and evaluating misinformation interventions that are both effective and accepted by the public. This proposed work will make several contributions to the research on countermeasures.

Theoretical

I will create a detailed categorization of misinformation countermeasures compared across a comprehensive list of features. This typology can be used by future researchers when developing and comparing possible interventions.

Academic

This thesis will have several academic contributions. The literature review in Chapter 1 of the thesis was presented at *SBP-BRiMS* as a poster in September 2023. It will additionally be submitted in fall 2023 as a full conference paper to a relevant conference.

The survey I developed will provide much-needed insight into behavioral user-based countermeasures, an understudied area of research. The pre-registration of the study on user-based countermeasures is currently under review at *Scientific Reports*. The survey will additionally provide insight into the popularity and perceived effectiveness of various platform and government misinformation interventions [45]. We plan to submit that work to the *Harvard Misinformation Review* or another relevant journal.

My work on developing the OMEN game has already been published as a Technical Report at Carnegie Mellon [46]. Analyzing the effectiveness of training in the OMEN context will be helpful to understand better the efficacy of training games in general, which is currently contested in the literature [12, 61]. It will be the first time training games are studied in a "train-as-you-work" environment. I plan to submit this study to a relevant conference or journal.

The remaining work in the last chapter of my thesis will combine this previous work and result in additional conference publications and presentations. This comprehensive trade-off

analysis between effectiveness and usability has not been done before. For a timeline of possible publications, see 5.1.

Datasets

First, I will collect data from over 1,000 participants on their social media behavior and opinions on user-based, platform, and government countermeasures. This survey data will provide critical evidence on the current status of user-based countermeasures and public opinion surrounding countermeasures. Next, I will create a dataset of pre and post-training quiz scores from OMEN participants. This dataset will contribute crucial information to the possible effectiveness of training games. Finally, I will create a curated dataset comparing countermeasures, including policy interventions, on many metrics and features, including efficacy and acceptance.

4.2 Limitations

There are several limitations to this proposal, specifically related to its scope and applicability across social media platforms and cultures. First, while the thesis intends to cover as many social media countermeasures as possible, it cannot include everything. Some types of misinformation or countermeasures may not be included, and I will not be able to add new interventions as they emerge. Next, I am limited to focusing on the social media behavior and opinions of American adults. The survey will ask participants about their behavior on the current list of the eleven most used social media platforms in the US as determined by Pew Research in 2021 [9]. Therefore, some emerging social media platforms may be missed. Additionally, the training game will be limited to college-educated American OMEN participants and will not be generalizeable to the general public. These studies are restricted to US citizens due to funding constraints

Chapter 5

Timeline

Figure 5.1 shows my proposed timeline from October 2023 to Feb 2025. It is broken up by chapters and tasks. In the fall of 2023, I will be submitting a citation network analysis paper, which is virtually complete. For Chapters 2 and 4, my pre-registration is under revision. If the pre-registration is accepted, that acceptance would occur mid or late Fall 2023, with the deployment of the survey and analysis of the results to begin immediately afterward. I plan to submit the final papers associated with the survey by the end of Spring 2024. Approximately half of both survey papers have already been written, as an introduction, literature review, methods, and analysis plan were required for the peer-reviewed pre-registration submission at *Scientific Reports*.

For Chapter 3, I expect to be able to run the OMEN survey in February 2024, with a complete write-up of the results by the end of Summer 2024. Because my final chapter needs to use some of the work from previous chapters, I will focus on it more heavily in the second half of 2024. Finally, I am reserving the fall and winter of 2024-2025 to write and defend my thesis.

Figure 5.1: Proposed Timeline.

Bibliography

[1] Zoë Adams, Magda Osman, Christos Bechlivanidis, and Björn Meder. (Why) Is Misinformation a Problem? *Perspectives on Psychological Science*, page 17456916221141344, February 2023. ISSN 1745-6916. doi: 10.1177/17456916221141344. URL https://doi.org/10.1177/17456916221141344. Publisher: SAGE Publications Inc. 1.1

[2] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Understanding the Effect of Deplatforming on Social Networks. In *ACM Web Science Conference*, WebSci '21, pages 187–195, New York, NY, USA, June 2021. Association for Computing Machinery. ISBN 978-1-4503-8330-1. doi: 10.1145/3447535.3462637. 5

[3] Jennifer Allen, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. Scaling up fact-checking using the wisdom of crowds. *SCIENCE ADVANCES*, 7(36), 2021. doi: 10.1126/sciadv.abf4393. 5

[4] Jennifer Allen, Cameron Martel, and David G. Rand. 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*, pages 1–19, New Orleans LA USA, April 2022. ACM. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3502040. 1.1, 3.1.1

[5] E Aronson, C Fried, and J Stone. Overcoming denial and increasing the intention to use condoms through the induction of hypocrisy. *American Journal of Public Health*, 81(12):1636–1638, December 1991. ISSN 0090-0036, 1541-0048. doi: 10.2105/AJPH.81.12.1636. URL https://ajph.aphapublications.org/doi/full/10.2105/AJPH.81.12.1636. 3.1.2

[6] Kevin Aslett, Andrew M. Guess, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances*, 8(18):eabl3844, May 2022. ISSN 2375-2548. doi: 10.1126/sciadv.abl3844. 1.1

[7] Kevin Aslett, Andrew M. Guess, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances*, 8(18):eabl3844, May 2022. ISSN 2375-2548. doi: 10.1126/sciadv.abl3844. 5

[8] Dennis Assenmacher, Derek Weber, Mike Preuss, André Calero Valdez, Alison Bradshaw, Björn Ross, Stefano Cresci, Heike Trautmann, Frank Neumann, and Christian Grimme.

Benchmarking Crisis in Social Media Analytics: A Solution for the Data-Sharing Problem. *Social Science Computer Review*, 40(6):1496–1522, December 2022. ISSN 0894-4393. doi: 10.1177/08944393211012268. 5

[9] Sara Atske. Social Media Use in 2021, April 2021. URL https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/. 4.2

[10] Sumitra Badrinathan and Simon Chauchard. "I Don't Think That's True, Bro!" Social Corrections of Misinformation in India. *The International Journal of Press/Politics*, page 19401612231158770, February 2023. ISSN 1940-1612. doi: 10.1177/19401612231158770. 1.2.3, 3.1.1, 3.3.1

[11] Sumitra Badrinathan and Simon Chauchard. "I Don't Think That's True, Bro!" Social Corrections of Misinformation in India. *The International Journal of Press/Politics*, page 19401612231158770, February 2023. ISSN 1940-1612. doi: 10.1177/19401612231158770. 5

[12] Melisa Basol, Jon Roozenbeek, and Sander van der Linden. Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News. *Journal of Cognition*, 3(1):1–9, 2020. ISSN 2514-4820. doi: 10.5334/joc.91. 1.3.2, 3.2.1, 4.1

[13] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like Trainer, Like Bot? Inheritance of Bias in Algorithmic Content Moderation. In Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri, editors, *Social Informatics*, Lecture Notes in Computer Science, pages 405–415, Cham, 2017. Springer International Publishing. ISBN 978-3-319-67256-4. doi: 10.1007/978-3-319-67256-4_32. 5

[14] Libby Bishop and Daniel Gray. Ethical Challenges of Publishing and Sharing Social Media Research Data. In Kandy Woodfield, editor, *The Ethics of Online Research*, volume 2 of *Advances in Research Ethics and Integrity*, pages 159–187. Emerald Publishing Limited, January 2017. ISBN 978-1-78714-486-6 978-1-78714-485-9. https://doi.org/10.1108/S2398-601820180000002007. 1.1

[15] Libby Bishop and Daniel Gray. Ethical Challenges of Publishing and Sharing Social Media Research Data. In Kandy Woodfield, editor, *The Ethics of Online Research*, volume 2 of *Advances in Research Ethics and Integrity*, pages 159–187. Emerald Publishing Limited, January 2017. ISBN 978-1-78714-486-6 978-1-78714-485-9. https://doi.org/10.1108/S2398-601820180000002007. 5

[16] Janice T. Blane, Daniele Bellutta, and Kathleen M. Carley. Social-Cyber Maneuvers During the COVID-19 Vaccine Initial Rollout: Content Analysis of Tweets. *Journal of Medical Internet Research*, 24(3):e34040, March 2022. doi: 10.2196/34040. URL https://www.jmir.org/2022/3/e34040. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. 3.2.3

[17] Leticia Bode and Emily K. Vraga. See Something, Say Something: Correction of

Global Health Misinformation on Social Media. *Health Communication*, 33(9):1131–1140, September 2018. ISSN 1041-0236. doi: 10.1080/10410236.2017.1331312. 1.2.3, 3.1.1

[18] Alexander Bor, Mathias Osmundsen, Stig Hebbelstrup Rye Rasmussen, Anja Bechmann, and Michael Bang Petersen. "Fact-checking" videos reduce belief in misinformation and improve the quality of news shared on Twitter, September 2020. 5

[19] Paul Burstein. The Impact of Public Opinion on Public Policy: A Review and an Agenda. *Political Research Quarterly*, 56(1):29–40, 2003. ISSN 1065-9129. doi: 10.2307/3219881. URL https://www.jstor.org/stable/3219881. Publisher: [University of Utah, Sage Publications, Inc.]. 3.3.1

[20] Kathleen M. Carley. ORA: A Toolkit for Dynamic Network Analysis and Visualization. In Reda Alhajj and Jon Rokne, editors, *Encyclopedia of Social Network Analysis and Mining*. Springer, 2017. 1.3.2

[21] Kathleen M. Carley. Social cybersecurity: an emerging science. *Computational and Mathematical Organization Theory*, 26(4):365–381, December 2020. ISSN 1572-9346. doi: 10.1007/s10588-020-09322-9. URL https://doi.org/10.1007/s10588-020-09322-9. 1.2.1

[22] Dustin Carnahan, Daniel E. Bergan, and Sangwon Lee. Do Corrective Effects Last? Results from a Longitudinal Experiment on Beliefs Toward Immigration in the U.S. *Political Behavior*, 43(3):1227–1246, September 2021. ISSN 1573-6687. doi: 10.1007/s11109-020-09591-9. 5

[23] Man-pui Sally Chan, Christopher R. Jones, Kathleen Hall Jamieson, and Dolores Albarracín. Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*, 28(11):1531–1546, November 2017. ISSN 0956-7976. doi: 10.1177/0956797617714579. 1.1

[24] Lesley Chiou and Catherine Tucker. Fake News and Advertising on Social Media: A Study of the Anti-Vaccination Movement, November 2018. 5

[25] Farhan Asif Chowdhury, Lawrence Allen, Mohammad Yousuf, and Abdullah Mueen. On Twitter Purge: A Retrospective Analysis of Suspended Users. In *ACM Web Conference*, pages 371–378, Taipei Taiwan, April 2020. ACM. ISBN 978-1-4503-7024-0. doi: 10.1145/3366424.3383298. 5

[26] Giovanni Luca Ciampaglia, Alexios Mantzarlis, Gregory Maus, and Filippo Menczer. Research Challenges of Digital Misinformation: Toward a Trustworthy Web. *AI Magazine*, 39(1):65–74, March 2018. ISSN 0738-4602, 2371-9621. doi: 10.1609/aimag.v39i1.2783. URL https://onlinelibrary.wiley.com/doi/10.1609/aimag.v39i1.2783. 3.3.2

[27] Committee on a Decadal Survey of Social and Behavioral Sciences for Applications to National Security, Board on Behavioral, Cognitive, and Sensory Sciences, Division of Behavioral and Social Sciences and Education, and National Academies of Sciences, Engineering, and Medicine. *A Decadal Survey of the Social and Behavioral Sciences: A Research Agenda for Advancing Intelligence Analysis*. National Academies Press, Washington, D.C., 2019. ISBN 978-0-309-48761-0. doi: 10.17226/25335. URL https:

//www.nap.edu/catalog/25335. 1.2.1

[28] Laura Courchesne, Julia Ilhardt, and Jacob N. Shapiro. Review of social science research on the impact of countermeasures against influence operations. *Harvard Kennedy School Misinformation Review*, 2, September 2021. doi: 10.37016/mr-2020-79. 1.1, 1.2.2, 1.2.3

[29] Nicholas Dias, Gordon Pennycook, and David G. Rand. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review*, January 2020. doi: 10.37016/mr-2020-001. 5

[30] Joan Donovan. Why social media can't keep moderating content in the shadows, November 2020. URL https://www.technologyreview.com/2020/11/06/1011769/social-media-moderation-transparency-censorship/. 3.3.1

[31] Ullrich K. H. Ecker, Joshua L. Hogan, and Stephan Lewandowsky. Reminders and repetition of misinformation: Helping or hindering its retraction? *Journal of Applied Research in Memory and Cognition*, 6(2):185–192, June 2017. ISSN 2211-369X, 2211-3681. doi: 10.1037/h0101809. 1.1

[32] Ziv Epstein, Gordon Pennycook, and David G. Rand. Will the Crowd Game the Algorithm? Using Layperson Judgments to Combat Misinformation on Social Media by Downranking Distrusted Sources. In *CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–11, New York, NY, USA, April 2020. Association for Computing Machinery. ISBN 978-1-4503-6708-0. doi: 10.1145/3313831.3376232. 5

[33] Valerie Fointiat. "I know what I have to do, but..." When hypocrisy leads to behavioral change. *Social Behavior and Personality: An International Journal*, 32(8):741–746, January 2004. ISSN 0301-2212. doi: 10.2224/sbp.2004.32.8.741. URL https://www.ingentaconnect.com/content/10.2224/sbp.2004.32.8.741. 3.1.2

[34] Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945, January 2020. ISSN 2053-9517. doi: 10.1177/2053951719897945. 5

[35] Kacper T Gradoń, Janusz A. Hołyst, Wesley R Moy, Julian Sienkiewicz, and Krzysztof Suchecki. Countering misinformation: A multidisciplinary approach. *Big Data & Society*, 8(1), January 2021. ISSN 2053-9517. doi: 10.1177/20539517211013848. 5

[36] Andrew M. Guess, Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 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, July 2020. doi: 10.1073/pnas.1920498117. 5

[37] Paweł Gwiaździński, Aleksander B. Gundersen, Michal Piksa, Izabela Krysińska, Jonas R. Kunst, Karolina Noworyta, Agata Olejniuk, Mikołaj Morzy, Rafal Rygula, Tomi Wójtowicz, and Jan Piasecki. Psychological interventions countering misinformation in social media: A scoping review. *Frontiers in Psychiatry*, 13, 2023. ISSN 1664-0640. doi: 10.3389/fpsyt.2022.974782. 1.2.3

[38] Todd C. Helmus and Marta Kepe. A Compendium of Recommendations for Countering Russian and Other State-Sponsored Propaganda. Technical report, RAND Corporation, June 2021. URL https://www.rand.org/pubs/research_reports/RRA894-1.html. 1.2.2, 1.2.3

[39] Robert A. Huber, Michael L. Wicki, and Thomas Bernauer. Public support for environmental policy depends on beliefs concerning effectiveness, intrusiveness, and fairness. *Environmental Politics*, 29(4):649–673, June 2020. ISSN 0964-4016. doi: 10.1080/09644016.2019.1629171. URL https://www.tandfonline.com/doi/10.1080/09644016.2019.1629171. Publisher: Routledge. 3.3.2

[40] Matthew O. Jackson, Suraj Malladi, and David McAdams. Learning through the grapevine and the impact of the breadth and depth of social networks. *Proceedings of the National Academy of Sciences*, 119(34):e2205549119, August 2022. doi: 10.1073/pnas.2205549119. 5

[41] Se-Hoon Jeong, Hyunyi Cho, and Yoori Hwang. Media Literacy Interventions: A Meta-Analytic Review. *The Journal of Communication*, 62(3):454–472, June 2012. ISSN 0021-9916. doi: 10.1111/j.1460-2466.2012.01643.x. 1.2.2

[42] Se-Hoon Jeong, Hyunyi Cho, and Yoori Hwang. Media Literacy Interventions: A Meta-Analytic Review. *The Journal of Communication*, 62(3):454–472, June 2012. ISSN 0021-9916. doi: 10.1111/j.1460-2466.2012.01643.x. 5

[43] Shagun Jhaver and Amy Zhang. Do Users Want Platform Moderation or Individual Control? Examining the Role of Third-Person Effects and Free Speech Support in Shaping Moderation Preferences, February 2023. arXiv:2301.02208 [cs]. 5

[44] Jialun Aaron Jiang, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. A Trade-off-centered Framework of Content Moderation. *ACM Transactions on Computer-Human Interaction*, 30(1):3:1–3:34, March 2023. ISSN 1073-0516. doi: 10.1145/3534929. 1.2.2, 5

[45] Catherine King, Samantha C. Phillips, and Kathleen M. Carley. Pre-analysis Plan: Predicting Support for Misinformation Countermeasures. Working paper. 1.2.3, 4.1

[46] Catherine King, Christine Sowa Lepird, and Kathleen M. Carley. Project OMEN: Designing a Training Game to Fight Misinformation on Social Media. 2021. URL http://reports-archive.adm.cs.cmu.edu/anon/isr2021/abstracts/21-110.html. 3.2.1, 4.1

[47] Nadejda Komendantova, Love Ekenberg, Mattias Svahn, Aron Larsson, Syed Iftikhar Hussain Shah, Myrsini Glinos, Vasilis Koulolias, and Mats Danielson. A value-driven approach to addressing misinformation in social media. *Humanities and Social Sciences Communications*, 8(1):1–12, January 2021. ISSN 2662-9992. doi: 10.1057/s41599-020-00702-9. URL https://www.nature.com/articles/s41599-020-00702-9. Number: 1 Publisher: Palgrave. 3.3.1

[48] Vasilis Koulolias, Gideon Mekonnen Jonathan, Miriam Fernandez, and Dimitris Sotirchos. *Combating Misinformation: An ecosystem in co-creation*. January 2018. 3.3.1

[49] Aleksandra Lazić and Iris Žeželj. A systematic review of narrative interventions: Lessons

for countering anti-vaccination conspiracy theories and misinformation. *Public Understanding of Science*, 30(6):644–670, May 2021. ISSN 0963-6625. doi: 10.1177/09636625211011881. 1.3.2

[50] Jeffrey Lees, John A Banas, Darren Linvill, Patrick C Meirick, and Patrick Warren. The Spot the Troll Quiz game increases accuracy in discerning between real and inauthentic social media accounts. *PNAS Nexus*, 2(4):pgad094, April 2023. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgad094. 3.2.1

[51] Stephan Lewandowsky and Sander van der Linden. Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology*, 0(0): 1–38, February 2021. ISSN 1046-3283. doi: 10.1080/10463283.2021.1876983. 1.1

[52] Stephan Lewandowsky and Sander van der Linden. Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology*, 0(0): 1–38, February 2021. ISSN 1046-3283. doi: 10.1080/10463283.2021.1876983. 5

[53] Stephan Lewandowsky, John Cook, Ullrich K. H. Ecker, Dolores Albarracín, Michelle A. Amazeen, Panayiota Kendeou, Doug Lombardi, Eryn J. Newman, Gordon Pennycook, Ethan Porter, David G. Rand, David N. Rapp, Jason Reifler, Jon Roozenbeek, Philipp Schmid, Colleen M. Seifert, Gale M. Sinatra, Briony Swire-Thompson, Sander Van Der Linden, Emily K. Vraga, Thomas Wood, and Maria S. Zaragoza. Debunking Handbook 2020. Technical report, Databrary, 2020. URL `http://databrary.org/volume/1182`. 1.2.3, 3.1.1, 5

[54] Sander van der Linden, Jon Roozenbeek, Rakoen Maertens, Melisa Basol, Ondřej Kácha, Steve Rathje, and Cecilie Steenbuch Traberg. How Can Psychological Science Help Counter the Spread of Fake News? *The Spanish Journal of Psychology*, 24:e25, 2021. ISSN 1138-7416, 1988-2904. doi: 10.1017/SJP.2021.23. 1.3.2

[55] Rakoen Maertens, Jon Roozenbeek, Melisa Basol, and Sander van der Linden. Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, 27(1):1–16, 2021. ISSN 1939-2192. doi: 10.1037/xap0000315. 5

[56] Maximilian Maier, František Bartoš, T. D. Stanley, David R. Shanks, Adam J. L. Harris, and Eric-Jan Wagenmakers. No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, 119(31):e2200300119, August 2022. doi: 10.1073/pnas.2200300119. 1.3.2

[57] Maximilian Maier, František Bartoš, T. D. Stanley, David R. Shanks, Adam J. L. Harris, and Eric-Jan Wagenmakers. No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences*, 119(31):e2200300119, August 2022. doi: 10.1073/pnas.2200300119. 5

[58] Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 748–757, December 2020. doi: 10.1109/BigData50022.2020.9377956. 3.4.3

[59] Amy Mitchell and Mason Walker. More Americans now say government should take

steps to restrict false information online than in 2018. *Pew Research Center*, August 2021. URL https://www.pewresearch.org/fact-tank/2021/08/18/more-americans-now-say-government-should-take-steps-to-restrict-false-information-online-than-in-2018/. 3.1.2, 3.3.2

[60] Ariana Modirrousta-Galian and Philip Anthony Higham. Gamified Inoculation Interventions Do Not Improve Discrimination Between True and Fake News: Reanalyzing Existing Research With Receiver Operating Characteristic Analysis, August 2022. 5

[61] Ariana Modirrousta-Galian and Philip Anthony Higham. Gamified Inoculation Interventions Do Not Improve Discrimination Between True and Fake News: Reanalyzing Existing Research With Receiver Operating Characteristic Analysis, August 2022. 3.2.1, 4.1

[62] Garrett Morrow, Briony Swire-Thompson, Jessica Montgomery Polny, Matthew Kopec, and John P. Wihbey. The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*, 73(10):1365–1386, 2022. ISSN 2330-1643. doi: 10.1002/asi.24637. 5

[63] Mohsen Mosleh, Cameron Martel, Dean Eckles, and David G. Rand. 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 *CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–13, New York, NY, USA, May 2021. Association for Computing Machinery. ISBN 978-1-4503-8096-6. doi: 10.1145/3411764.3445642. 1.3.2

[64] Sheryl Wei Ting Ng. Self- and Social Corrections on Instant Messaging Platforms. 2023. 3.1.2

[65] Konrad Niklewicz. Weeding Out Fake News: An Approach to Social Media Regulation. *European View*, 16(2):335–335, December 2017. ISSN 1781-6858, 1865-5831. doi: 10.1007/s12290-017-0468-0. 5, 5

[66] Brendan Nyhan, Jason Reifler, Sean Richey, and Gary L. Freed. Effective Messages in Vaccine Promotion: A Randomized Trial. *Pediatrics*, 133(4):e835–e842, March 2014. ISSN 0031-4005. doi: 10.1542/peds.2013-2365. 1.3.2

[67] Tomasz Oleksy, Anna Wnuk, Dominika Maison, and Agnieszka Łyś. Content matters. Different predictors and social consequences of general and government-related conspiracy theories on COVID-19. *Personality and Individual Differences*, 168:110289, January 2021. ISSN 01918869. doi: 10.1016/j.paid.2020.110289. URL https://linkinghub.elsevier.com/retrieve/pii/S0191886920304797. 1.1

[68] Andrea E. O'Rear and Gabriel A. Radvansky. Failure to accept retractions: A contribution to the continued influence effect. *Memory & Cognition*, 48(1):127–144, July 2019. ISSN 1532-5946. doi: 10.3758/s13421-019-00967-9. 5

[69] Gordon Pennycook and David G. Rand. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50, July 2019. ISSN 0010-0277. doi: 10.1016/j.cognition.2018.06.011. URL http://www.sciencedirect.com/science/article/pii/S001002771830163X. 3.1.2

[70] Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand. The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science*, 66(11):4944–4957, November 2020. ISSN 0025-1909. doi: 10.1287/mnsc.2019.3478. 5

[71] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. Shifting attention to accuracy can reduce misinformation online. *Nature*, pages 1–6, March 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03344-2. 5

[72] Christina Peter and Thomas Koch. When Debunking Scientific Myths Fails (and When It Does Not): The Backfire Effect in the Context of Journalistic Coverage and Immediate Judgments as Prevention Strategy. *Science Communication*, 38(1):3–25, February 2016. ISSN 1075-5470, 1552-8545. doi: 10.1177/1075547015613523. 5

[73] Steve Rathje, Claire Robertson, William J. Brady, and Jay J. Van Bavel. People think that social media platforms do (but should not) amplify divisive content, October 2022. URL `https://psyarxiv.com/gmun4/`. 5

[74] Timothy S. Rich, Ian Milden, and Mallory Treece Wagner. Research note: Does the public support fact-checking social media? It depends who and how you ask. *Harvard Kennedy School Misinformation Review*, November 2020. doi: 10.37016/mr-2020-46. URL `https://misinforeview.hks.harvard.edu/?p=3861`. 3.3.2

[75] Alex Rochefort. Regulating Social Media Platforms: A Comparative Policy Analysis. *Communication Law and Policy*, 25(2):225–260, April 2020. ISSN 1081-1680, 1532-6926. doi: 10.1080/10811680.2020.1735194. 3.3.1

[76] Jon Roozenbeek and Sander van der Linden. Breaking Harmony Square: A game that "inoculates" against political misinformation. *Harvard Kennedy School Misinformation Review*, November 2020. doi: 10.37016/mr-2020-47. 3.2.1

[77] Jon Roozenbeek, Sander van der Linden, and Thomas Nygren. Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review*, 1(2), February 2020. doi: 10.37016//mr-2020-008. 1.1, 1.3.2

[78] Emily Saltz, Soubhik Barari, Claire R. Leibowicz, and Claire Wardle. Misinformation interventions are common, divisive, and poorly understood. *Harvard Kennedy School Misinformation Review*, October 2021. doi: 10.37016/mr-2020-81. 3.1.2

[79] Ian Skurnik, Carolyn Yoon, Denise C. Park, and Norbert Schwarz. How Warnings about False Claims Become Recommendations. *Journal of Consumer Research*, 31(4):713–724, March 2005. ISSN 0093-5301. doi: 10.1086/426605. 5

[80] Brian G. Southwell, Emily A. Thorson, and Laura Sheble. *Misinformation and mass audiences*. 3.1.2

[81] Jeff Stone and Nicholas C. Fernandez. To Practice What We Preach: The Use of Hypocrisy and Cognitive Dissonance to Motivate Behavior Change. *Social and Personality Psychology Compass*, 2(2):1024–1051, 2008. ISSN 1751-9004. doi: 10.1111/j.1751-9004.2008.00088.x. URL

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-9004.2008.00088.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-9004.2008.00088.x. 3.1.2

[82] Edson C Tandoc, Darren Lim, and Rich Ling. Diffusion of disinformation: How social media users respond to fake news and why. *Journalism*, 21(3):381–398, March 2020. ISSN 1464-8849. doi: 10.1177/1464884919868325. 3.1.2, 3.1.2

[83] Li Qian Tay, Mark J. Hurlstone, Tim Kurz, and Ullrich K. H. Ecker. A comparison of prebunking and debunking interventions for implied versus explicit misinformation. *British Journal of Psychology*, 113(3):591–607, 2022. ISSN 2044-8295. doi: 10.1111/bjop.12551. 1.2.3

[84] Benjamin Toff and Nick Mathews. Is Social Media Killing Local News? An Examination of Engagement and Ownership Patterns in U.S. Community News on Facebook. *Digital Journalism*, 0(0):1–20, October 2021. ISSN 2167-0811. doi: 10.1080/21670811.2021.1977668. 5

[85] Joshua A. Tucker, Yannis Theocharis, Margaret E. Roberts, and Pablo Barberá. From Liberation to Turmoil: Social Media and Democracy. *Journal of Democracy*, 28(4):46–59, October 2017. doi: 10.1353/jod.2017.0064. URL https://www.journalofdemocracy.org/articles/from-liberation-to-turmoil-social-media-and-democracy/. 1.1

[86] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. Contestability For Content Moderation. *Proceedings of the ACM on Human-Computer Interaction*, 5 (CSCW2):318:1–318:28, October 2021. doi: 10.1145/3476059. 5

[87] Nathan Walter and Sheila T. Murphy. How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, 85(3):423–441, July 2018. ISSN 0363-7751. doi: 10.1080/03637751.2018.1467564. 2

[88] Nathan Walter, John J. Brooks, Camille J. Saucier, and Sapna Suresh. Evaluating the Impact of Attempts to Correct Health Misinformation on Social Media: A Meta-Analysis. *Health Communication*, 36(13):1776–1784, November 2021. ISSN 1532-7027. doi: 10.1080/10410236.2020.1794553. 1.2.3, 3.1.1

[89] Nathan Walter, John J. Brooks, Camille J. Saucier, and Sapna Suresh. Evaluating the Impact of Attempts to Correct Health Misinformation on Social Media: A Meta-Analysis. *Health Communication*, 36(13):1776–1784, November 2021. ISSN 1532-7027. doi: 10.1080/10410236.2020.1794553. 5

[90] Benjamin R. Warner and Ryan Neville-Shepard. Echoes of a Conspiracy: Birthers, Truthers, and the Cultivation of Extremism. *Communication Quarterly*, 62(1):1–17, January 2014. ISSN 0146-3373. doi: 10.1080/01463373.2013.822407. URL https://doi.org/10.1080/01463373.2013.822407. Publisher: Routledge _eprint: https://doi.org/10.1080/01463373.2013.822407. 1.1

[91] Thomas Wood and Ethan Porter. The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. *Political Behavior*, 41(1):135–163, March 2019. ISSN 1573-6687. doi: 10.1007/s11109-018-9443-y. 5

[92] Kamya Yadav. Countering Influence Operations: A Review of Policy Proposals Since 2016. Technical report, Carnegie Endowment for International Peace, November 2020. URL https://carnegieendowment.org/2020/11/30/countering-influence-operations-review-of-policy-proposals-since-2016-pub-83333. 3.3.1

[93] Kamya Yadav. Platform Interventions: How Social Media Counters Influence Operations. Technical report, Carnegie Endowment for International Peace, January 2021. URL https://carnegieendowment.org/2021/01/25/platform-interventions-how-social-media-counters-influence-operations-pub-83698. 1.2.3, 3.3.1, 5

Countermeasures Definitions

This appendix describes the labels given to countermeasures papers in the citation network analysis project described in Chapter 1.

Content Distribution

- *Content Distribution* - How content is distributed on social media. This includes limiting how many people you can forward a message to [40, 73].
- *Redirection* - A type of content distribution where users are redirected to other content or no content when searching for something. For example, a user searching for COVID-19 leading to a CDC information box [93].
    - *Nudging* - A type of content distribution or redirection where people are nudged in some way. Typically, this involves reminding people about accuracy. For example, pop-ups asking if the user is sure they want to post something even if they haven't opened the link [57, 71].

Content and Account Moderation

- *Content Moderation* - A general category of interventions related to how content is shown or not shown on social media. This includes fact-checking, narrative counterspeech, among others [44, 86].
    - *Fact-Checking* - The process of verifying information. This verification can be done by experts, journalists, platforms, and/or users and includes multi-modal fact-checking, such as fact-checking videos [18, 89].
    - *Debunking* - Debunking is a stronger form of fact-checking, where context and coherence is typically given in addition to verifying or correcting content. It can also be described as a "narrative intervention" [53, 72].
    - *Misinformation Detection* - The algorithmic detection of misinformation. Usually for the purposes of content moderation [35].
    - *Algorithmic Content Moderation* - Automated content moderation. This can include automated fact-checking, downranking of content, removing of content, or labeling of content [13, 34].
    - *Continued-Influence Effect* - This category related to the effectiveness of moderation or corrections, possible backfire effects or lack thereof, and the process of debiasing individuals [22, 91].

- *Account Moderation* - Moderation involving a user account. Examples are suspending, banning users, or shadowbanning users [25].
    - *Deplatforming* - A specific type of account moderation where users are completely removed from a platform or multiple platforms as a way to limit the spread of their content [2].

Content Labeling

- *Content Labeling* - This category includes all general types of misinformation disclosure. Content labels are often used to display fact-checks or additional context on a post. This intervention is related to general *Content Moderation*, *Fact-Checking*, and *Debunking* [62].
    - *Crowdsourcing* - Crowdsourcing typically involves asking regular individuals to verify information and label content rather than asking journalists or expert fact-checkers [3, 32].
    - *Source Credibility* - Disclosing or labeling the credibility of a post's source [7, 29].

Advertising

Advertising policy encompasses items such as banning political ads, requiring ads to go through a fact-checking service before posting, or banning certain advertisers [24].

Media Support

Investing in local news; or promoting local or reliable news on social media platforms [84].

Media Literacy and Awareness

- *Media Literacy* - This category involves any educational or training effort meant to increase the public's civic reasoning and critical thinking skills when engaging with media messages [36, 42].
    - *Fake News Games* - Games that are designed to help players detect misinformation and improve their critical thinking skills [55, 60].
    - *Inoculation* - Often known as "pre-bunking", inoculation involves warning messages or other interventions meant to prevent people from later believing misinformation [52].
    - *Proactive Warning* - A warning about the possibility of being misled, either through a label or a media literacy training [70, 79].
- *Data Sharing* - How researchers can get access to high-quality, relevant data from social media platforms and other researchers, while maintaining privacy and considering the ethics of the studies involved [8, 15].

User-based Countermeasures

- *User-based Countermeasures* - This category involves people seeing or hearing misinformation, and how they respond to it in real-time. It also includes community moderation [11, 43].
    - *Reporting* - Users can report users or their posts [65].

    - *Social Corrections* - Users employing fact-checking or debunking directly with a poster of misinformation. This includes publicly commenting on a post or private messaging the poster [11].
    - *Retraction* - This category includes when users or organizations retract misinformation they posted, and how that affects individuals who have already seen the misinformation [68].

Other Interventions

- *Government Regulation* - This label encompasses any laws, rules, or regulations at local, state, or federal levels. It may involve requiring platforms to adopt previously discussed interventions. Government-specific interventions also include breaking up technology companies, regulating platforms like media or utility companies, or controlling speech on social media platforms [65].
- *Combining Interventions* - Papers that specifically compare the impact of using multiple interventions at once with using one intervention.

Other Qualitative Labels

- *Review Article* - This label is given to papers that reviews other papers. A review article could review papers in a specific area or they can be broader.
- *Acceptance* - These papers focus on user acceptance, the popularity of interventions, general public opinion, or political feasibility.

Other Quantitative Labels

- *Meta-Analysis* - A meta-analysis is a type of review paper that specifically analyzes all the previous results in a certain section of the literature.
- *Effectiveness* - Any paper that discusses or directly measures the effectiveness or one or more interventions.